Novel statistical modelling approaches for pesticide residues.

Jacobus Roelofs

A Thesis presented for the degree of
Doctor of Philosophy

Statistics and Probability Group
Department of Mathematical Sciences
Durham University
United Kingdom

August 2013

Dedication

To my wonderful wife Vicki

and

my fantastic son William

Novel statistical modelling approaches for pesticide residues.

Jacobus Roelofs

Abstract

Plant protection products play an important role in protecting our food supply against pests, diseases and weeds. As global food demand rises, their role in maintaining the quality and quantity of our food production is likely to increase in the absence of other control methods. To manage the risks associated with pesticide usage, EU laws regulate the placing of plant protection products on the market and the monitoring of pesticide residues in food. This involves assessing the potential risks associated with human dietary exposure by conducting dietary risk assessments which take both consumption patterns and residue levels of pesticides in and on food items into account. Residue levels will vary from one food item to the next so we need to know what the distribution of residues over food items is in order to assess how high residue levels can be.

In this thesis we introduce novel statistical approaches that can be used to obtain better estimates of the variation and uncertainty in pesticide residue levels on raw agricultural products. The first approach uses monitoring data and pesticide usage information to model the correlation in pesticide residue levels when multiple pesticides have been used. Next we introduce an approach that can be used to describe the variation in log-residue levels in units, assuming that multiple data sets share a common shape. The final model describes both within-field and between-field variation of residue levels. These new approaches, which provide promising alternatives to existing methods, can be implemented in existing dietary risk assessment software and will expand the suite of models available to risk assessors when assessing dietary exposure to pesticides.

Declaration

I declare that the research presented in this thesis is, to the best of my knowledge, original. Where other work is quoted, due reference has been made.

Acknowledgements

I would like to thank my supervisor Dr. Peter Craig for all his guidance, encouragement and support during the development of this work. He has also provided numerous suggestions helping me with the scientific communication of the approaches presented in this thesis. I would also like to thank both Sara and Gavin Montgomery for their love and encouragement as well as providing me with comments on draft versions of this thesis and innumerable cups of tea throughout this project! I would also like to express my gratitude to my parents and sister for their love and support. My biggest thanks go to my wife Vicki for her enduring love, support and help. Without her, this 7 year project would have most likely ended without this thesis and she will be glad to know that her proofreading days are now finally over! Last but not least, I would like to thank my little boy William for sleeping through many of the hours spent on this work and providing me with not-so-subtle reminders that there are more important things in life than this thesis (like reading the Gruffalo)!

Contents

List of Figures

List of Tables

List of Abbreviations

ARfD	Acute Reference Dose
CAC	Codex Alimentarius Commission
cGAP	Critical Good Agricultural Practice
DAG	Directed Acyclic Graph
DAR	Draft Assessment Report
DP	Dirichlet Process
DPMN	Dirichlet Process Mixture of Normal distributions
EC	European Commission
EFSA	European Food Safety Authority
EU	European Union
GAP	Good Agricultural Practice
HR	Highest Residue in Supervised Trials
IESTI	International Estimate of Short Term Intake
LOD	Limit of Determination
MCMC	Markov Chain Monte Carlo
MRL	Maximum Residue Level
NOAEL	No-observed-adverse-effect-level

OECD	Organisation for Economic Co-operation and Development
PPR	Panel on Plant Protection Products and their Residues
PRAPeR	Pesticide Risk Assessment Peer Review
PUS	Pesticide Usage Survey
RAC	Raw Agricultural Commodity
RMS	Rapporteur Member State
SCoFCAH	Standing Committee on the Food Chain and Animal Health
STMR	Supervised Trial Median Residue
VF	Variability Factor
WHO	World Health Organisation

Chapter 1

Introduction to Human Dietary Risk Assessment

1.1 Introduction

Pesticides are used to protect crops before and after harvest from infestation by pests and plant diseases. A pesticide is any substance, preparation or organism prepared or used, to protect plants or wood or other plant products from harmful organisms, to regulate the growth of plants, to give protection against harmful creatures, or to render such creatures harmless (FEPA, 1985). A possible consequence of pesticide use on food crops may be the presence of pesticide residues in or on treated products. Residue levels will vary from one food item to the next and to account for this, we need to know what the distribution of residues is over food items in order to assess how high residue levels can be.

To assess the dietary risk associated with pesticide residues, information is needed about the residue levels associated with food items and the consumption of food items. This chapter will describe the regulatory context for dietary risk assessment in the European Union (EU) as well as the pesticide registration process. We will describe the current deterministic approach to dietary risk assessment and the recently developed probabilistic alternatives. We will also discuss several issues with the

quality and quantity of data available and with the existing modelling approaches. Finally, we present the motivation for this thesis followed by a short overview of how we propose to overcome some of the obstacles associated with current practices which we then develop further in Chapters 3, 4 and 5.

1.2 Regulatory Context

The regulation of pesticides, commonly referred to as plant protection products, in the EU was first harmonised under Council Directive 91/414/EEC (EC, 1991). This Directive established agreed criteria for considering the safety and effectiveness of formulated plant protection products. The Directive set out a two-stage assessment system which focuses on a consideration of the safety of active substances at the EU level and (once safety of the active substance has been established) the authorisation of formulated products at a national level.

1.2.1 Active Substance Authorisation

The two most important regulatory tools in the EU for plant protection products are Directive 1107/2009 (EC, 2009) on the placing of plant protection products on the market and Regulation 396/2005 (EC, 2005) on maximum residue levels of pesticides allowed in food and animal feed. Directive 1107/2009 regulates the use of plant protection products and their residues in food and it provides procedures for approval of active substances and plant protection products containing these substances. This Directive states that substances cannot be used in plant protection products unless an appropriate risk assessment has shown that the substance is without unacceptable risk to people or the environment. The Directive aims to harmonise the authorisation process of plant protection products within the EU and to establish a list of active substances (Regulation 540/2011; EC, 2011a), that have been shown to be without unacceptable risk. The process for deciding whether an active substance can be included in the list of approved active substances eligible for use in plant protection products in the EU involves all the Member States, the European Food Safety Authority (EFSA) and the European Commission (EC). Once a

substance is included in the list of approved active substances Member States may authorise the use of products containing them (see Section 1.2.2).

The active substance authorisation process starts with an application being made by a company, the notifier, for the inclusion of a new or existing active substance in the list of approved active substances. Authorisations can be granted for a fixed period of up to 10 years. After this period, the authorisation may be renewed after verification that the standards then in force are adhered to. An application needs to be supported by a dossier which contains the required data (as specified in Regulation 545/2011; EC, 2011b) including information on the physical and chemical properties of the active substance and its effects on target pests and on non-target organisms. As these properties may depend on characteristics of the plant protection product in which the active substance is used, detailed information on at least one proposed plant protection product must be included to support the proposed use or uses. The dossier will include a risk assessment for any possible effects on workers/operators, consumers, the environment and non-target plants and animals. On behalf of the EC, a Rapporteur Member State (RMS) will evaluate the dossier in the areas of physical chemical properties, analytical methods, mammalian toxicology, operator exposure, environmental fate and ecotoxicology. The evaluation of the submitted studies, a risk assessment and a proposal for inclusion or non-inclusion of the active substance in the approved list of substances is summarised in a Draft Assessment Report (DAR).

The RMS submits the DAR to the Pesticide Risk Assessment Peer Review (PRAPeR) unit of EFSA. EFSA was established in 2002 as an independent European Agency whose role includes providing independent scientific advice to the EC and European Community Member States concerning plant protection products. The PRAPeR unit is responsible for making arrangements for the distribution of the DAR to all Member States and for collecting comments from both Member States and the general public, the latter via open public consultations. The RMS will respond to the comments received and the responses will be evaluated by EFSA experts. Com-

ments that were not addressed satisfactorily may be discussed in expert meetings with experts drawn from Member States and EFSA. The outcome of the expert discussions will be recorded in EFSA's draft conclusion document which will be circulated to all Member States before it is finalised. EFSA then presents a comprehensive summary of the risk assessment to the EC, Member States and the notifier in a report which will be considered by the Member States and the EC. Depending on the conclusions and a consideration of risk management options, the EC will then propose whether or not to include the substance in Regulation 540/2011 (EC, 2011a) subject to a vote by Member States. In formulating a proposal for a decision, the EC may consult with Member States at the Standing Committee on the Food Chain and Animal Health (SCoFCAH). In special cases, clarifications may also be sought from EFSA on aspects of the risk assessment, e.g. by referring open issues to EFSA's independent Panel on Plant Protection Products and their Residues (PPR) for further consideration. In addition, confirmatory data requirements may be identified to support decision making about plant protection products after inclusion of the active substance in the list of approved active substances.

Once an active substance has been approved, Member States must ensure that all authorised plant protection products which contain this active substance, comply with Directive 1107/2009. This ensures that authorisations issued in all Member States are assessed to the same standards. After a decision to remove an active substance from Regulation 540/2011, Member States must apply for withdrawal of products containing the active substance within a timescale defined in the decision.

1.2.2 Plant Protection Product Authorisation

Once approval is granted for the active substance at the EU level, Member States may approve the uses of a specific product if all the data and/or information on the safety, efficacy and, where relevant, humaneness of the pesticide are considered to be acceptable. Before any pesticide can be used, sold, supplied, advertised or stored it must be approved for use. Pesticide approvals may at any time be subject to review, amendment, suspension or revocation. Revocation of approval may occur

for various reasons, e.g. the identification of safety concerns or an approval holder's failure to meet a data submission deadline. On expiry or revocation of approvals it becomes unlawful to advertise, sell, supply, store or use the products.

1.2.2.1 Maximum Residue Levels (MRLs)

To assess whether pesticides are applied in accordance with the conditions of use set by Member States, legal limits on residues in or on food are set which are referred to as maximum residue levels (MRLs). If residue levels in food items are above the MRL for a particular product, this may suggest that the product was not applied to crops in accordance with the conditions of use set by the Member State's approval. Regulation 396/2005 (EC, 2005) establishes the MRLs of pesticides permitted in products of plant or animal origin intended for human or animal consumption. The Regulation replaces all national MRLs with harmonised EU MRLs for all food items. It facilitates the harmonisation of pesticide MRLs whilst ensuring better consumer protection throughout the EU. The EC decided to set the MRLs for active substances which are no longer used in agriculture in or outside the EU at the limit of determination (LOD), the lowest level surveillance laboratories can measure. For the remaining substances that are still in use, temporary EU MRLs have been set at the highest national level MRLs, indicating that MRLs are primarily intended as trading standards. Where uses of pesticides are not authorised at the EU level (e.g. because the product is considered to be unsafe) or authorised use does not result in detectable levels of residues, the MRL is set at the LOD. The MRL is also set to the LOD for crops on which there are no uses of the pesticide.

In addition to statutory EU MRLs, international non-statutory (Codex) levels are set for a wide variety of pesticide/commodity combinations. The Codex Alimentarius Commission (CAC), responsible for setting Codex MRLs, is an international body that aims to protect the health of consumers, ensure fair trade practices in the food trade and promote co-ordination of all food standards work undertaken by international governmental and non-governmental organisations. Codex sets MRLs for countries which do not have their own MRL-setting capacity and aims for the

harmonisation of MRL-setting. The Codex MRLs may help inspection services to decide whether imported agricultural products containing traces of residues can be further traded. However, where produce is marketed within the EU and an EU MRL exists, it is the EU MRL that must be complied with. Regulation 396/2005 (EC, 2005) states that MRLs set at the international level by the CAC should be considered when EU MRLs are being set. To harmonise the MRL setting process even further, the Organisation for Economic Co-operation and Development (OECD) developed a MRL calculation procedure to support experts in the derivation of MRLs (OECD, 2011b).

1.2.2.2 Surveillance Programmes

Directive 396/2005/EC (EC, 2005) states that Member States shall establish multi-annual national control programmes for pesticide residues. These surveillance programmes aim to monitor the levels of pesticide residues in food to ensure that residue levels do not exceed the statutory MRLs for approved products as MRL exceedance may indicate that there are incidents of misuse. EC Directive 2002/63/EC (EC, 2002) specifies sampling procedures for these surveillance programmes. The programmes are designed to select the majority of food items at random with the remainder coming from targeted sampling based on, e.g. the violation rate in previous years. If the results of the monitoring programmes suggest that pesticides are not being applied in accordance with the approved conditions of use, Member States may take enforcement action.

1.3 The Pesticide Registration Process

Pesticide registration involves an assessment of a population's dietary intake of a pesticide. In this section we first provide a detailed overview of the data available and how these data are used in dietary risk assessments for the pesticide registration process. Then we will outline both the deterministic and currently available probabilistic approaches for calculating dietary intake. For brevity, we will restrict our focus to acute (short-term) intake assessment.

1.3.1 Data

In this section we will briefly discuss the types of data that are available for dietary risk assessment for the pesticide registration process and how they are obtained.

1.3.1.1 Residue Levels

The EU framework for risk assessment of pesticides results in the collection of two types of residue level data related to human dietary risk assessment. Before approval is granted, notifiers have to provide supervised field trial data which are used in the risk assessment that is conducted as part of the DAR. Following approval, pesticide residue levels will be monitored in food products to determine any MRL exceedance and to indicate whether unauthorised pesticides have been applied.

Supervised Field Trial Data

For the authorisation of a new use, the only residue data available come from a number of supervised field trials. These trials are conducted to determine the magnitude of the pesticide residue in or on raw agricultural commodities (RACs) and are designed to reflect pesticide use patterns that lead to the highest possible residues under 'Critical Good Agricultural Practice' (cGAP). This is the GAP selected to represent the worst-case use scenario that produces the highest possible field residues on crop commodities. It usually includes the maximum use-rate, the maximum number of applications and the minimum re-treatment and pre-harvest intervals (OECD, 2011a). Supervised field trial data are used to propose MRLs and to provide the Supervised Trial Median Residue ($STMR$) and Highest Residue (HR) values for use in intake assessments. Generally, composite samples consisting of several units of a raw agricultural commodity are obtained from a supervised field trial (OECD, 2009).

EC (1997) and OECD (2009) provide guidelines for supervised field trials and give an overview of a wide range of considerations that need to be taken into account when conducting them. Field trial characteristics include:

Number of Trials: The precise number of trials necessary is difficult to determine

in advance of a preliminary evaluation of the trial results. Assuming comparability can be established between production areas (e.g. climate, application techniques, growing seasons, etc.), a minimum of eight trials representative of the proposed growing area is required for major crops. For minor crops normally four trials representative of the proposed growing area are required. If comparability cannot be established, more trials should be conducted to represent the variation in conditions.

Site Selection: Supervised field trials which are carried out in open fields should include data from four different sites in the same growing season. For applications under glass, a single site is sufficient as the conditions are controlled. Trials should be conducted in regions where the crops are predominantly grown commercially and should reflect the main types of agricultural practice, especially if this has a significant impact on residue levels. Furthermore, the sites should be chosen to reflect variations in weather conditions, different types of soil and the special characteristics of each crop.

Plot Size: The plot size depends on the crop but should be large enough to allow application of the test substance in a manner which reflects routine use and such that sufficient representative samples can be obtained.

Post-harvest Treatment: Records should be kept on post-harvest treatments and storage location conditions for those crops that are routinely treated or stored after harvesting (e.g. potatoes, seeds, etc.).

Application: Supervised field trials should be based on the highest proposed rate of application consistent with GAP. Test substance applications should not be made in strong wind, during rain or when rainfall is expected shortly after application. The formulation should be the intended formulation of the product for the crop or commodity. The maximum proposed label rate, the maximum number of applications and minimum treatment interval should be used when applying the test substance. Application timing is determined by plant growth stage and/or the number of days prior to harvest. If a specific minimum pre-harvest interval is indicated on the label (e.g. ‘Do not apply

this product less than 14 days prior to harvest.'), it should be used in the field trials.

Sampling of RACs: For the purpose of MRL setting, samples taken from supervised field trials should be of the whole RAC as it is used in the food supply chain. The residue level on the edible portion of the commodity needs to be obtained for use in dietary risk assessment (WHO, 1997). For plants or plant products with inedible skin (such as citrus, banana, kiwi, pineapple) a separate analysis of flesh and skin should be performed on some samples in order to provide data on the distribution of residues between flesh and skin (EC, 1997). For some crops, there may be more than one RAC (e.g. maize). Guidelines for the sampling strategy for RACs from supervised field trials are provided in EC (1997).

Monitoring Data

Residue level data may also be available from monitoring surveys. These surveys do not only focus on pesticides that have been approved but may also test for pesticides that have not been approved in order to assess compliance with approval regulations. EC Directive 2002/63/EC (EC, 2002) specifies sampling procedures for the official control of pesticide residues in and on products of plant and animal origin. The procedure is based on taking a representative sample from a 'lot'. A 'lot' is defined as a quantity of a food material delivered at one time and presumed to have uniform characteristics such as origin, producer, variety, etc. The guidelines specify the quantity to sample, both in terms of the total weight and the number of units. The number of units do not necessarily correspond to the number of units that are sampled in supervised field trials: for example, in supervised field trials a composite sample of cucumbers will consist of 12 units whereas in monitoring surveys the number of units is at least 5. However, there is little information available on how commodities and pesticides should be selected for inclusion in monitoring programmes. EFSA (2011) states that many countries determine the sampling frequency of different commodities based on the results of previous monitoring programmes (monitoring of similar crops to determine trends in residue levels), food

consumption figures and exceedances in previous years. Therefore, the extent of monitoring programmes varies between countries and different amounts of data will be available.

1.3.1.2 Consumption Data

For dietary intake assessments, consumption data is obtained from dietary surveys. The most basic survey is a food frequency survey in which participants record or recall the number of occasions each food was consumed over a specified period of time (Brandstetter et al., 1999). Another type of survey is a 24 hour recall study in which the quantities consumed are retrieved in the course of an interview. The interviewer may use appropriate memory aids (e.g. photographs of prepared dishes and/or calibrated portion sizes) and information on cooking methods, recipes and labels of industrially prepared foods may also be retrieved (Lallukka et al., 2001). A further type of survey is a dietary record survey which involves recording the amount of food consumed in a specified period of time. These surveys can either be based on weighing all foods prior to their consumption or comparing the food with photographs of calibrated portion sizes (Gregory et al., 2000; Hoare et al., 2004; Ocké et al., 2007; VCP, 1998).

Figure 1.1 shows an overview of the general characteristics of dietary surveys and a few examples of surveys that have been conducted in EU countries. To obtain an EU-wide conservative intake estimate for dietary risk assessments, it is important to obtain a representative sample of consumption in each country as EU subpopulations may have different dietary habits.

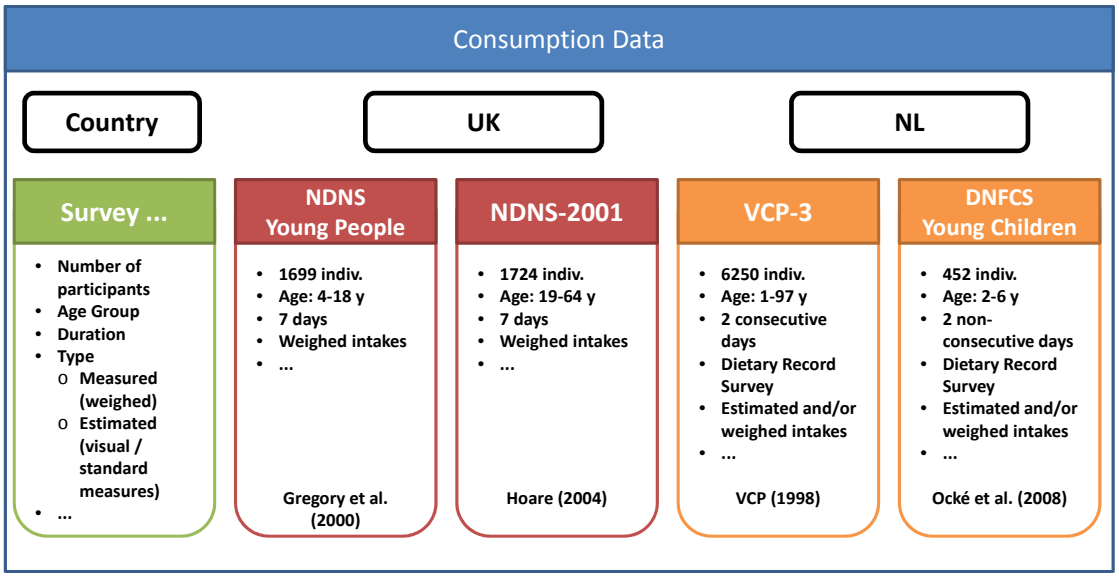


Figure 1.1 – *Examples of existing dietary survey data.*

Figure 1.2 provides an overview of how information from dietary surveys are processed before they can be used in dietary risk assessments. For each person a daily record of which food items were consumed during various eating events (e.g. a pizza for dinner) is available. For dietary risk assessments, we need to estimate how many units of RACs were consumed and how much each of them weighs. Therefore, these data may have to be converted from a portion size to a weight-based amount (using photographs of food items of various portion sizes, e.g. if the portion consumed is similar to the photograph of a medium pizza, a weight of 300 grammes of pizza is assigned to the eating event). Processed food items will have to be converted into ingredients (e.g. tomato puree, mushroom slices), which then need to be converted into RACs (e.g. tomatoes, mushrooms). This is done using generic recipe databases and may depend on the food item's brand. Conversion into RACs is necessary because residue data are collected at the RAC level.

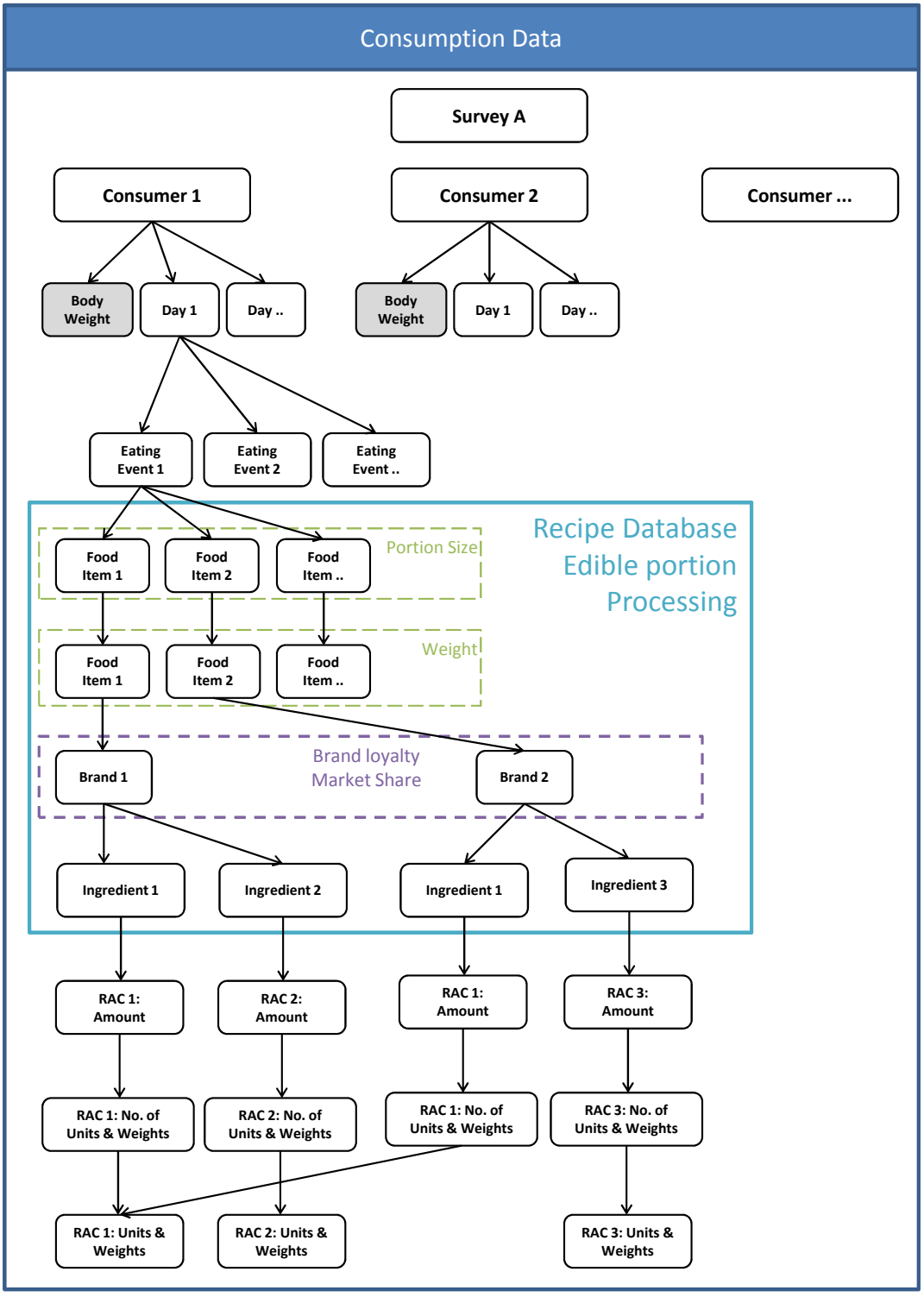


Figure 1.2 – *Generic approach for modelling of consumption data in dietary risk assessments.*

1.3.2 Current Approaches

The current approach for dietary risk assessment as part of the pesticide registration process is deterministic and involves three steps:

1. Conduct supervised field trials to provide information on residue levels in and on RACs.

2. Deterministic intake assessment using the International Estimate of Short Term Intake (IESTI) equations. These are based on conservative consumption estimates and conservative residue levels obtained from supervised field trials.

3. Comparison of the intake assessment with an acceptable intake estimate leading to acceptance or rejection of the pesticide use and the MRL.

We discuss each step in detail in the following sections and a summary of the process is shown in Figure 1.3.

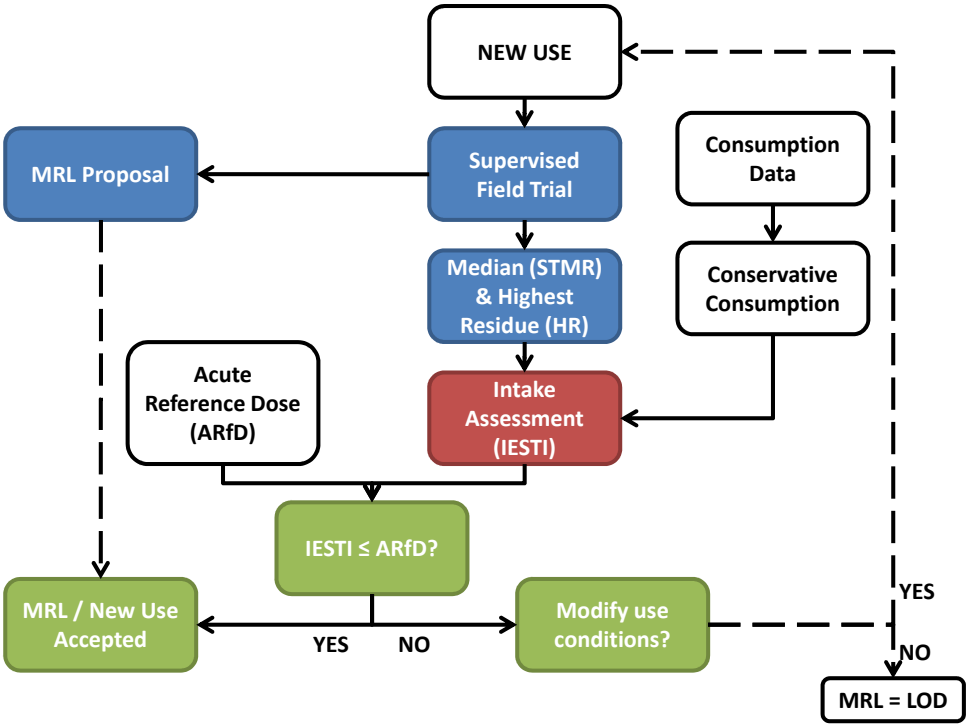


Figure 1.3 – *Use of supervised field trial data for dietary risk assessment and MRL setting.*

1.3.2.1 Residues from Supervised Field Trials

Supervised field trial data (see Section 1.3.1.1) are used to propose MRLs and to provide the Supervised Trials Median Residue (*STMR*) and Highest Residue (*HR*) for use in intake assessments (blue boxes in Figure 1.3).

1.3.2.2 Intake Assessment

Dietary risk assessment for pesticides focuses on effect levels and intake estimates in order to establish that pesticide usage is unlikely to lead to impacts on health when a high-residue unit is consumed or when someone consumes a treated product over a longer period. Intake estimation is based on two factors: residue levels on food items and consumption amounts of food items. Regulation EC 396/2005 (EC, 2005) states that the acute exposure of consumers to pesticide residues via food products should be evaluated taking into account the guidelines published by the World Health Organisation (WHO, 1997).

Intake assessments are based on the following simple equation:

$$\text{Intake (mg/kg/day)} = \frac{\text{Amount Consumed (kg/day)} \times \text{Concentration (mg/kg)}}{\text{Body weight (kg)}}$$

where consumption is divided by body weight to enable a comparison with the outcome of a toxicological effect assessment. The World Health Organisation (WHO) proposed the IESTI equations as a measure of acute dietary exposure (JMPR, 2002). To calculate the IESTI the following definitions are used:

LP	Largest portion provided (kg food/day).
$STMR$	Supervised trials median residue (mg/kg food).
$STMR_P$	Supervised trials median residue (mg/kg food) in processed commodity, calculated by multiplying the $STMR$ in the raw commodity by a processing factor.
HR	Highest residue (mg/kg food) in composite sample of edible portion from the supervised field trials from which the proposed MRL and $STMR$ were derived.

HR_P Highest residue (mg/kg food) in the processed commodity, calculated by multiplying the HR in the raw commodity by a processing factor.

bw Average consumer body weight (kg), from the country that provided the dietary survey with the selected largest portion, LP.

U Unit weight (kg) of edible portion, converted from the RAC provided by a country in the region where the supervised field trials were carried out that resulted in the highest residue level.

v The variability factor, v, is a measure used to reflect the variability of residue levels in or on individual commodity units and is defined as the 97.5^{th} percentile of the distribution of unit residues divided by the mean residue level (EFSA, 2005). It is applied to account for the fact that some of the units making up the composite sample may have had higher residue levels than the residue level of the composite sample itself.

For the deterministic IESTI calculations, the 97.5^{th} percentile consumption value of a RAC is often used as the LP (JMPR, 2002). This means that 2.5% of the population is consuming a larger portion of the RAC than the LP. However, as the IESTI equations consist of some conservative estimates (e.g. HR) and residue levels from supervised field trials are assumed to be higher than residue levels in food items available on the market, it is unclear what level of protection is achieved.
The IESTI is calculated using one of 3 standard equations, depending on the type of commodity involved (JMPR, 2002):

Case 1

This case is used for commodities for which a meal-sized portion consists of a number of units that is similar to the number of units in a composite sample (e.g. peanuts, grapes). The concentration of residue in a composite sample (raw or processed) reflects that in a meal-sized portion of the commodity (unit weight <25g).

$$IESTI = \frac{LP \times (HR \text{ or } HR_P)}{bw}$$

Case 2

This case reflects the situation where a consumer eats a few units (but less than the number in a composite sample) in one day, one of them possibly having high residue levels (e.g. apples). A meal-sized portion, such as a single piece of fruit or vegetable, might have a higher residue than the composite (unit weight of the whole portion is >25g). Standard variability factors, v, are applied in the equation unless sufficient data are available on residues in single units to calculate a more realistic variability factor.

$$IESTI = \begin{cases} \frac{U \times (HR \text{ or } HR_P) \times v + (LP - U) \times (HR \text{ or } HR_P)}{bw} & \text{if } U \leq LP \\ \frac{U \times (HR \text{ or } HR_P) \times v}{bw} & \text{if } U > LP \end{cases}$$

It is clear that the higher the unit weight, U, the higher the intake is. As weights may vary considerably between units, care must be taken when selecting a value for U.

When data are available on residues in single units and allow for the estimation of the highest residue in a single unit, HR^{unit}, the equations become:

$$IESTI = \begin{cases} \frac{U \times (HR^{unit} \text{ or } HR_P^{unit}) + (LP - U) \times (HR \text{ or } HR_P)}{bw} & \text{if } U \leq LP \\ \frac{U \times (HR^{unit} \text{ or } HR_P^{unit})}{bw} & \text{if } U > LP \end{cases}$$

Case 3

In this case, the number of units is larger than the number of units in a composite sample and the residue level is assumed to be similar to the median of the composite samples from the supervised field trial (e.g. orange juice, tomato soup). When a processed commodity is bulked or blended, the $STMR_P$ value represents the probable highest concentration of residue.

$$IESTI = \frac{LP \times STMR_P}{bw}$$

The deterministic IESTI equations are currently used for pesticide registration as illustrated in the red box in Figure 1.3.

1.3.2.3 Decision on pesticide approval and MRL

Data from toxicological tests on the pesticides are used to derive an 'Acute Reference Dose' (ARfD). The ARfD is the amount of a chemical that can be consumed at one meal or on one day in the practical certainty, on the basis of all known facts, that no harm will result (JMPR, 2002). It provides a measure of exposure that relates to the hazards occurring during short-term exposure and can be obtained from short-term (repeated daily doses for 14-28 days), sub-chronic and reproductive toxicity tests that provide an estimate of the no-observed-adverse-effect-level (NOAEL), a 'safe' dose for a group of experimental animals. The ARfD is obtained by dividing a NOAEL by a safety factor, usually 100, to account for interspecies differences and human variability in sensitivity (Renwick, 2002). This 100-fold safety factor has been attributed to Lehman and Fitzhugh (1954) who stated that 'the chemical additive should not occur in the total human diet in a quantity greater than 1/100 of the amount that is the maximum safe dosage in long-term animal experiments' (Renwick and Lazarus, 1998; Dorne and Renwick, 2005). Lehman and Fitzhugh (1954) emphasised the arbitrariness of the value by stating that 'The 100-fold margin of safety is a good target but not an absolute yardstick as a measure of safety. There are no scientic or mathematical means by which we can arrive at an absolute value. However, this factor of 100 appears to be high enough to reduce the hazard of food additives to a minimum and at the same time low enough to allow some use of chemicals which are necessary in food production or processing'. This statement is still valid today despite several attempts to justify the chosen value (Vermeire et al., 1999).

If the consumer intake is below the ARfD, then the proposed MRL and pesticide use is accepted, assuming that the pesticide does not have detrimental effects on non-target organisms. If not (i.e. calculated intake is higher than the ARfD), the use conditions will have to be modified to reduce the residue levels on the commodity. Examples of modifications include lowering the dose (providing that it will still be effective), extending the period between treatment and harvest and/or applying the pesticide to a different crop altogether. This process is illustrated in the green

boxes in Figure 1.3.

In this section, we have discussed the deterministic IESTI approach for the pesticide registration process, which is currently the most commonly used approach for dietary risk assessment (Paul Hamey, Chemicals Regulation Directorate; personal communication, 21 January 2013). In the next section, we detail alternative probabilistic approaches.

1.3.3 Probabilistic Approaches

In recent years there has been a growing interest in the application of probabilistic techniques to estimate consumer exposure to chemicals in food. In contrast to the deterministic methodology, probabilistic techniques allow the distribution of intakes for multiple individuals in a specified population to be estimated, taking into consideration the variability in food consumption between individuals and the variability in occurrence of residues in food commodities. As in the deterministic IESTI equations, estimating intake from one commodity for a single person on a single day requires the multiplication of the amount of commodity they consumed by the concentration of pesticide it contained, followed by a division by the person's body weight. To assess how often that person's intakes exceed the ARfD, this process can be repeated for every day of the year. If we want to assess what proportion of a population exceeds the ARfD, we need to repeat this calculation for each person in the population. Since this is not possible in practice, dietary exposure models are based on the principle that, if we have a representative sample from the population, we should be able to make inferences about characteristics of the whole population.

For dietary risk assessment, probabilistic approaches infer these characteristics by taking descriptions of the variation in consumption and body weights for multiple people and multiple days and combining them with a description of the variation in residue levels, selected at random. Consumption and body weight data are derived from national dietary surveys and residue concentrations are derived from supervised field trials or monitoring programmes, depending on whether the risk assessment is

part of the registration process or not.
The basic procedure is as follows:

1. Select one 'person-day' record from a dietary survey, comprising consumption and body weight. The consumption and body weight data are sampled together to account for the perceived dependencies between those quantities.

2. Sample a single concentration at random from a distribution describing the variation in pesticide residue levels.

3. Calculate the modelled intake for this person-day by multiplying consumption with concentration and dividing this product by body weight.

4. Repeat steps 1-3 for a large number of person-days, calculating a modelled intake for each.

5. Determine the percentage of modelled intakes for all the person-days that are below the ARfD for the pesticide.

Until EFSA (2012) recently developed guidelines on the use of probabilistic methodology for modelling dietary exposure to pesticide residues, little guidance existed on how probabilistic dietary modelling should be conducted. EFSA (2012) proposes a tiered approach for probabilistic dietary risk assessments and focuses on a 'basic' assessment which may be refined if it results in uncertainty about the risk associated with pesticide exposure. This 'basic' assessment consists of two model runs, a pessimistic model run that is expected to overestimate intake and an optimistic model run that should lead to an underestimate of the intake. The idea is that if the former does not raise any concern for risk managers, the 'true' dietary intake should also not raise concerns. If the optimistic model indicates an unacceptable level of risk, it is considered that refining the model is unlikely to be worthwhile. Various probabilistic dietary risk assessment models have been developed (CREMe, McNamara et al., 2003; MCRA, De Boer and Van der Voet, 2011; Uni-HB, EFSA, 2007b).

Most models used in probabilistic dietary risk assessment include several of the following characteristics:

- **Residue Levels**
  - **Data**: For a proposed new use, typically only supervised field trial data on composites of food items are available. Each composite sample consists of several units of a raw agricultural commodity from a supervised field trial. If the pesticide is already used for other commodities, monitoring data may be available for those commodities. If a product has been approved, monitoring data can be used to either assess the risk associated with a high residue event (i.e. one of the monitored samples has residue levels above the MRL) or for an evaluation of risk associated with pesticide exposure.

    It is important to note that concentration data are often used as actual residue levels, not accounting for measurement errors and reporting/rounding errors. Data below the limit of determination may be modelled using simple replacement rules (e.g. set to LOD, half the LOD or zero) or by more advanced modelling that treats them as latent (censored) values from either a residue level distribution or a mixture distribution, allowing for a proportion of these values to be true zeros.
  - **Choice of Model**: Currently pesticide residue levels may be modelled with empirical or parametric distributions. In the former case, composite residue samples are resampled with replacement. Sometimes a bootstrap approach (Efron, 1979) is applied to account for uncertainty. Bootstrapping involves resampling the data with replacement to generate new 'data sets' of the same size which can be described by empirical distributions. To model the variation in residue levels these empirical distributions are then subsequently sampled with replacement. In the parametric case, a (set of) distribution(s) is fitted to the residue data and samples from this (set of) distribution(s) are drawn to generate estimates of the mean residue level. EFSA (2012) recommends using either an empirical distri-

bution or a Lognormal distribution although more advanced models have been suggested that make use of extreme value theory (Kennedy et al., 2011).

- **Unit variation**: Unit variation can be modelled using two different approaches (EFSA, 2012) depending on the data available:
  - **Sample-based**: This approach comes from interpreting each of the composite samples as the average concentration of a population of a finite number of units (e.g. the potatoes in a bag of potatoes or a bunch of bananas). We can describe the variation in the mean residue levels using an empirical or parametric distribution, F, assuming composite data are representative of the field mean. Once we have generated a new mean $\overline{R}$ from F, the finite number of units, n, implies that there is an upper bound on the unit distribution: the highest possible residue is now equal to $n \times \overline{R}$ (i.e. the case where all of the residue is contained in one unit). EFSA (2012) suggests that in this case a Beta distribution should be used to sample a unit residue value.
  - **Lot-based**: This approach can be thought of as having m composite sample values based on taking n units (e.g. potatoes) from each of the m fields. In contrast to the sample-based approach, this method assumes that there are an infinite number of units in each field. We can again use an empirical or parametric distribution, F, to describe the variation in mean residues. To sample a unit residue level for a unit from a random field, a Lognormal distribution is assumed with the mean value sampled from F and the variance calculated using this mean and a variability factor, representing variation in residue levels between units. The value of the variability factor depends on the type of data. For supervised field trial data, the variability factor is sampled from a Lognormal distribution based on unit field trial data (EFSA, 2005) or fixed at a value of 3 or 6.83 (EFSA, 2007a). For monitoring data the variability factor is sampled from a Lognormal

distribution based on unit monitoring data (EFSA, 2005) or fixed at 6.83, 5 or 1 (EFSA, 2007a).

- **Food Processing**: Residue levels are likely to be affected by various processing steps before the raw agricultural commodity is consumed. Dietary risk assessment models use fixed values of processing factors, defined as the ratio of the concentration in processed and unprocessed food, when processing information is available.

- **Consumption**

  - **Data**: Consumption data are taken from dietary surveys for various age groups and are obtained from a wide range of survey types (see Section 1.3.1.2).

  - **Choice of Model**: Variation in consumption is typically modelled empirically (EFSA, 2012), resampling the observed consumption data as recorded in a dietary survey with replacement, rather than by fitting parametric models to the data. This approach retains potentially complex patterns in the data, in particular correlations between consumption of different foods. However, modelling a variable empirically using the observed data is likely to underestimate the maximum intake. This is because it is unlikely that the survey recorded the most extreme eating event in the population for every commodity. An alternative would be to use parametric approaches, which allow values higher than the highest observed consumption amount, but this would require modelling of dependencies. In order to model dependencies using parametric approaches, many observations are needed. As these are often not available for food types that are consumed rarely this approach may only be reasonable for some food types (e.g. staple foods consumed frequently such as bread or potatoes). One approach to model consumption parametrically is to use a latent Gaussian model (Allcroft2007, Chatterjee2008). Rather than introducing a parameter to account for non-consumption events the model uses an underlying multivariate Gaussian distribution such that the part

of the distribution below a defined threshold corresponds to zero consumption.

- **Unit Weights**: The total amount consumed (in kg food/day) needs to be converted into the number of items consumed so we can account for the effect of unit variation in residue levels on intake.
- **Recipes**: Dietary consumption surveys record data on food items 'as eaten' whereas dietary risk assessment models are based on residue levels on raw agricultural commodities. Therefore, consumption data from surveys need to be converted to (units of) RACs. This conversion consists of two steps: a) identify which ingredients are used and b) for each ingredient, convert the amount (e.g. flour, tomato puree) to a RAC (e.g. wheat, tomatoes) using standard recipes (e.g. a pizza contains 17% wheat and 8% tomatoes, etc.).
- **Body Weight**: Information on body weight comes from the consumption surveys. To account for the dependency of consumption and body weight, both quantities are often sampled together.

- **Model Characteristics**
  - **Population**: Dietary exposure assessments may focus on the whole population or on various subgroups of the population. The latter could refer to only those individuals who consume the commodity in question, vulnerable groups (e.g. children, pregnant women, etc.) or groups that are expected to have higher exposures from other routes (e.g. operators, workers, etc.).
  - **Monte Carlo**: Monte Carlo approaches are often used to obtain population intake distributions by sampling from the consumption and residue level distributions.
  - **Uncertainty**: Typically uncertainty in consumption and residue data is quantified using bootstrap or parametric approaches (EFSA, 2012). Uncertainty for other factors (e.g. processing factors) is generally not quantified with the exception of the variability factor.

- **Model Output**: Probabilistic dietary risk assessment methods will result in an intake distribution. If a probabilistic intake assessment replaced the deterministic IESTI equations, the outcome would be a probability that the ARfD is exceeded (with a confidence or credibility statement).

In this section we have discussed the data and models available for the pesticide registration process. In the next section we will discuss issues with both.

1.4 Discussion of current procedures

In this section we will raise several concerns with regard to the data and methodologies used in current procedures for dietary risk assessment.

1.4.1 Data

1.4.1.1 Residue levels

- **Purpose of data collection**: Data on residue levels in food items comes from either supervised field trials or from monitoring programmes, neither of which are collected for the purpose of dietary risk assessment. The fact that residue data are not generated with dietary risk assessments in mind, leads to the following more specific issues:
  - **Supervised Field Trial - Composite Data**: The most common pre-registration data set consists of a small set of composite data from supervised field trials. These composite samples may provide a conservative estimate of residue levels that consumers are unlikely to be exposed to. The reason for this is that the trials are conducted under cGAP conditions which aim to minimise residue loss, thus leading to higher residue levels than we would expect for RACs available on the market. However, the level of conservatism of supervised trial data is difficult to assess because of various factors that may make residue levels in food as consumed by the general population higher or lower (e.g. farmers may not comply with GAP procedures, local conditions may be different than those in

the trials, not every unit on the market is treated, longer time between harvest and consumption may lead to lower residue levels due to degradation processes, etc.). A further issue is that supervised field trial data are collected at the composite level so the data do not provide information on residue levels for food items that may be consumed as individual units (e.g. apples).

- **(Supervised) Field Trial - Unit Data**: Unit data from supervised field trials are relatively scarce. However, even if they were available, they would suffer from the same conservatism issues as the composite data from supervised field trials. Some unit data, which are useful to describe the variation in residue levels between units, are available from field trials (Ambrus, 2006). Field trials are different to supervised field trials in that they are conducted under normal agricultural practice with two deviations. The first is that they are designed to facilitate the detection of residue levels. As a consequence, field trials might either be conducted at higher application rates than normal or use a shorter time between application and harvesting. The second deviation is that pesticides are often applied in mixtures, so-called tank mixes, to assess whether variation in residue levels is pesticide-specific.

  Field trial data sets have been used to estimate variability factors (EFSA, 2005). However, unit data collected from (supervised) field trials under controlled conditions, may not include as many sources of variation as residue levels observed in units obtained from real applications under a variety of weather conditions, application equipment, local practices, etc. Therefore, variability factors calculated using (supervised) field trial data may underestimate the true variation in residue levels on units.
  Unit data cannot be used directly in dietary risk assessment as they do not include between-field variation. However, they would be useful if information about between-field variation could be obtained from other sources. For example, in principle we could use composite data from

supervised field trials to describe the variation between fields as long as we account for the fact that composite samples, based on very few units, only provide an estimate of the field mean.

- **Monitoring - Composite Data**: Monitoring programmes generally result in composite data, which would provide a more realistic residue level estimate than those obtained from supervised field trials if they had been sampled at random from food items available to consumers. However, monitoring programmes tend to be a mixture of surveillance sampling, in which samples are collected at random and enforcement sampling, in which samples are taken based on suspicions about the safety or non-compliance with the legal limits of a product and/or as a follow-up of violations found previously (EFSA, 2011). Samples taken as part of the EU coordinated programme are considered to be surveillance samples whereas enforcement samples are taken as part of national programmes (EFSA, 2011). Existing residue level databases do not distinguish data obtained from targeted sampling from those obtained from random sampling. As a result, unless the data obtained from monitoring programmes are labelled as being obtained using a random sampling approach, they should not be regarded as a random sample of pesticide residue levels as experienced by consumers. However, guidelines for probabilistic dietary risk assessment currently assume that monitoring data are a random sample (EFSA, 2012).

  Another issue with the collection of monitoring data is that the proportion of samples obtained from various sources as part of monitoring programmes may not reflect availability to consumers. For example, in 2008, 29 out of 48 cherry samples taken in the UK originated from Spain (PRC, 2008; PRC, 2009), whereas it is unknown what proportion of cherries consumed by the UK population are of Spanish origin. In addition, monitoring samples are generally taken from retail outlets to mimic the selection of food by consumers. This may not be representative for the

residue levels on RACs that are used in processed food, e.g. tomatoes in pizza, as tomatoes used in pizzas may come from a different source or be subject to different treatments (pesticide application, storage time and conditions, etc.) than tomatoes sold on the shelf. As little is known about the origins and/or treatment history of units within a composite, i.e. whether they originate from the same field or multiple fields, composite residue data should be treated with care when inferring residue level distributions.

- **Monitoring - Unit Data**: Unit data are rarely collected as part of regulatory monitoring programmes. In one publicly available study (Hill and Reynolds, 2002), units were only measured if positive residue levels were found in a composite sample. Therefore unit data obtained from this study were a biased sample from the residue level distribution. Consequently we cannot use the data obtained from this study as if they were representative of food items that are available to consumers. When we have unit data, we cannot always infer whether the variation observed in unit residue levels is caused by a proportion of untreated units in the sample or whether the variation is caused by variation in application factors, crop and environmental factors and/or dissipation factors. As a consequence, unit data from monitoring programmes should be treated with care and may only be suitable for estimating variability factors. However, as they are a biased sample from the upper tail of the residue distribution, they are likely to underestimate the true variability in residue levels. In addition, if the proportion of non-treated units is very different to the proportion of non-treated food items considered in the dietary risk assessment, the variability factor may not provide a good estimate of unit variation.

Given that the currently collected residue level data cannot easily be used to model residue levels on food items, it would be sensible to reconsider what data should be collected for use in the pesticide registration process. If for a new use, supervised field trials were to be conducted in such a way that

unit data were collected from multiple fields, whilst recording which data was obtained from which field, these data could be used to model within-field and between-field variation. The residue level estimates, obtained from these data, would still be conservative as they would not account for untreated food items and the trials are conducted according to cGAP. However, this would be an improvement on current practice.

If these data were available, surrogate residue data sets, such as the field trial data used to derive variability factors (EFSA, 2005), would not have to be used to model unit variation. However, if the principles for data collection do not change, we need to make sure that the data that are available are treated appropriately in dietary risk assessments.

- **Residue level variation in composite samples**: Monitoring programmes provide estimates of residue levels in composite samples. Combining treated and untreated units of a commodity will lead to a reduction of residue levels: if a sample consists of twelve apples, three with a residue level at twice the MRL of chemical A and nine untreated, this will result in residue levels of half the MRL of A in the composite sample, indicating that there is no reason for regulatory action, despite some of the units having residue levels of twice the MRL on them.

- **Dealing with censored data**: A common issue with residue data, particularly those obtained in monitoring programmes, is that many samples will contain residue levels that are not quantifiable and are reported as less than the limit of determination (LOD). The LOD is the lowest concentration at which quantitative results can be reported with a high degree of confidence. It is important to realise that <LOD values are only reflecting our technical abilities to measure residues. Even unquantifiable concentration levels may lead to adverse effects and therefore it is important to deal with <LOD data appropriately. An often proposed solution to deal with <LOD data is to replace them by $k \times LOD$ where $k \in \{0, 1/2, 1\}$ (EFSA, 2012; OECD, 2011b). However, this does not take into account the distribution shape of the under-

lying population and may in fact violate distributional assumptions that are often made when conducting probabilistic dietary risk assessments.

Some probabilistic approaches are more suitable for dealing with <LOD data than others. For example, empirical methods based on resampling the data with replacement cannot offer an alternative method to replacing <LOD values with another value, whereas assuming a distribution allows samples to be imputed for values below the LOD. When modelling monitoring data it may be appropriate to replace <LOD with a zero if information is available on the proportion of untreated food items available on the market. In this case, a <LOD result may indicate that either no residues were present in the sample or the pesticide was present but concentrations were too low to quantify. Paulo et al. (2005) introduced a mixture model approach which specifically addresses this case.

EFSA (2010b) explored various statistical approaches for fitting distributions to left-censored data sets. Their conclusion was that when there are >25 censored values in data sets consisting of <50 samples, or when more than 80% of the data are censored, no probabilistic assessment should be conducted. However, they did not consider Bayesian approaches which can deal with high levels of censoring and account for the uncertainty caused by the censored data. An analysis of UK monitoring data sets (PRC, 2010; PRC, 2011a; PRC, 2011b; PRC, 2011c) that are not completely censored showed that, on average, 93% of values were reported as below the LOD as shown in Figure 1.4. EFSA (2010b) recommends that when data consists of >80% censored data, similar food categories should be pooled together or more data should be collected. As collecting more data is not likely to increase the proportion of positive samples, EFSA (2012) suggest that <LOD data should be replaced by 0 or by the LOD, stating that the latter is conservative. However, this ignores the fact that even though replacing the values with the LOD will increase the mean it also reduces the variance, making it unclear what the overall effect on the residue distribution is.

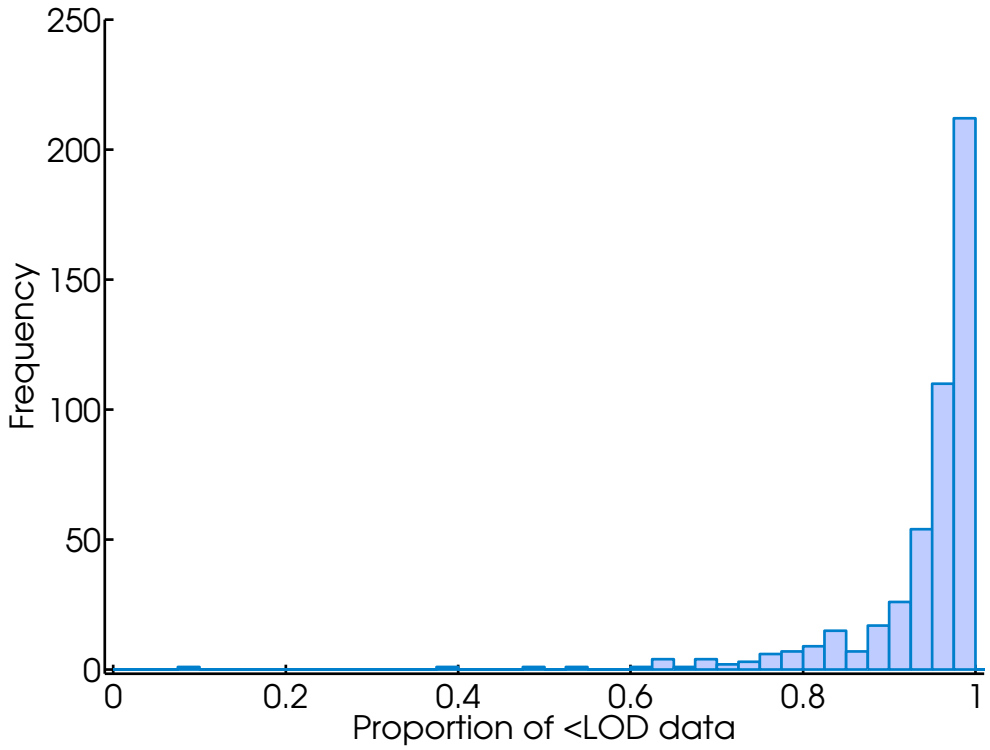


Figure 1.4 *– Proportion of data <LOD in data collected as part of the 2010 UK residue monitoring programme. Chemicals for which 100% of the data were below the LOD were excluded as in those cases the pesticide may not be registered for use on that crop.*

- **Reporting error**: Residue level data are often reported after rounding (either to n_s significant figures or n_d decimal places). As a result, many of the values in a data set may be repeated, which might suggest that the population distribution is discrete. If the rounding method applied to the data is known, Bayesian methods can be used to account for the uncertainty introduced by rounding (see Chapter 4 for details).

- **Measurement error**: The effect of measurement error on estimating residue levels is often ignored, perhaps due to the laboratory process conforming to the relevant international standards. Kennedy and Hart (2009) provide a general approach that allows for the integral modelling of measurement uncertainty in dietary risk assessments. Their analysis indicates, however, that the effect of

measurement error may be significantly smaller than the uncertainty caused by the limited number of composite data. As a result, EFSA (2012) considers it unnecessary to take measurement error into account.

- **Sample Size**: Sample sizes for both types of data (supervised field trials and monitoring programmes) tend to be very small compared with the number of food items consumed. Figure 1.5 shows the number of trials conducted for a large set of pesticides obtained from EFSA draft risk assessment reports. The most common number of field trials is 8 and the median number of field trials is 10. When very few trials are conducted, it is essential to quantify the uncertainty resulting from the small number of data in an appropriate manner, e.g. using Bayesian approaches.

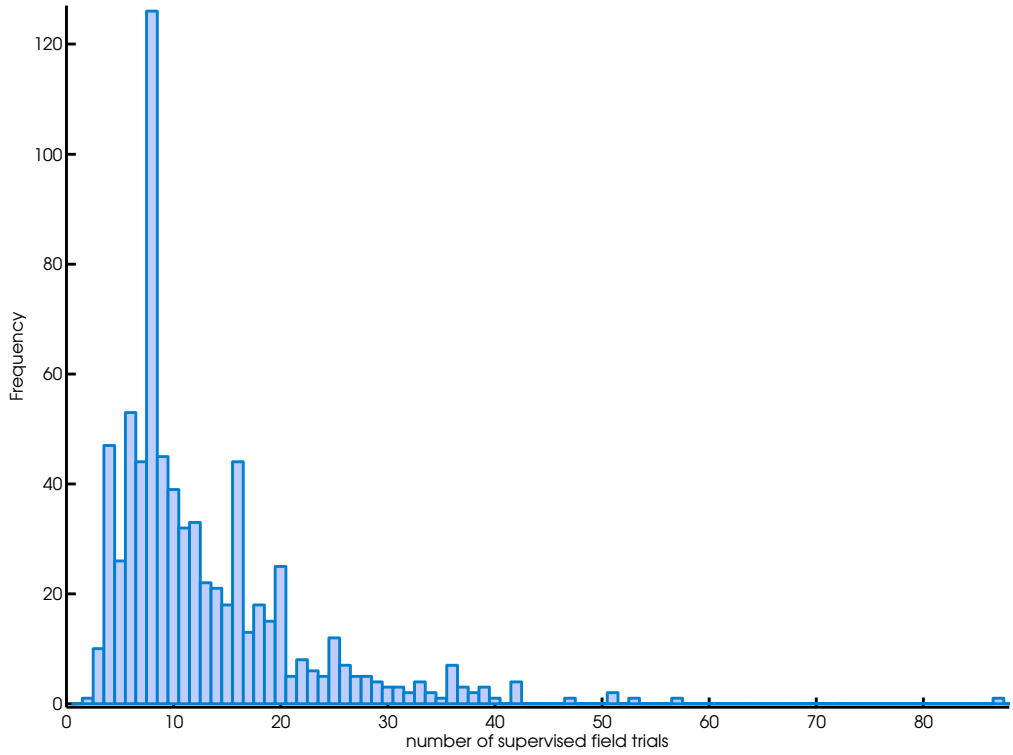


Figure 1.5 – *Overview of the number of field trials conducted for 730 pesticides and reported in draft risk assessment reports.*

With regards to monitoring data, 416 active substances (SANCO, 2012) are approved for use in the EU which may be applied to up to 383 food commodities (WHO, 2012). In addition to these approved pesticides, monitoring

programmes will also have to focus on pesticides that are not approved to assess whether they have been used illegally. In 2009, 10,553 composite samples were analysed in an EU-coordinated programme focusing on 138 pesticides and 10 different food commodities (EFSA, 2011). In addition, 67,978 samples were analysed as part of national monitoring programmes, focusing on 834 distinct pesticides in 300 different food commodities. Apart from the fact that various commodities have not been monitored at all, this shows that sample sizes in monitoring studies are small. However, one could argue that food items come from a finite number of sources and that one may be prepared to make the assumption that all products from one source are likely to have received the same treatment. If so, one could in theory obtain a reliable estimate of pesticide residue levels of food products on the market from a small sample provided that it was representative of pesticide residue levels on all food products.

1.4.1.2 Consumption data

- **Age of surveys**: Dietary surveys provide a snapshot of people's diets for a specific period of time. It is questionable how relevant historical dietary records are for current risk assessments as available products and dietary habits change over time (e.g. consumption of bottled water and ready meals has increased in recent years).

- **Sample Size**: Food consumption surveys are expensive and time-consuming as they may require face-to-face contact (interviews, physical measurements), analyses of food samples and analyses of dietary records. As a result sample sizes are kept low, particularly when considering seasonality in consumption and variation in consumption patterns in the population. Therefore, they may not capture the extensive variation in consumption patterns between individuals and sub-groups (e.g. based on age). A small stratified sample may be sufficient because dietary surveys are designed to be representative of the population of interest. However, it is difficult to assess how representative they are because not every individual selected for consumption surveys will take part and non-respondents will not always be replaced (Hoare et al., 2004).

- **Bias**: Food surveys are often run with volunteers and even though individuals are selected to create a representative sample of the population, high levels of non-response increase the potential for bias. In the 2002 UK National Diet and Nutrition Survey (Hoare et al., 2004), only 47% of the selected individuals completed a full 7-day dietary record. Assessing and dealing with bias is particularly difficult when there is little or no information on subgroups within the study population.

  Another type of bias is caused by the fact that people's behaviour may be affected by their involvement in surveys. They may be reluctant to record sensitive or taboo subjects and therefore either decide not to record them or they may change their behaviour. A simple example of this in dietary records is that people may record lower consumption amounts for foods that are considered to be socially unacceptable.

- **Minor Foods**: Food surveys are only able to reliably record the consumption of food types that form a major part of our diet (i.e. staple foods). They tend to underestimate the consumption of minor food items or food items that are only consumed on a seasonal basis. EFSA (2012) suggest that if a consumption survey does not contain records of a rarely eaten food, the consumption amounts could be estimated from consumption data of related food types that may have been recorded in the survey.

1.4.2 Modelling

- **IESTI Equations**: Although the IESTI equations are simplistic, if the results are interpreted appropriately, they may be useful to manage the dietary risk associated with pesticide intake. However, when interpreting the outcome of an IESTI-based risk assessment, the following should be considered:

  - **Conservatism**: The deterministic IESTI equation does not provide an indication of how conservative it is. It is expected that combining conservative estimates for residue levels and consumer consumption will result in a conservative intake estimate. EFSA (2007a) explored what level of

conservatism was obtained by using the IESTI equation. The level of conservatism was assessed in terms of the proportion of the EU population that would be exposed to a dose not exceeding the ARfD. This was estimated in case studies for 13 pesticides, 8 countries and various subgroups (based on age) of the population using a range of probabilistic dietary exposure models. In the case studies, the IESTI results were compared with the output of the probabilistic models and residue levels from monitoring studies. The comparison indicated that the level of protection, i.e. the number of person-days with intakes below the ARfD, was at least 99% and above 99.9% for most probabilistic models for the total population. However, as we do not know what level of protection is achieved by the probabilistic methods used in the study, we do not know what the true level of protection is.

- **IESTI assumes exposure to a single chemical**: The IESTI equations assume that consumers will only be exposed to a pesticide on one commodity at a time. This relies on the assumptions that multiple commodities have not been treated by the same pesticide, that multiple commodities would not all have high residue levels at the same time and that the consumption of large portions of these commodities in a short period of time is unlikely.

  A brief analysis of residue level data from the UK monitoring programme shows that several pesticides have detectable residue levels on multiple crops (PRC, 2010; PRC, 2011a; PRC, 2011b; PRC, 2011c). For example, imidacloprid was detected on thirteen out of twenty crops (including broccoli, cabbage, cherries, grapes, lettuce, nectarines and peaches). In fact, out of 134 pesticides analysed, 85 were detected in two or more crops. This suggests that the first assumption conflicts with the available residue data. Given that the same pesticide appears to be used on multiple crops, it is possible that consumers are exposed to a particular pesticide multiple times by eating portions of different commodities.

- **Estimation of LP**: One issue with the estimation of LP is that it appears to be generally accepted to use the 97.5^{th} percentile commodity consumption value (JMPR, 2002). However, for a commodity for which fewer than 40 consumption days have been recorded in dietary surveys, it is not possible to calculate the 97.5^{th} percentile without extrapolation and/or making distributional assumptions. Van der Velde-Koerts et al. (2011) state that no guidelines are available about how many consumers are needed per commodity to get an accurate value for the LP. They suggest using the 95^{th} percentile commodity consumption value if there are between 20 and 40 consumption days recorded. If less then 20 consumption days are available, the 90^{th} percentile is suggested and if less than 10 consumption days are reported, the maximum value is suggested. This practice does not appear to be conservative as it is unlikely that with sample sizes as small as these, the full spectrum of consumption behaviour will have been captured by dietary surveys.

- **Use of IESTI for MRL setting**: One feature of MRL setting using the IESTI equations is that an MRL may be higher than the effect threshold (e.g. ARfD). The OECD MRL calculator suggests that the MRL should be proposed as the maximum of the highest residue, the mean + 4 × standard deviation and 3 × mean $\times(1 - \frac{2F}{3})$, where the residue data are obtained from supervised field trials and F is the fraction of censored data. Therefore, the proposed MRL may be greater than the highest residue, which is used in the IESTI equation. If the IESTI equation results in an intake just below the effect threshold, then the MRL will be accepted and the pesticide will be approved. If the residue level on commodities on the market are below the MRL but above the highest residue observed in the supervised field trial, consumers may be exposed above the effect threshold. However, as the value is below the MRL there is no legal issue of non-compliance, demonstrating that MRLs are trading standards.

- **Recipe conversion factors**: Consumption surveys record food as eaten by individuals which includes processed food whereas for dietary risk assessment we are interested in which raw agricultural commodities (RACs) individuals have consumed. Therefore recipe databases have to be used to make assumptions about the proportions of RACs used in processed food. As recipe information for food products is commercially sensitive and will vary between products and possibly even batches of products, (manufacturers may introduce a 'new recipe' for various reasons, e.g. reducing sugar or salt levels), these conversions are often based on simple models and standard conversion factors for commodities. Converting ingredients to RAC amounts can be complex due to various types of processing and little is known about the impact of these conversions on the overall exposure estimate.

  Another issue is that some food items can be either bought as processed food or prepared at home, e.g. apple pie. Therefore it is important to consider whether the ingredients came from different sources or whether the ingredients were bought in a single purchase. In the latter case, residue levels for multiple food items may be similar and may be the driving force behind extreme exposure events, e.g. an individual who buys apples with higher than average residue levels and consumes some of them as whole apples, some of them in a homemade apple pie and some of them as homemade apple juice.

- **Distribution choice**: EFSA (2012) recommends either using an empirical distribution or a Lognormal distribution to model composite residue data. Given the small sample size of composite residue data, empirical distributions are unlikely to be a viable approach for describing the variation in residue levels. EFSA justified the suggestion of a Lognormal distribution by referring to an analysis by Boon et al. (2003) on 10 data sets (consisting of 5-66 composite samples) for which a Lognormal distribution could not be rejected. Given the type of study, the small number of data sets and small number of samples per data set, it is questionable whether this provides sufficient evidence to recommend a Lognormal distribution in general. EFSA (2012) discussed other approaches to model unit variation in residue levels because the Lognor-

mal distribution, often used in combination with variability factors (VF), was found to be inappropriate for multiple unit data sets. EFSA concluded that further simulations are needed to assess how residue distributions may best be represented. Whether this can be done with parametric distributions or whether non-parametric approaches are necessary is one area of research that will be addressed as part of this thesis.

EFSA (2012) proposes the use of empirical distributions for consumption data. Given that consumption data sample sizes are much larger than residue data sample sizes, this may be a reasonable choice provided that the consumption data are representative of the population of interest. However, it still may not provide a reasonable estimate of some individual's extreme consumption habits.

- **Bootstrap approaches**: Some existing dietary risk assessment models, e.g. MCRA (De Boer and Van der Voet, 2011), rely on bootstrap techniques to quantify uncertainty resulting from small data sets and empirical distributions to describe the variation in consumption and residue levels. Bootstrap techniques can be useful and their simplicity has made them a popular choice in dietary risk assessment. However, it is important to know their limitations and to avoid using them inappropriately. The main idea in statistical inference is that a sample is used to learn about a population's characteristics. This can be done with a wide range of inference techniques, the bootstrap being one of them. However, certain population characteristics may be poorly estimated by bootstrap approaches, particularly for small to medium sample sizes. Corrections may be applied to counter the estimation bias introduced by small sample sizes, but they require additional assumptions and computations.

  In dietary risk assessments, bootstrapping is used to describe uncertainty about the consumption and residue level distributions. For consumption distributions, the number of data may be sufficient to estimate certain parameters (e.g. mean, median, non-extreme percentiles) of the population consumption distribution. However, as it is unclear whether the risks associated with dietary intake are caused by extreme consumers, extreme residue levels or a

combination of both, distributional approaches need to be able to provide realistic estimates of extreme consumer intakes. With an EU population size exceeding 500 million, the number of data from surveys is relatively small and unlikely to be representative of the whole population so it is doubtful that bootstrapping approaches will be able to predict extreme intakes.

More worrying is the fact that bootstrap approaches are also used to model residue levels. Figure 1.6 shows the results from simulation studies where $n =$ 2, 4, 8 and 100 samples were taken from a $\mathcal{N}(0, 1)$ distribution.

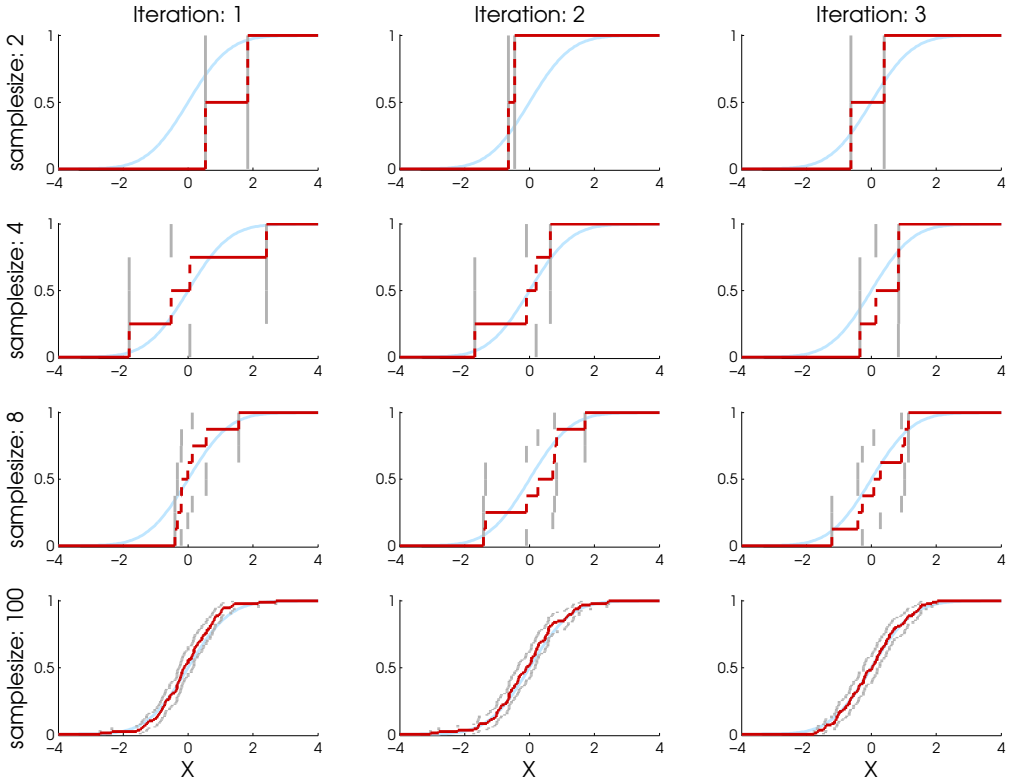


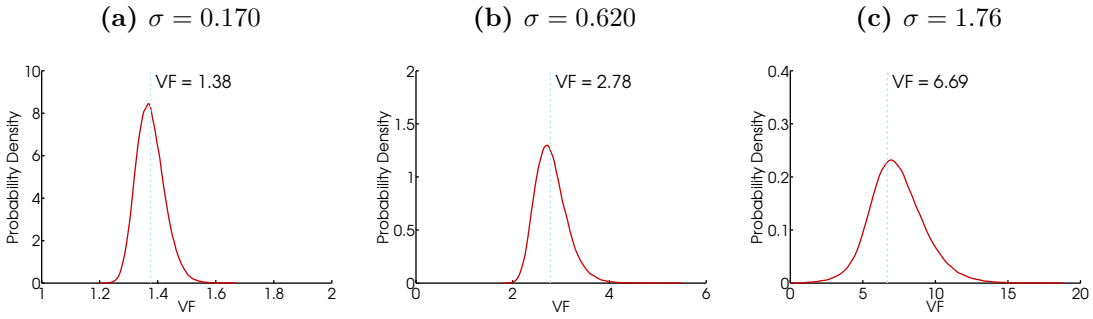
Figure 1.6 – *Results of bootstrapping samples X of size $n = 2$, 4, 8 and 100 from a standard Normal distribution. The blue line indicates the target $\mathcal{N}(0, 1)$ distribution, the red line indicates the median distribution, obtained from generating 10,000 bootstrap samples. The grey dashed lines indicate the 95% confidence interval. For each value of n, the simulation was repeated 3 times to demonstrate the impact of the original n values on the performance.*

These sample sizes were chosen to represent typical sample sizes of residue data sets for which bootstrap approaches are currently applied in dietary risk assessment. The samples were then bootstrapped 10,000 times and the median estimates are displayed as cumulative distribution functions together with a 95% confidence interval and the target distribution. The exercise was repeated 3 times for each sample size. It is obvious that for small sample sizes, the distribution produced by bootstrapping provides a very poor estimate of the underlying distribution. For a sample size of 100, one could argue that the bootstrap distribution starts to reflect the target distribution. However, it is still far from perfect and when we look at the tails of the distribution, it is clear that the population distribution is not represented very well. If we believe that dietary risk is caused by either extreme consumption amounts, extreme residue levels or a combination of both, more advanced techniques are needed to describe the variation in both.

- **Derivation of Variability Factors**: Variability factors (VFs) were derived using various unit residue data sets (EFSA, 2005). The VF is defined as the 97.5^{th} percentile of the distribution of unit residues divided by the mean residue level (EFSA, 2005). Sample sizes used to determine VFs are relatively small to estimate the 97.5^{th} percentile (e.g. the median number of unit values in unit residue studies is around 120) and therefore may lead to a poor estimate of the VF. In the following simulation study, we show that VFs estimated from a typical field trial sample size provide a poor estimate of the true VF. We generated 120 log unit residue levels from a $\mathcal{N}(\log(100), \sigma^2)$ distribution with $\sigma = 0.170$, 0.620 and 1.76, corresponding to the 2.5^{th}, 50^{th} and 97.5^{th} percentiles of the standard deviation of the unit field trial data discussed in Chapter 4. Then we estimated the ratio of the 97.5^{th} percentile and the mean of the back-transformed samples, i.e. the VF. We repeated this 10,000 times to obtain a distribution of VF estimates. Figure 1.7 shows a kernel density plot (Silverman, 1981) of the VF estimates (red line), together with the true VF (blue dashed line) for each value of σ. It is clear that using a small sample to estimate the VF results in some uncertainty around the VF estimate. Ideally,

approaches that make use of VFs should account for this uncertainty.

Figure 1.7 *– Kernel density plots of simulated variability factors. VF distributions (red lines) were obtained by simulating unit residue data for various values of σ. The true VF is indicated by the blue dashed line.*



Another important question is whether the population VF varies between crops/pesticides. Hamilton et al. (2004) assumed that the observed variation in estimated VFs is the result of sampling error (i.e. uncertainty about the VF as a result of small sample sizes), whereas EFSA (2005) showed that the population VF varies between data sets. The models describing unit variation in residue levels, presented in Chapters 4 and 5, are based on this latter observation.

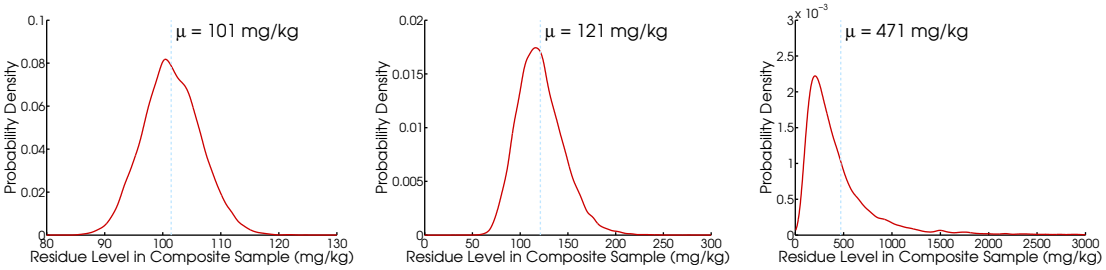
- **Treatment of composite supervised field trial data as field means**: Current probabilistic models assume that the composite values obtained from supervised field trials are field means and that fitting a distribution to the composite samples from multiple fields can be regarded as the field mean distribution. A composite sample obtained in a supervised field trial consists of a small number of units (e.g. 12 for apples) and may therefore provide a poor estimate of the field mean residue level. Figure 1.8 shows kernel density plots of simulated composite samples obtained from repeatedly sampling 12 log unit residue values from a $\mathcal{N}(\log(100), \sigma^2)$ distribution, where $\sigma = 0.170$, 0.620 and 1.76, as before. The blue dashed lines indicate the true mean of the unit distribution and the red kernel density line shows the distribution of simulated composite samples. It is clear that the resulting composite samples

are poor estimates of the true field mean because, for example, for $\sigma = 0.620$ composite samples ranged from 50 to 250 mg/kg whereas the true mean is 121 mg/kg.

The variation in composite samples is a result of the variation in residue levels on individual units and the small number of units making up a composite sample. We should account for both these factors when aiming to obtain a field mean distribution.

Figure 1.8 – *Kernel density plots of simulated composite samples. The blue dashed lines represent the true means of the unit distributions which were used to simulate the composite samples. The red lines indicate the distribution of simulated composite samples.*

(a) $\sigma = 0.170$ **(b)** $\sigma = 0.620$ **(c)** $\sigma = 1.76$



Composite samples are currently used together with variability factors (VFs) to simulate unit residue levels. However, the definiton of the VF implies that it should be applied to an estimate of the field mean. When inferring the field mean distribution from composite samples we first need to 'remove' the unit variation component in the observed variation in composite samples to obtain a distribution which only describes the variation in field mean residue levels. Therefore in this thesis we present an approach which allows us to estimate residue levels on units in this way (see Chapter 5).

- **Mixtures of pesticides**: Up until recently, dietary risk assessments focused on a single pesticide at a time. However, with increasing numbers of pesticides being approved, concerns have been raised about exposure to multiple pesticides (EFSA, 2009; Van Klaveren et al., 2009). This can occur because

consumers are exposed via a range of commodities, because commodities may have been treated with a range of pesticides or because commodities which have received different treatments are mixed before they arrive on the market. Monitoring programmes already show that some samples contain residues from multiple pesticides (PRC, 2011a; PRC, 2011b). As monitoring programmes focus on composite samples, it is unclear whether a composite sample consists of units that have received different treatments or whether units have been treated with various pesticides. Pesticide Usage Survey (PUS) data (Fera, 2011) indicate that the latter is not uncommon, so an important question is how we should assess the risk associated with exposure to multiple pesticides. Apart from the fact that no legal framework exists for dealing with these cumulative risk assessments, current dietary risk models have only been developed to deal with a single pesticide at a time. An important consideration is how to model dependencies between residue levels when a commodity has been treated with multiple pesticides.

- **Interpretation of model**: A problem with current acute probabilistic dietary risk assessments is that they are based on the concept of 'person-days' as a result of the way dietary surveys are treated in existing models. Current models tend to resample person-days and consequently the output of the assessment represents variation between person-days, not individuals. Model outcomes indicating that 1% of 'person-days' are above the effect threshold level could refer to every individual of the population experiencing an intake above the threshold for 3.65 days annually (on average), refer to one individual in a hundred experiencing an intake above the threshold for every day of the year or, more likely, somewhere in between. It would be much better if consumer intakes were modelled on an individual basis, so that we could estimate for each individual the probability of exceeding the threshold value and ideally by how much on a daily basis. It should then be possible to extrapolate this to the population level, e.g. x% of the population will have a y% probability of being exposed to a dose equal to n times the threshold level. This would allow risk-managers to interpret the outcome of dietary risk assessments more

easily. The downside of this would be that risk managers would have to derive a set of politically sensitive acceptability criteria, e.g. a product is considered to be safe if up to 0.01% of the population is exposed to the ARfD level on one day of the year unless a) 0.001% of the population is exposed to more than three times the ARfD on one day or b)..., etc. In absence of these criteria, one could try to assess what level of protection the current regulations offer in an exercise similar to the EFSA (2007a) study into what level of conservatism was obtained by using the IESTI equation.

- **Summary of output**: With regards to model outputs, the focus is currently on the likelihood of exceeding a toxicological threshold, e.g. proportion of the population exceeding the ARfD (EFSA, 2012). Although this is an important statistic, it would be more informative if this were accompanied by an assessment of how extreme exceedances are. This information can be obtained from some current dietary risk assessment models as illustrated in EFSA (2007b).

- **Validation of dietary risk assessment models**: Another issue with the use of probabilistic modelling in dietary risk assessment is that ideally the models should be validated before being used in a regulatory context. However, dietary risk models cannot be validated as a whole because it would require knowledge of the population exposure distribution for a wide range of scenarios. Gibney and Van der Voet (2003) suggested that a probabilistic dietary exposure model is fit for purpose when (a) the modelled exposure does not underestimate the true exposure and (b) the modelled exposure was lower than the IESTI approach. However, there is no reason why the IESTI results should provide an upper bound on exposure. In fact EFSA (2007a) already showed that the IESTI equation only corresponds to high percentiles of exposure estimates from various probabilistic models. A more thorough validation exercise could consist of validating various parts of the exposure models. If we know that the consumption model and the residue level models are valid, then we can assume that a model consisting of both model parts is validated as well, assuming that dependencies are dealt with appropriately. Validating sub-

models has the added advantage that if more advanced, validated approaches become available for parts of the model, they can be added without having to revalidate the whole model. Given that, even for sub-models, we often do not know what the truth is, the only validation options available are either using synthetic data or using a large number of case studies. In the latter case, the model's output can be 'validated' if the model output seems reasonable according to some pre-defined criteria.

Currently available probabilistic dietary exposure models have attracted criticism for resulting in unrealistic exposure estimates (EFSA, 2012). One reason for this may be that little information is available to estimate the parameters for consumption and residue level distributions. When dietary risk models have tried to account for this lack of information, the population intake distributions became very uncertain and high intake levels, that were labelled as unrealistic, were observed (EFSA, 2012). However, it may not always be clear when an exposure estimate is unrealistic and when it is a realistic extreme case.

The issues presented in this section could have a significant impact on the outcome of the risk estimate and should therefore be considered carefully by risk assessors. However, the issues not related to the modelling of pesticide residues are considered to be outside the scope of this thesis and thus are not discussed further.

1.5 Motivation for Thesis

This thesis will focus on the development of novel approaches for describing the variation and uncertainty in pesticide residues on raw agricultural products. The main reason for selecting this particular area of the overall pesticide registration process is that current methods, both deterministic and probabilistic approaches, for acute human dietary risk assessment are based on very basic models for residue levels in food items which neither reflect the data well nor provide an adequate quantification of uncertainty.

Deterministic risk assessment is by far the most commonly used risk assessment approach but the use of probabilistic risk assessment approaches is likely to increase now that guidance documents (EFSA, 2012) and software tools (CREMe, McNamara et al., 2003; MCRA, De Boer and Van der Voet, 2011) are available. However, the implementation of probabilistic approaches for dietary risk modelling is still in its infancy and many issues have not been dealt with appropriately. As deterministic risk assessments are also based on probabilistic elements, e.g. percentiles of the consumption and residue level distributions, there is a need for robust approaches that can be used to model the available data in a more appropriate way. Currently, most probabilistic risk assessments are based on strong distributional assumptions, e.g. Lognormal and other parametric distributions to account for the variability in residue levels. For most risk assessments there are relatively few data available and so there is little evidence to support these distributional choices for individual crop/pesticide populations. More importantly, the current approaches for dietary risk assessment do not make best use of the available data as each field trial data set is analysed separately. In other words, when conducting a risk assessment, the analysis is not making use of available information from previous analyses of pesticide residue levels on food items.

Another issue is that current probabilistic models are based on poor modelling choices which fail to account for the lack of unit data. For example, the definition of variability factors can only be justified if they are applied to estimates of the field mean and take into account appropriate distribution shapes for units. As neither is the case, more advanced approaches are needed to account for unit variation.

One aim of this thesis is to solve some of the issues mentioned in this chapter. However, obtaining estimates of dietary risk is very challenging given financial and practical constraints on data collection practices. These constraints affect the estimation of the variability in consumption amounts and residue levels and thus intake amounts and emphasise the importance of quantifying uncertainties associated with these quantities.

1.6 Overview of Thesis

This thesis aims to improve the way in which residue data are modelled in dietary risk assessments, although some of the methods will be applicable to consumption data as well. The approaches presented in this thesis aim to obtain better estimates of the variation in residue data and can be used to improve both deterministic and probabilistic risk assessment approaches. With regards to deterministic assessments, the distributions can be used to obtain better estimates of the conservative parameters that are used in routine deterministic risk assessments. Chapter 2 provides an introduction to the mathematical concepts which are used extensively throughout the thesis. Chapter 3 introduces a novel approach to modelling the correlation in residue levels of multiple pesticides which makes use of monitoring data and pesticide usage information. Chapters 3 and 4 introduce new approaches to model pesticide residues on raw agricultural products for the registration process. Chapter 4 introduces a model that can be used to describe unit variation in residue levels and accounts for censored data and reporting errors. Unlike current models describing the variation in residue levels, this model aims to learn the distribution shape from data. Chapter 5 presents an approach to model within-field and between-field variation of residue levels in a way that does not overestimate the variation in supervised field trial data and accounts for uncertainty. Finally, Chapter 6 provides an overview of all the new approaches and identifies further research needs.

Chapter 2

Bayesian approaches for Dirichlet Process Mixture Models

This chapter aims to introduce several concepts that are used extensively throughout this thesis. The approaches introduced in Chapters 3, 4 and 5 rely heavily on Bayesian techniques, Markov Chain Monte Carlo (MCMC) and either the Dirichlet distribution or Dirichlet Process Mixture models. This chapter will provide an overview of these concepts.

2.1 Bayesian Inference

Bayesian inference is based on Bayes' rule to express our uncertainty about a parameter of interest θ given some form of evidence. Bayes' rule is based on updating our beliefs, expressed in a prior distribution $p(\theta)$, with data, y, using the likelihood function $p(y|\theta)$:

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)}$$

Often prior distributions are selected from standard distribution families to facilitate calculations in Bayesian inference. Conjugate prior distributions are a common choice as they result in a posterior distribution that is from the same family as the prior distribution. This is particularly helpful if the family is easy to characterise. If it is not, we need to use numerical approaches such as Monte Carlo methods.

2.1.1 Monte Carlo methods

If the posterior distribution, $p(\theta)$, is not in a form from which we can calculate summary statistics of interest (e.g mean, credible intervals, etc.) we may instead have to use numerical approaches to draw samples from the distribution, a process often referred to as Monte Carlo simulation (Metropolis and Ulam, 1949; Von Neumann, 1951). We can then calculate summary statistics from these samples to characterise the distribution. For example, if we are interested in the mean of the distribution, we can calculate this using N samples, $\mathbf{x}$, and the fact that $\mathbb{E}[\theta] \approx \frac{\sum_{i=1}^{N} x_i}{N}$. In a Bayesian context, Monte Carlo simulation can be helpful if we are interested in the joint posterior distribution of multiple variables. To illustrate this, let us consider the posterior distribution of the mean, μ, and standard deviation, σ, of a $\mathcal{N}(\mu, \sigma^2)$ distribution. Assuming a prior distribution $\pi(\mu, \sigma) \propto \frac{1}{\sigma}$, the joint posterior is given by:

$$p(\mu, \sigma|\mathbf{y}) \propto \sigma^{-(n+1)} \exp\left[-\frac{(n-1)s^2 + n(\mu - m)^2}{2\sigma^2}\right]$$

where s is the standard deviation of the data vector (denoted by a bold typeface throughout this thesis), $\mathbf{y}$, m is the mean and n is the sample size (Box and Tiao, 1973). If we cannot easily obtain a summary statistic of interest from this distribution we can factorise it and generate samples using the marginal distribution, $p(\sigma|\mathbf{y})$, and the conditional distribution, $p(\mu|\sigma, \mathbf{y})$:

$$\sigma|\mathbf{y} \sim \frac{\sqrt{n-1}s}{\chi^2_{n-1}}$$
$$\mu|\sigma, \mathbf{y} \sim \mathcal{N}\left(m, \frac{\sigma^2}{n}\right)$$

where χ^2_ν is a Chi-squared distribution with ν degrees of freedom. In many cases we cannot factorise the joint posterior distribution and then we have to use other numerical simulation techniques, including acceptance/rejection sampling, importance sampling and MCMC, which are discussed in the following sections.

2.1.1.1 Acceptance/Rejection Sampling

Another approach for generating random samples from a probability density function $f(\theta)$ is acceptance/rejection sampling. Let $g(\theta)$ be a probability density function

that is easy to sample from and that satisfies the following condition for all θ:

$$f(\theta) \leq cg(\theta)$$

where $c > 0$. We can obtain samples from $f(\theta)$ using the following algorithm:

1. Generate $\theta^* \sim g(\theta)$.
2. Generate $u \sim \text{Uniform}(0, 1)$.
3. Accept θ^* as a sample from $f(\theta)$ if $u \leq \frac{f(\theta^*)}{cg(\theta^*)}$.

To prove that this process results in samples from $f(\theta)$ we make use of Bayes theorem:

$$p(\theta|\text{accept}) = \frac{p(\text{accept}|\theta)p(\theta)}{p(\text{accept})}$$

We know that:

$$\begin{aligned} p(\text{accept}|\theta) &= \frac{f(\theta)}{c \times g(\theta)} \\ p(\theta) &= g(\theta) \\ p(\text{accept}) &= \int_\theta p(\text{accept}|\theta)p(\theta)d\theta \\ &= \frac{1}{c}\int_\theta f(\theta)d\theta = \frac{1}{c} \end{aligned}$$

which leads to:

$$p(\theta|\text{accept}) = \frac{\frac{f(\theta)}{c \times g(\theta)}g(\theta)}{\frac{1}{c}} = f(\theta)$$

The difficulty with implementing this approach is finding a distribution, $g(\theta)$, that satisfies the condition $f(\theta) \leq c \times g(\theta)$ whilst using a value for c that does not result in a high frequency of rejections. For this reason, it is recommended that the $g(\theta)$ is similar to the target distribution. To overcome the problem of finding a suitable $g(\theta)$ over the whole sampling space, adaptive acceptance-rejection sampling has been proposed to sample from log-concave distributions (Gilks and Wild, 1992). This approach is based on the idea that the target distribution can be approximated by enclosing it using piecewise-exponential functions. The more samples generated, the better the approximation of $f(\theta)$. The advantage of adaptive acceptance-rejection sampling is that sampling new values of θ will become more efficient in time. A

disadvantage is that as well as being restricted to log-concave functions, we need to differentiate $f(\theta)$ to obtain the slope of the tangent line at θ^*. Gilks (1992) suggests a derivative-free alternative if we cannot differentiate $f(\theta)$. This approach is generally less useful for multivariate problems as the efficiency of the algorithm may become very low.

2.1.1.2 Importance Sampling

Importance sampling is an alternative technique that can be used to estimate properties of a distribution from which we cannot sample directly. It is a variance reduction technique which is based on the idea that some samples in a Monte Carlo simulation will have more impact on the estimation of the parameter of interest than others. The aim of importance sampling is therefore to sample these 'important' values more frequently to reduce the estimator variance. It makes use of the observation that the expected value of an arbitrary function of θ, $h(\theta)$, where $\theta \sim p(\theta)$ is given by:

$$\mathbb{E}[h(\theta)] = \int_{-\infty}^{\infty} h(\theta)p(\theta)d\theta \tag{2.1}$$

We can use ordinary Monte Carlo simulation to estimate $\mathbb{E}[h(\theta)]$ if we can sample easily from $p(\theta)$. If we cannot sample from $p(\theta)$, but we can sample from a probability density function $g(\theta)$, we can rewrite Equation 2.1 as:

$$\mathbb{E}[h(\theta)] = \int_{-\infty}^{\infty} \frac{h(\theta)p(\theta)}{g(\theta)} g(\theta)d\theta$$

We can now estimate $\mathbb{E}[h(\theta)]$ using:

1. Sample $\theta \sim g(\theta)$.
2. Calculate $w(\theta) = \frac{p(\theta)}{g(\theta)}$.
3. $\widehat{\mathbb{E}[h(\theta)]} = \frac{1}{N}\sum_{i=1}^{N} h(\theta)w(\theta)$.

Unlike acceptance-rejection sampling, every sampled value is used. The disadvantage of importance sampling is that if a few of the $w(\theta)$ are much larger than the others, those values will determine $\widehat{\mathbb{E}[h(\theta)]}$ and the result will behave as if it was estimated from a small sample. To overcome this behaviour, a distribution shape for

$g(\theta)$ needs to be selected in such a way that $p(\theta)$ is slightly smaller than $g(\theta)$ in the tails. Analogously to acceptance-rejection sampling, the choice of $g(\theta)$ determines the performance of this approach.

Sequential importance samplers are an extended version of the importance sampling algorithm above that can be used to sample from K-dimensional distributions:

1. Sample $\theta_1 \sim g(\theta_1)$.
2. Sample $\theta_k \sim g(\theta_k|\theta_1, \ldots, \theta_{k-1})$ for $k = 2, \ldots, K$.
3. Calculate:

$$\begin{aligned} w(\theta_1, \ldots, \theta_k) &= \frac{p(\theta_1, \ldots, \theta_{k-1})}{g(\theta_1, \ldots, \theta_{k-1})} \frac{p(\theta_1, \ldots, \theta_k)}{p(\theta_1, \ldots, \theta_{k-1}) g(\theta_k|\theta_1, \ldots, \theta_{k-1})} \\ &= w(\theta_1, \ldots, \theta_{k-1}) w(\theta_k|\theta_1, \ldots, \theta_{k-1}) \end{aligned}$$

4. $\widehat{\mathbb{E}[h(\boldsymbol{\theta})]} = \frac{1}{N} \sum_{i=1}^{N} h(\boldsymbol{\theta}) w(\boldsymbol{\theta})$.

One problem with the sequential importance sampler is that if the weight, w_k, for sample θ_k is small, the weights for $\theta_{k+1}, \ldots, \theta_K$ will be small as well due to the multiplicative character of the weights. Sequential importance resampling algorithms aim to overcome this by resampling the samples proportionally to their weights (Kitagawa, 1996).

2.1.1.3 Markov Chain Monte Carlo Sampling

If we cannot sample easily from $p(\theta)$, but we can evaluate the density function, Markov Chain Monte Carlo (MCMC) sampling may provide a solution to generate samples from $p(\theta)$. MCMC approaches aim to generate samples from a probability distribution by constructing a Markov chain, $X_1, \ldots, X_n$ whose equilibrium distribution is $p(\theta)$. A Markov chain is a discrete-time stochastic process $X_1, X_2, \ldots$ taking values in an arbitrary state space and having the property that the conditional distribution of X_{n+1} depends only on the present state X_n. In other words, MCMC approaches make use of $P(X_{n+1}, X_n) = P(X_{n+1}|X_n)P(X_n)$ and $P(X_{n+1}|X_1, \ldots, X_n) = P(X_{n+1}|X_n)$. For a more detailed overview of MCMC we

refer to Gamerman and Lopes (2006) and Roberts and Casella (2005).

Limitations of MCMC approaches include:

1. Subject to regularity conditions, the Markov chain will converge to the distribution of choice but initial samples may be from a different distribution. As a result, a number of 'burn-in' samples will have to be discarded.

2. Depending on the shape of the posterior distribution and the transition structure of the Markov chain, it may take a long time before the sampling space is fully explored.

3. Samples obtained using MCMC algorithms are correlated. If we want to reduce this correlation, we will have to discard many of the samples.

Well-known MCMC approaches include the Metropolis-Hastings algorithm, Gibbs sampling and slice sampling, which will be discussed in the following sections.

Metropolis-Hastings Algorithm

The Metropolis-Hastings algorithm (Metropolis et al., 1953; Hastings, 1970) can be used to generate samples from a probability density function $p(\theta)$ from which we cannot generate samples directly. Let $q(\theta^*|\theta^{(t)})$ be an arbitrary distribution that we can sample from, which we will call the proposal distribution. Given an arbitrarily chosen starting value $\theta^{(0)}$, we can generate a new sample $\theta^{(t+1)}$, given the most recent sample $\theta^{(t)}$ using the following steps:

1. Generate a proposal value $\theta^* \sim q\left(\theta^*|\theta^{(t)}\right)$.

2. Calculate the acceptance probability $\alpha = \min\left\{\frac{p(\theta^*)q\left(\theta^{(t)}|\theta^*\right)}{p\left(\theta^{(t)}\right)q\left(\theta^*|\theta^{(t)}\right)}, 1\right\}$. Note that for symmetrical proposal distributions, e.g. $\theta^*|\theta^{(t)} \sim \mathcal{N}(\theta^{(t)}, \sigma^2)$, $q\left(\theta^*|\theta^{(t)}\right) = q\left(\theta^{(t)}|\theta^*\right)$ and this becomes $\alpha = \min\left\{\frac{p(\theta^*)}{p\left(\theta^{(t)}\right)}, 1\right\}$.

3. Set $\theta^{(t+1)} = \theta^*$ with probability α and $\theta^{(t)}$ with probability $1 - \alpha$.

Metropolis-Hastings samplers are popular because they can be used even if the normalising constant is unknown. The reason for this is that the normalisation constant

will cancel out in the acceptance ratio. In addition to the generic limitations for MCMC approaches mentioned above, the main problem with Metropolis-Hastings algorithms is that in multivariate problems it may be hard to find an efficient proposal distribution. As a result the acceptance probability may be low, resulting in a slow exploration of the sampling space.

Gibbs sampling

Gibbs sampling (Geman and Geman, 1984; Gelfand and Smith, 1990) can be used to sample from a multivariate distribution $p(\theta_1, \ldots, \theta_n|y)$ when we cannot sample directly from the distribution itself. It can be seen as a special case of the Metropolis-Hastings algorithm. In multivariate cases, we want to generate a Markov chain with stationary distribution $p(\theta_1, \ldots, \theta_n|y)$. However, in many cases it is easier to sample from the posterior conditional distributions:

$$\begin{aligned} p(\theta_i|\theta_1, \ldots \theta_{i-1}, \theta_{i+1}, \ldots, \theta_n) &= \frac{p(\theta_1, \ldots, \theta_n)}{p(\theta_1, \ldots \theta_{i-1}, \theta_{i+1}, \ldots, \theta_n} \\ &\propto p(\theta_1, \ldots, \theta_n) \end{aligned}$$

This follows from the observation that the denominator is independent of θ_i and is therefore a normalisation constant. The easiest way to obtain a conditional distribution is to ignore all factors that are not dependent on θ_i as they are part of the normalisation constant. If this results in a familiar distribution form, we can sample from the conditional distribution directly. If not, we can use sampling approaches that do not require the normalisation constant for those variables, for example we can use a Metropolis-Hastings step within a Gibbs sampler.

Gibbs sampling is useful if the conditional distributions of a variable are known and consists of the following steps for a model with three random variables, θ_1, θ_2 and θ_3:

1. Generate $\theta_1^{(t+1)} \sim p\left(\theta_1|\theta_2^{(t)}, \theta_3^{(t)}\right)$.
2. Generate $\theta_2^{(t+1)} \sim p\left(\theta_2|\theta_1^{(t+1)}, \theta_3^{(t)}\right)$.
3. Generate $\theta_3^{(t+1)} \sim p\left(\theta_3|\theta_1^{(t+1)}, \theta_2^{(t+1)}\right)$.

Gibbs samplers are a special case of Metropolis-Hastings samplers with acceptance probability one:

$$\frac{p\left(\boldsymbol{\theta}^{*}\right)}{p\left(\boldsymbol{\theta}^{(t)}\right)}\frac{q\left(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{*}\right)}{q\left(\boldsymbol{\theta}^{*}|\boldsymbol{\theta}^{(t)}\right)} = \frac{p(\theta_i^{*}|\boldsymbol{\theta}_{-i}^{*})p(\boldsymbol{\theta}_{-i}^{*})}{p\left(\theta_i^{(t)}|\boldsymbol{\theta}_{-i}^{(t)}\right)p\left(\boldsymbol{\theta}_{-i}^{(t)}\right)}\frac{p\left(\theta_i^{(t)}|\boldsymbol{\theta}_{-i}^{*}\right)}{p\left(\theta_i^{*}|\boldsymbol{\theta}_{-i}^{(t)}\right)}$$
$$= 1$$

where $\boldsymbol{\theta}_{-i} = \{\theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_n\}$ and $\boldsymbol{\theta}_{-i}^{*}$ is equal to $\boldsymbol{\theta}_{-i}^{(t)}$.

An alternative to the standard Gibbs sampler is the 'blocked Gibbs sampler' where variables can be grouped together and samples are taken from their joint distributions. For the model above we can use a 'blocked Gibbs sampler' which samples θ_1 and θ_2 together:

1. Generate $\left(\theta_1^{(t+1)}, \theta_2^{(t+1)}\right) \sim p\left(\theta_1, \theta_2|\theta_3^{(t)}\right)$.
2. Generate $\theta_3^{(t+1)} \sim p\left(\theta_3|\theta_1^{(t+1)}, \theta_2^{(t+1)}\right)$.

Alternatively, a 'collapsed Gibbs sampler' may be useful if it is easier to sample from a marginal distribution than from the full conditional distribution of one of the variables. Let us again consider the above model but now use the following collapsed Gibbs sampler:

1. Generate $\theta_1^{(t+1)} \sim p\left(\theta_1|\theta_3^{(t)}\right)$.
2. Generate $\theta_2^{(t+1)} \sim p\left(\theta_2|\theta_1^{(t+1)}, \theta_3^{(t)}\right)$.
3. Generate $\theta_3^{(t+1)} \sim p\left(\theta_3|\theta_1^{(t+1)}, \theta_2^{(t+1)}\right)$.

Note that the first two steps result in $p\left(\theta_1, \theta_2|\theta_3^{(t)}\right)$ which we used in the first step of the 'blocked Gibbs sampler'. In this algorithm we integrated out θ_2 from the joint distribution $p(\theta_1, \theta_2|\theta_3)$ to obtain a sample of θ_1 assuming that it is easier to sample from $p\left(\theta_1|\theta_3^{(t)}\right)$ than from $p\left(\theta_1|\theta_2^{(t)}, \theta_3^{(t)}\right)$. Note that the opposite of collapsed Gibbs sampling is often used as well: in many models auxiliary variables are added to facilitate sampling. If the auxiliary variables are latent, i.e. not observed and inferred from other variables through a mathematical model, this is referred to as data augmentation (Tanner and Wong, 1987).

Slice sampler

A special case of a sampler that makes use of data augmentation is the slice sampler (Neal, 2003). In the univariate case, where we want to sample from $p(\theta)$, the idea is to sample from the two-dimensional region that lies under $p(\theta)$. This can be achieved by introducing an auxiliary variable h and defining a joint distribution over θ and h that is uniform over the region $U = \{\theta, h\} : 0 < h < p(\theta)$ below the curve defined by $p(\theta)$. Using Gibbs sampling, we can obtain samples from $p(\theta)$ as follows:

1. Sample $h^{(t+1)} \sim \text{Uniform}\left(0, p\left(\theta^{(t)}\right)\right)$ which defines a 'horizontal' slice $S = \{\theta : h < p(\theta)\}$.

2. As it may be difficult to find the whole region S, it has been suggested to instead find an interval $S' = (L, U)$ around $\theta^{(t)}$.

3. Sample $\theta^* \sim \text{Uniform}(L, U)$.

4. Set $\theta^{(t+1)} = \theta^*$ if $p(\theta^*) \geq h$ and return to Step 3 if $p(\theta^*) < h$.

To be most efficient, S' should be the smallest interval that contains S, but this is often impossible. Therefore, estimates of S can be obtained from e.g. the domain of θ or by stepwise increasing/decreasing the interval S' until a point outside/inside the interval is obtained.

The main advantage of a slice sampler over a Gibbs sampler is that it does not require the (conditional) distributions to be in a form that we can easily sample from. The main advantage in comparison with a Metropolis-Hastings sampler is that we do not have to tune the proposal distribution as a slice sampler will dynamically adjust the scale of the proposal distribution, depending on the current value $\theta^{(t)}$. The main disadvantage of a slice sampler is that it can be hard to find the interval S and that it may not be as efficient as a Gibbs sampler for multivariate problems.

2.1.1.4 Conclusion

As discussed, several approaches exist to obtain samples from a probability density function. Depending on the characteristics of the distribution of interest, some methods may be easier to implement or more efficient than others. These sampling techniques can be combined to obtain the most efficient sampling approach for a given model, e.g. Metropolis-Hastings steps within a Gibbs sampler.

2.2 The Dirichlet distribution

The Dirichlet distribution is a probability distribution over the C-dimensional standard simplex $\Delta^C = \{(\theta_1, \ldots, \theta_m) : \theta_j \geq 0, \sum_{j=1}^m \theta_j = 1\}$, where $C = m - 1$. The standard $m - 1$ simplex is the space of all discrete probability distributions on m possible outcomes. The Dirichlet distribution is a family of continuous, multivariate distributions with a single parameter vector $\boldsymbol{\gamma}$. It is the multivariate generalisation of the Beta distribution and is often used in a Bayesian context as the conjugate prior in problems involving a Multinomial likelihood. If we let

$$\boldsymbol{\theta} = \{\theta_1, \ldots, \theta_{m-1}\}$$
$$\boldsymbol{\theta} \sim \text{Dirichlet}(\gamma_1, \ldots, \gamma_m)$$

then the probability density function of the Dirichlet distribution is given by:

$$p(\theta_1, \ldots, \theta_{m-1}; \gamma_1, \ldots, \gamma_m) = \frac{\Gamma(\sum_{j=1}^m \gamma_j)}{\prod_{j=1}^m \Gamma(\gamma_j)} \prod_{j=1}^m \theta_j^{\gamma_j - 1}$$

Note that $\theta_m = 1 - \sum_{j=1}^{m-1} \theta_j$ as the θ_js need to sum to one. The Dirichlet distribution is sometimes represented using two parameters: a concentration parameter $\gamma_0 = \sum \gamma_j$ and a base measure $\{\gamma'_1, \ldots, \gamma'_m\}$ with $\gamma'_j = \frac{\gamma_j}{\gamma_0}$.

2.2.1 Derivation

Let $\mathbf{w} = \{w_1, \ldots, w_m\}$ and $w_i \overset{\text{i.i.d.}}{\sim} \text{Gamma}(\gamma_i, 1)$ with $\gamma_i > 0$. Let us define the normalising constant $W = \sum_{i=1}^m w_i$ and $\theta_i = \frac{w_i}{W}$ for $i \in \{1, \ldots, m-1\}$, leading to

$\boldsymbol{\theta} \in \Delta^{m-1}$. To determine the distribution of $\boldsymbol{\theta}$, we can use our knowledge about $\mathbf{w}$. We know that:

$$\begin{aligned} p(\mathbf{w}) &= \frac{1}{\prod_{i=1}^{m} \Gamma(\gamma_i)} \exp\left[-\sum_{i=1}^{m} w_i\right] \prod_{i=1}^{m} w_i^{\gamma_i - 1} \\ &= \frac{1}{\prod_{i=1}^{m} \Gamma(\gamma_i)} \exp\left[-W\right] \prod_{i=1}^{m} w_i^{\gamma_i - 1} \end{aligned}$$

We can find the distribution of $\boldsymbol{\theta}$ by changing variables from $\mathbf{w}$ to $(\boldsymbol{\theta}, W)$ using the fact that the determinant of the Jacobian equals W^{m-1}:

$$p(\theta_1, \ldots, \theta_{m-1}, W) = \frac{1}{\prod_{i=1}^{m} \Gamma(\gamma_i)} \exp[-W] W^{\left(\sum_{i=1}^{m} \gamma_i\right) - 1} \prod_{i=1}^{m} \theta_i^{\gamma_i - 1}$$

To obtain the marginal distribution of $\boldsymbol{\theta} = \{\theta_1, \ldots, \theta_{m-1}\}$, we need to integrate over W.

$$\begin{aligned} p(\theta_1, \ldots, \theta_{m-1}) &= \frac{1}{\prod_{i=1}^{m} \Gamma(\gamma_i)} \left(\prod_{i=1}^{m} \theta_i^{\gamma_i - 1}\right) \int_0^{\infty} \exp(-W) W^{\left(\sum_{i=1}^{m} \gamma_i\right) - 1} dW \\ &= \frac{\Gamma(\sum_{i=1}^{m} \gamma_i)}{\Gamma(\gamma_1) \ldots \Gamma(\gamma_m)} \theta_1^{\gamma_1 - 1} \ldots \theta_m^{\gamma_m - 1} \end{aligned} \tag{2.2}$$

As the integral $\int_0^{\infty} \exp(-W) W^{\left(\sum_{i=1}^{m} \gamma_i\right) - 1} dW$ is in the form of a Gamma$(\sum_{i=1}^{m} \gamma_i, 1)$ probability density function bar the normalisation constant, $\Gamma(\sum_{i=1}^{m} \gamma_i)$, we can easily integrate over W. In Equation 2.2 we recognise the Dirichlet probability density function.

2.2.2 Relation to other distributions

Before we explore how the Dirichlet distribution can be used, we first describe the relationship between the Dirichlet distribution and the Gamma and Beta distributions.

2.2.2.1 Relation to the Gamma distribution

Given the previously explored relationship between the Gamma distribution and the Dirichlet distribution (see Section 2.2.1), it will be interesting to see how several characteristics of the Gamma distribution affect the properties of the Dirichlet distribution. This will be useful when explaining the stick-breaking algorithm for the

Dirichlet process in Section 2.3.3.1.

We will first explore the summation property of the Gamma distribution. Let X_1 and X_2 be two independent random variables, with distributions $X_1 \sim \text{Gamma}(\gamma_1, b)$ and $X_2 \sim \text{Gamma}(\gamma_2, b)$. The joint probability density function can be written as:

$$p(x_1, x_2) = \frac{b^{\gamma_1+\gamma_2}}{\Gamma(\gamma_1)\Gamma(\gamma_2)} \exp\left[-(x_1 + x_2)b\right] x_1^{\gamma_1-1} x_2^{\gamma_2-1}$$

Now, define $u = x_1 + x_2$ and $v = \frac{x_1}{x_1+x_2}$ so that $x_1 = uv$ and $x_2 = u(1-v)$. Reparameterising the joint distribution requires the determinant of the Jacobian matrix:

$$J(x_1, x_2) = \begin{vmatrix} \frac{dx_1}{du} & \frac{dx_1}{dv} \\ \frac{dx_2}{du} & \frac{dx_2}{dv} \end{vmatrix} = \begin{vmatrix} v & u \\ 1-v & -u \end{vmatrix} = |-vu - (u - vu)| = u$$

This leads to:

$$p(u, v) = \frac{b^{\gamma_1+\gamma_2}}{\Gamma(\gamma_1)\Gamma(\gamma_2)} \exp[-ub] u^{\gamma_1+\gamma_2-1} v^{\gamma_1-1} (1-v)^{\gamma_2-1}$$

The sum of X_1 and X_2, u, has density:

$$\begin{aligned} p(u) = \int p(u, v)dv &= \int_0^1 \frac{b^{\gamma_1+\gamma_2}}{\Gamma(\gamma_1)\Gamma(\gamma_2)} \exp[-ub] u^{\gamma_1+\gamma_2-1} v^{\gamma_1-1} (1-v)^{\gamma_2-1} dv \\ &= \frac{b^{\gamma_1+\gamma_2} \exp[-ub] u^{\gamma_1+\gamma_2-1}}{\Gamma(\gamma_1)\Gamma(\gamma_2)} \int_0^1 v^{\gamma_1-1} (1-v)^{\gamma_2-1} dv \end{aligned}$$

Apart from the normalising constant, the integral is equal to the Beta(γ_1, γ_2) probability density function, so the integral equals $\frac{\Gamma(\gamma_1)\Gamma(\gamma_2)}{\Gamma(\gamma_1+\gamma_2)}$. Therefore:

$$p(u) = \frac{b^{\gamma_1+\gamma_2} \exp[-ub] u^{\gamma_1+\gamma_2-1}}{\Gamma(\gamma_1 + \gamma_2)}$$

in which we recognise a Gamma$(\gamma_1 + \gamma_2, b)$ distribution. So the sum of two independent Gamma random variables with the same rate/scale parameter has a Gamma distribution.

The summation property of the Gamma distribution can be translated into the aggregation property of the Dirichlet distribution. In section 2.2.1 we showed that if we had independent samples $w_i \sim \text{Gamma}(\gamma_i, 1)$, the variables $\theta_i = \frac{w_i}{W}$

for $i = 1, \ldots, m-1$ followed a Dirichlet$(\gamma_1, \ldots, \gamma_m)$ distribution. If we define $w_* = w_1 + w_2$, we know from the previous section that $w_* \sim \text{Gamma}(\gamma_1 + \gamma_2, 1)$. Extending this to $\theta_* = \frac{w_1+w_2}{W}$ leads to:

$$(\theta_*, \theta_3, \ldots, \theta_{m-1}) \sim \text{Dirichlet}(\gamma_1 + \gamma_2, \gamma_3, \ldots, \gamma_m)$$

Note that as well as aggregating θ_is, the Dirichlet variables can also be split, i.e. there exist θ_{1+} and θ_{1-} such that:

$$(\theta_{1^+}, \theta_{1^-}, \ldots, \theta_{m-1}) \sim \text{Dirichlet}(\gamma_{1^+}, \gamma_{1^-}, \ldots, \gamma_m)$$

with $\gamma_{1^+} + \gamma_{1^-} = \gamma_1$ and $\theta_1 = \theta_{1+} + \theta_{1-}$.

2.2.2.2 Relation to the Beta Distribution

We can use the aggregation property of the Dirichlet distribution to derive the marginal distribution of θ_i:

$$\theta_i, \sum_{j \neq i} \theta_j \sim \text{Dirichlet}\left(\gamma_i, \sum_{j \neq i} \gamma_j\right)$$

As $\sum_{j \neq i} \gamma_j = \gamma_0 - \gamma_i$ and $\sum_{j \neq i} \theta_j = 1 - \theta_i$, the marginal distribution for θ_i is:

$$p(\theta_i) = \frac{\Gamma(\gamma_0)}{\Gamma(\gamma_i)\Gamma(\gamma_0 - \gamma_i)} \theta_i^{\gamma_i - 1} \left(\sum_{j \neq i} \theta_j\right)^{\gamma_0 - \gamma_i - 1} = \frac{\Gamma(\gamma_0)}{\Gamma(\gamma_i)\Gamma(\gamma_0 - \gamma_i)} \theta_i^{\gamma_i - 1} (1 - \theta_i)^{\gamma_0 - \gamma_i - 1}$$

in which we recognise a Beta $(\gamma_i, \gamma_0 - \gamma_i)$ distribution.

2.2.3 Properties of the Dirichlet distribution

2.2.3.1 Mean and Variance

The mean of a Dirichlet distribution can easily be derived from the Beta marginal distributions, $\theta_i \sim \text{Beta}(\gamma_i, \gamma_0 - \gamma_i)$:

$$\mathbb{E}[\theta_i] = \frac{\Gamma(\gamma_0)}{\Gamma(\gamma_i)\Gamma(\gamma_0 - \gamma_i)} \int_0^1 \theta_i \theta_i^{\gamma_i - 1} (1 - \theta_i)^{\gamma_0 - \gamma_i - 1} = \frac{\gamma_i}{\gamma_0} \tag{2.3}$$

The variance is given by:

$$\text{Var}[\theta_i] = \int (\theta_i - \mathbb{E}[\theta_i])^2 f(\theta_i) d\theta_i = \frac{\gamma_i(\gamma_0 - \gamma_i)}{\gamma_0^2(\gamma_0 + 1)} \tag{2.4}$$

It is clear from the properties of the Dirichlet distribution that small values of γ_0 favour more dispersed distributions and that as $\gamma_0 \to \infty$, the probabilities are known, i.e. the Dirichlet distribution approximates a Dirac delta function at $\{\frac{\gamma_1}{\gamma_0}, \ldots, \frac{\gamma_m}{\gamma_0}\} = \{\gamma'_1, \ldots, \gamma'_m\}$.

2.2.4 Random Number Generation

In this section we present three different ways to generate random values from the Dirichlet distribution. Even though we use the random number generator based on normalised Gamma random variates, discussed next, we will discuss two alternative approaches for completeness.

2.2.4.1 Using the Gamma distribution

One way to generate samples, $\boldsymbol{\theta}$, from a Dirichlet$(\gamma_1, \ldots, \gamma_n)$ distribution is based on the relationship with the Gamma distribution (see Section 2.2.1). The algorithm is as follows:

1. For $i = 1$ to n, repeat $y_i \sim \text{Gamma}(\gamma_i, 1)$.
2. $\theta_i = \frac{y_i}{\sum_{i=1}^{n} y_i}$.

2.2.4.2 Using the Pólya Urn scheme

The Pólya urn scheme is related to the Dirichlet-Multinomial distribution. In this model an urn contains balls of m colours. After a draw of a ball of a particular colour, the ball is put back together with an extra ball of the same colour. In the bivariate case, the Dirichlet-Multinomial distribution is known as the Beta-Binomial distribution. Let us consider an urn with red and black balls. Before the first draw, the probability of drawing a red ball is given by $\frac{\gamma_R}{\gamma_R+\gamma_B}$, where the parameters γ_R and γ_B are the number of red and black balls respectively. If the first ball drawn is red, the probability of a red ball in the second draw is $\frac{\gamma_R+1}{\gamma_R+\gamma_B+1}$ and as a consequence the probability of drawing a sequence (red, red) is $\left(\frac{\gamma_R}{\gamma_R+\gamma_B}\right)\left(\frac{\gamma_R+1}{\gamma_R+\gamma_B+1}\right)$. Let X_n be a random variable denoting the number of red balls, k, after n draws. As $P(X_n = k)$

does not depend on the order in which the balls are drawn (easily shown) it can be written as:

$$\begin{aligned} P(X_n = k) &= \binom{n}{k} \frac{\prod_{i=1}^{k} [\gamma_R + i - 1] \prod_{i=1}^{n-k} [\gamma_B + i - 1]}{\prod_{i=1}^{n} [\gamma_R + \gamma_B + i - 1]} \\ &= \frac{\Gamma(n+1)}{\Gamma(n + \gamma_R + \gamma_B)} \frac{\Gamma(k + \gamma_R)}{\Gamma(k+1)} \frac{\Gamma(n - k + \gamma_B)}{\Gamma(n - k + 1)} \frac{\Gamma(\gamma_R + \gamma_B)}{\Gamma(\gamma_R)\Gamma(\gamma_B)} \end{aligned}$$

In this we recognise the Beta-Binomial distribution, which in a Bayesian context can be obtained as the posterior predictive distribution when a Binomial likelihood function is combined with a Beta prior distribution on the probability parameter.

Using Stirling's approximation, $x! = \Gamma(x+1) \approx \sqrt{2\pi x} \left(\frac{x}{e}\right)^x$ as $x \to \infty$, we can explore the distribution of X_n as n becomes very large. This leads to:

$$\begin{aligned} \frac{\Gamma(x+a)}{\Gamma(x+b)} &\approx \frac{\sqrt{2\pi(x+a-1)} \left(\frac{x+a-1}{e}\right)^{x+a-1}}{\sqrt{2\pi(x+b-1)} \left(\frac{x+b-1}{e}\right)^{x+b-1}} \\ &\approx x^{a-b} \end{aligned}$$

as $x \to \infty$ (and $x >> a$ and $x >> b$). If $n \to \infty$, $k \to \infty$ and $(n-k) \to \infty$ we can rewrite the Beta-Binomial distribution as:

$$P\left(X_n = k\right) = \frac{\Gamma(\gamma_R + \gamma_B)}{\Gamma(\gamma_R)\Gamma(\gamma_B)} k^{\gamma_R - 1} n^{1 - \gamma_R - \gamma_B} (n-k)^{\gamma_B - 1}$$

As $n \to \infty$, the proportion of red balls, $\theta_n = \frac{k}{n}$. becomes effectively continuous. Change of variables with $dk = n\, d\theta$ leads to:

$$\begin{aligned} p\left(\frac{X_n}{n} = \theta_n\right) &= \frac{\Gamma(\gamma_R + \gamma_B)}{\Gamma(\gamma_R)\Gamma(\gamma_B)} (n\theta_n)^{\gamma_R - 1} n^{1 - \gamma_R - \gamma_B} (n - \theta_n n)^{\gamma_B - 1} n \\ &= \frac{\Gamma(\gamma_R + \gamma_B)}{\Gamma(\gamma_R)\Gamma(\gamma_B)} \theta_n^{\gamma_R - 1} (1 - \theta_n)^{\gamma_B - 1} \end{aligned}$$

in which we recognise a Beta(γ_R, γ_B) distribution. As a result, sampling using a Pólya urn scheme converges to samples from a Beta(γ_R, γ_B) distribution as $n \to \infty$. Analogously to the Beta-Binomial, sampling from the Dirichlet-Multinomial distribution converges to a Dirichlet distribution as the number of draws $n \to \infty$.

2.2.4.3 Using a Stick-breaking Scheme

A third alternative to generate samples, θ, from a Dirichlet distribution is based on a stick-breaking algorithm (Connor and Mosimann, 1969) and uses Beta distributions.

Let $\beta_i \sim \text{Beta}(a_i, b_i)$ for $i = 1, \ldots, m-1$ and $\beta_m = 1$. Let us define:

$$\theta_i = \beta_i \prod_{j=0}^{i-1}(1-\beta_j) = \left(1 - \sum_{j=1}^{i-1}\theta_j\right)\beta_i$$

with $\sum_{i=1}^{m}\theta_i = 1$ and $\beta_0 = 0$ and define $S_i = \sum_{j=1}^{i}\theta_j$ with $S_0 = 0$, we obtain:

$$\theta_i = (1 - S_{i-1})\beta_i$$

We know that:

$$p(\beta_1, \ldots, \beta_{m-1}) \propto \prod_{i=1}^{m-1} \beta_i^{a_i-1}(1-\beta_i)^{b_i-1}$$

Changing variables to θ_i (using the fact that the determinant of the Jacobian matrix is $\prod_{i=1}^{m-1}(1-S_{i-1})^{-1}$) results in:

$$\begin{aligned} p(\theta_1, \ldots, \theta_{m-1}) &\propto \prod_{i=1}^{m-1} \frac{1}{1-S_{i-1}} \left(\frac{\theta_i}{1-S_{i-1}}\right)^{a_i-1} \left(1 - \frac{\theta_i}{1-S_{i-1}}\right)^{b_i-1} \\ &\propto \prod_{i=1}^{m-1} \left(\sum_{j=i}^{m}\theta_j\right)^{b_{i-1}-(a_i+b_i)} \theta_i^{a_i-1} \frac{\left(\sum_{j=i+1}^{m}\theta_j\right)^{b_i-1}}{\left(\sum_{j=i}^{m}\theta_j\right)^{b_{i-1}-1}} \end{aligned}$$

where $\theta_i = S_i - S_{i-1}$ and $1 - S_{i-1} = \left(1 - \sum_{j=1}^{i-1}\theta_j\right) = \sum_{j=i}^{n}\theta_j$. With:

$$\begin{aligned} \prod_{i=1}^{m-1} \frac{\left(\sum_{j=i+1}^{m}\theta_j\right)^{b_i-1}}{\left(\sum_{j=i}^{m}\theta_j\right)^{b_{i-1}-1}} &= \frac{(\theta_2 + \cdots + \theta_m)^{b_1-1}}{(\theta_1 + \cdots + \theta_m)^{b_0-1}} \frac{(\theta_3 + \cdots + \theta_m)^{b_2-1}}{(\theta_2 + \cdots + \theta_m)^{b_1-1}} \cdots \\ &\quad \frac{(\theta_{m-1} + \theta_m)^{b_{m-2}-1}}{(\theta_{m-2} + \theta_{m-1} + \theta_m)^{b_{m-1}-1}} \frac{(\theta_m)^{b_{m-1}-1}}{(\theta_{m-1} + \theta_m)^{b_{m-2}-1}} \\ &= (\theta_m)^{b_{m-1}-1} \end{aligned}$$

we obtain:

$$p(\theta_1, \ldots, \theta_{m-1}) \propto \theta_m^{b_{m-1}-1} \prod_{i=1}^{m-1} \theta_i^{a_i-1} \left(\sum_{j=i}^{m}\theta_j\right)^{b_{i-1}-(a_i+b_i)}$$

which is known as the generalised Dirichlet distribution (Connor and Mosimann, 1969). If we set $b_{i-1} = a_i + b_i$ for $i = 2, \ldots, m-1$ and $b_{m-1} = a_m$, we find:

$$p(\theta_1, \ldots, \theta_{m-1}) \propto \theta_m^{a_m-1} \prod_{i=1}^{m-1} \theta_i^{a_i-1}$$

So:

$$\theta_1, \ldots, \theta_{m-1} \sim \text{Dirichlet}(a_1, \ldots, a_m)$$

If we define $b_i = \sum_{j=i+1}^{m} a_j$ and sample $\beta_i \sim \text{Beta}(a_i, b_i)$, we can obtain samples from a Dirichlet distribution by setting $\theta_1 = \beta_1$ and $\theta_i = \beta_i \prod_{j=1}^{i-1}(1 - \beta_j)$.

The basic idea behind this approach is that a stick of unit length can be broken sequentially into N pieces of different length in such a way that the lengths of the pieces follow a Dirichlet distribution. In other words, the Dirichlet distribution can be used to specify the expected value of the relative length of each piece. The stick-lengths will be sampled from Beta distributions as defined above. A Dirichlet$(4, 2, 1)$ indicates that for a large number of sticks, the average 'first' stick length will be twice as long as the 'second' stick, which will in turn be twice as long as the third stick.

2.2.5 Bayesian Inference using Dirichlet distributions

2.2.5.1 Posterior Distribution for multinomial trials

For multinomial trials, the probability of an observation X being in category k can be written as: $P(X = k|\boldsymbol{\theta}) = \theta_k$ where $\sum_{k=1}^{m} \theta_k = 1$ and m is the number of categories. The likelihood for n_k observations in category k in $N = \sum_{k=1}^{m} n_k$ is given by:

$$p(\mathbf{n}|\boldsymbol{\theta}) = N! \prod_{k=1}^{m} \frac{\theta_k^{n_k}}{n_k!}$$

The Dirichlet distribution is the conjugate prior for the event probability parameter $\boldsymbol{\theta}$:

$$p(\boldsymbol{\theta}|\boldsymbol{\gamma}) \propto \prod_{k=1}^{m} \theta_k^{\gamma_k - 1}$$

where $\boldsymbol{\gamma} = \{\gamma_1, \ldots, \gamma_m\}$. This will lead to the posterior:

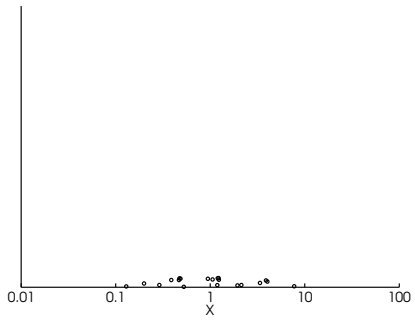
$$p(\boldsymbol{\theta}|n_1, \ldots, n_m, \boldsymbol{\gamma}) \propto \prod_{k=1}^{m} \theta_k^{\gamma_k + n_k - 1}$$

in which we recognise a Dirichlet$(\gamma_1 + n_1, \ldots, \gamma_m + n_m)$ distribution. If $\boldsymbol{\theta}$ is considered as a sample from a Dirichlet distribution of possible parameter vectors of

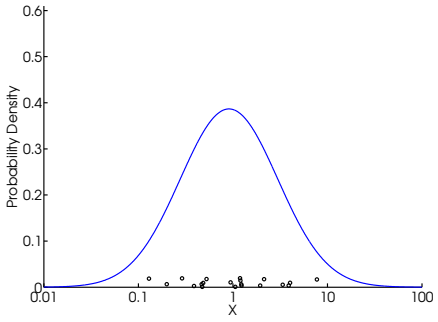
a Multinomial distribution, then this will lead to a distribution over distributions. The idea of a distribution over distributions can be explained using Figure 2.1.

Figure 2.1 – *Illustration of a distribution over distributions.*

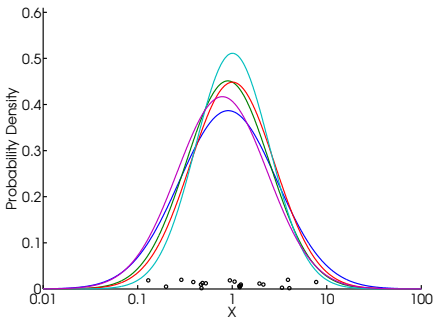
(a) Data set (n = 20) which is assumed to be a sample from a Normal distribution.



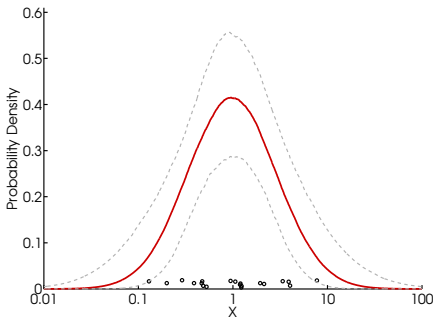
(b) A single Normal distribution that can be fitted credibly to the data.



(c) Other Normal distributions that can credibly be fitted to the data.



(d) Median distribution (red solid line) and 95% credible intervals (grey dashed lines) over the set of distributions that can be used to describe the data.



Suppose we have a data set of size 20 for a random variable X which we believe to be a sample from a Normal distribution. We do not know what the parameters of the Normal distribution are, so we use a Bayesian framework to account for our uncertainty. A possible Normal distribution that could be credibly fitted to the data is given in Figure 2.1b. However, given the small sample size other Normal distributions may fit the data just as well or even better. Figure 2.1c shows a set of five Normal distributions that can all be credibly used to describe the data, but we

can extend this to more distributions. Figure 2.1d summarises this by plotting the median and the 95% credible intervals. A $100(1-\alpha)\%$ credible interval is an interval $\mathcal{C}$ on the parameter space Θ such that $\int_{\mathcal{C}} p(\theta|y)d\theta = 1-\alpha$, where $p(\theta|y)$ represents the posterior distribution. In this thesis we will use equal-tailed credible intervals where the probability of θ being below the interval is equal to the probability of θ being above it $(\alpha/2)$. In other words we can calculate the 95^{th} credible interval using the 2.5^{th} and 97.5^{th} percentiles of the set of distributions for each value of x and this provides us with a distribution over (Normal) distributions.

It is important to realise that the distribution over the Normal distributions represents uncertainty around the Normal density functions. In other words, we assume the variability in the data is constrained to have a Normal distribution form so that the distribution over the Normal density functions only describes our lack of knowledge about the parameters μ and σ of the Normal densities. Analogously to the posterior distributions of the uncertain parameters of the Normal distribution, the Dirichlet distribution can be thought of as the natural probability distribution of the uncertain parameter vector $\boldsymbol{\theta}$ of a Multinomial distribution.

2.2.5.2 Predictive distribution for multinomial trials

The predictive distribution for the next observation X^{N+1} can be derived as:

$$\begin{aligned} P(X^{N+1} = k|\mathbf{X}) &= \int_0^1 p\left(X^{N+1}|\theta\right) p(\theta|X)d\theta \\ &= \frac{\gamma_k + n_k}{\gamma_0 + N} \end{aligned}$$

2.2.6 Applications of the Dirichlet distribution

In this section we discuss the use of Dirichlet distributions in finite mixture models and for clustering problems.

2.2.6.1 Finite Mixture Models

From Discrete to Continuous

The fact that the Dirichlet distribution provides a distribution over discrete distributions is obvious from the different representations discussed previously. For example, in the Pólya urn representation, a finite, fixed number of coloured balls could be observed and in the stick-breaking representation the number of pieces of stick was fixed. Therefore, to make the Dirichlet distribution useful to describe continuous data, we need to define a mixture distribution, consisting of continuous distributions as mixture components whose weights follow a Dirichlet distribution.

Generative

Let us first consider a data generating process which assumes that data can be associated with C components of the same family of distributions in a mixture distribution. Let w_k be the probability that data will be generated from component k and let data within each component be distributed as $f(\cdot|\theta_k)$, where $\theta_k \overset{\text{i.i.d.}}{\sim} G_0$, a distribution over the component parameters. To generate a value from this process, we will first select a component k from the distribution of components with probability $\mathbf{w} = \{w_1, \ldots, w_C\}$. Next we can generate a data value $y_i \sim f(\cdot|\theta_k)$. This results in the finite mixture model $p(y|\boldsymbol{\theta}, \mathbf{w}) = \sum_{k=1}^{C} w_k f(y|\theta_k)$. The continous mixture model can be represented as mixing a discrete distribution on the space of component parameters θ with a continuous distribution $f(y|\boldsymbol{\theta})$:

$$
\begin{aligned}
y_k|\theta_k &\sim f(\cdot|\theta_k) \\
p(\theta_k) &= \sum_{k=1}^{C} w_k \delta_{\theta_k} \\
\theta_1, \ldots, \theta_C &\sim G_0 \qquad\qquad i.i.d. \\
\{w_1, \ldots, w_{C-1}\} &\sim \text{Dirichlet}(\gamma_1, \ldots, \gamma_C)
\end{aligned}
$$

where δ_{θ_k} is a measure with a point mass of one at θ_k. Using this representation we can generate data by first sampling C parameters θ_i from G_0 and subsequently sampling the data, y, from $f(y|\theta_i)$.

Another representation makes use of a latent Multinomial variable $\mathbf{K}$, a component assignment variable. If $K_i = k$, the i^{th} data value is considered to be associated with component k:

$$\begin{aligned} p(y_i|\boldsymbol{\theta}, \mathbf{w}, \mathbf{K}) &= \sum_{k=1}^{C} p(K_i = k|\mathbf{w}) p(y_i|K_i = k, \boldsymbol{\theta}) \\ &= \sum_{k=1}^{C} w_k p(y|\theta_k) \end{aligned}$$

This leads to:

$$\begin{aligned} y_i|\boldsymbol{\theta}, \mathbf{K} &\sim f(\cdot|\theta_{K_i}) \\ K_i|\mathbf{w} &\sim \text{Multinomial}(w_1, \ldots, w_C) \\ \theta_1, \ldots, \theta_C &\sim G_0 \\ w_1, \ldots, w_{C-1} &\sim \text{Dirichlet}(\gamma_1, \ldots, \gamma_C) \end{aligned} \tag{2.5}$$

Bayesian Inference

If we want to fit a mixture model in a Bayesian context, we will need to define prior distributions for $\mathbf{w}$ and $\boldsymbol{\theta}$. As we saw in the previous section, a Dirichlet distribution can be used as a conjugate prior distribution for $\mathbf{w}$ so we only need to define G_0, the prior distribution for $\boldsymbol{\theta}$. We want to obtain either the posterior distribution $p(\boldsymbol{\theta}, \mathbf{w}|\mathbf{y})$ or the predictive distribution $p(y^{n+1}|y_1 \ldots, y_n)$. Since no analytical solution exists, we need to generate samples from these distributions for inference purposes. The simplest MCMC sampling scheme, a collapsed Gibbs sampler (see Section 2.1.1.3), makes use of the latent component assignment variable $\mathbf{K}$. Samples from the posterior distributions of the parameters of a finite mixture model can now be obtained using the following collapsed Gibbs sampling scheme:

$$\begin{aligned} &p(\mathbf{K}|\mathbf{w}, \boldsymbol{\theta}, \mathbf{y}) \\ &p(\mathbf{w}|\mathbf{K}, \mathbf{y}) \\ &p(\boldsymbol{\theta}|\mathbf{K}, \mathbf{y}) \end{aligned}$$

where $\mathbf{y} = \{y_1, \ldots, y_n\}$. We explain each step in the following sections.

p(K|w, θ, y)

The distribution of component allocations is given by:

$$p\left(K_i = k|\mathbf{w}, \boldsymbol{\theta}, \mathbf{y}\right) = \frac{w_k f(y_i|\theta_k)}{\sum_{k=1}^{C} w_k f(y_i|\theta_k)} \tag{2.6}$$

p(w|K, y)

Let $m_k = \sum_{i=1}^{n} \delta_{K_i,k}$ (i.e. the number of data values assigned to component k) and $\mathbf{m} = \{m_1, \ldots, m_C\}$. Given $\mathbf{m}$, the posterior distribution of $\mathbf{w}$ is:

$$\mathbf{w}|\mathbf{K}, \mathbf{y} \sim \text{Dirichlet}(\gamma_1 + m_1, \ldots, \gamma_C + m_C)$$

p(θ|K, y)

The introduction of $\mathbf{K}$ allows the parameters $\boldsymbol{\theta}$ for each component to be updated separately, which is easier if G_0 and $f(\mathbf{y}|\boldsymbol{\theta})$ are conjugate. For a mixture of Normal components, $\mathcal{N}(\theta_k, 1/\tau)$, choosing a conjugate Normal distribution $G_0 = p(\theta_k) \sim \mathcal{N}(\mu_0, 1/\tau_0)$ for the location parameters, θ_k, means that they can be updated using a simple Bayesian step:

$$p(\theta_k|\mathbf{K}, \mathbf{y}) \sim \mathcal{N}\left(\frac{\mu_0\tau_0 + \tau\sum_{i=1}^{N} y_i\delta_{K_i,k}}{\tau_0 + m_k\tau}, \frac{1}{\tau_0 + m_k\tau}\right) \tag{2.7}$$

where $\sum_{i=1}^{N} y_i\delta_{K_i,k}$ is the sum of the data allocated to component k, τ is the known precision parameter for the Normal components and μ_0 and τ_0 are the mean and precisions of the Normal prior distribution. However this only allows us to learn about the mean of the Normal components k. If we want to learn about both the mean and variance we can use a Normal-Gamma conjugate prior distribution for G_0:

$$\begin{aligned}\pi(\mu, \tau) &= \text{Normal-Gamma}(\mu, \tau|\mu_0, \kappa, \alpha, \beta) \\ &= \mathcal{N}(\mu|\mu_0, (\kappa\tau)^{-1})\text{Gamma}(\tau|\alpha, \beta)\end{aligned}$$

Using the m_k observations, y, assigned to component k, the posterior is given by:

$$\begin{aligned}p(\mu_k, \tau_k|\mathbf{K}, y) \propto \mathcal{N}&\left(\mu; \frac{\kappa\mu_0 + \sum_{i=1}^{N} y_i\delta_{K_i,k}}{\kappa + m_k}, \frac{1}{(\kappa + m_k)\tau}\right) \times \\ &\text{Gamma}\left(\tau; \alpha + \frac{m_k}{2}, \beta + \frac{1}{2}\sum_{i=1}^{N}(y_i - \bar{y}_k)^2 + \frac{\kappa m_k(\bar{y}_k - \mu_0)^2}{2(\kappa + m_k)}\right)\end{aligned}$$

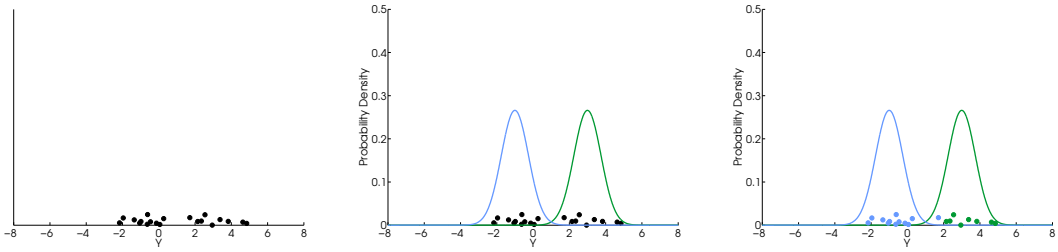
where $\bar{y}_k = m_k^{-1} \sum_{i=1}^{N} y_i \delta_{K_i,k}$. We can illustrate the Bayesian inference steps for the Dirichlet mixture of Normal distributions model more intuitively in Figure 2.2. Figure 2.2a shows a sample obtained from a distribution that we will be modelling using a Dirichlet mixture of Normal distributions. For graphical purposes, we have limited the number of components in the mixture to two. In Figure 2.2b, we have sampled $\boldsymbol{\theta}$ which define the locations of the two Normal components, $p(\boldsymbol{\theta}|\mathbf{y}, \mathbf{K})$. In Figure 2.2c, data are assigned to the Normal distributions in the mixture according to $p(K_i|\mathbf{w}, \boldsymbol{\theta})$. The green dots have been assigned to the green distribution and similarly the blue dots have been assigned to the blue distribution by the model. The last steps are to assign weights, $\mathbf{w}$, to the components (Figure 2.2d) and to update the component parameters. Repeating steps b to d twice will lead to two other realisations (Figures 2.2e and 2.2f). Figure 2.2g shows four realisations from the posterior Dirichlet mixture distribution. This is then summarised by calculating a median and a 95% credible interval, shown in Figure 2.2h.

Figure 2.2 – *Graphical representation of updating the parameters of a mixture model with two Normal components.*

(a) Data

(b) Two Normal components with location parameters θ_1 and θ_2 and equal weights $w_1 = w_2 = 0.5$

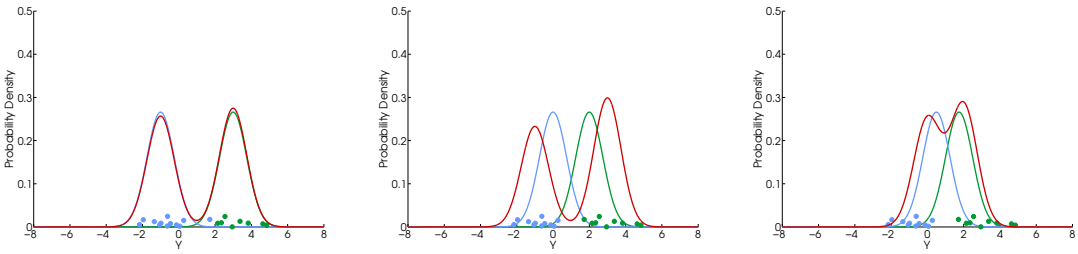
(c) Allocation of each data value to one of the two components



(d) Model output (red line) after weights have been updated.

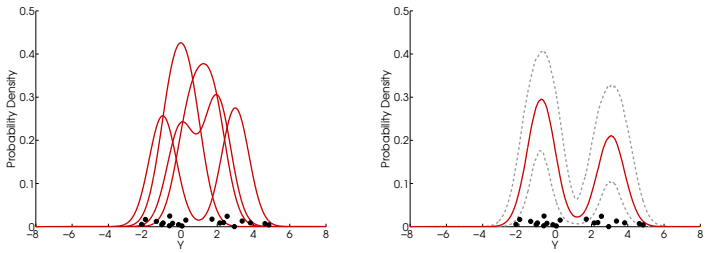
(e) Component parameters are updated using the data assigned to each component.

(f) Another iteration of the Monte Carlo approach.



(g) Model outputs after four iterations.

(h) The mixture distribution can be summarised using the median and a 95% credible interval.



2.2.6.2 Clustering

The use of Dirichlet distributions in mixture models can be extended to mixtures of multivariate distributions. The probability density function for $\mathbf{x}$ of a d-dimensional multivariate Normal distribution is:

$$\frac{1}{2\pi^{\frac{d}{2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}}\exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})^T\right]$$

where $\boldsymbol{\mu}$ is the mean and $\boldsymbol{\Sigma}$ is the covariance matrix. In a Bayesian setting, we can obtain samples from the posterior distribution $p(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\mathbf{x})$ using Gibbs sampling if we select conjugate prior distributions. The conjugate prior distribution of a multivariate Normal distribution with mean vector $\boldsymbol{\mu}$ and precision matrix $\boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{-1}$ is the Normal-Wishart distribution $\mathcal{NW}(\mu_0, \kappa_0, \Psi_0, \nu_0)$:

$$\begin{aligned}\pi(\mu,\sigma) &= \mathcal{N}\left(\boldsymbol{\mu}|\boldsymbol{\mu_0}, (\kappa_0\boldsymbol{\Lambda})^{-1}\right)\mathcal{W}(\boldsymbol{\Lambda}|\boldsymbol{\Psi_0}, \nu_0)\\ &= \frac{1}{\mathbf{Z}}|\boldsymbol{\Lambda}|^{\frac{1}{2}}\exp\left[-\frac{\kappa_0}{2}(\boldsymbol{\mu}-\boldsymbol{\mu_0})\boldsymbol{\Lambda}(\boldsymbol{\mu}-\boldsymbol{\mu_0})^T\right]|\boldsymbol{\Lambda}|^{\frac{\kappa_0-d-1}{2}}\exp\left[-\frac{1}{2}\mathrm{tr}(\boldsymbol{\Psi_0}^{-1}\boldsymbol{\Lambda})\right]\end{aligned}$$

where $\boldsymbol{\Psi_0}$ is the prior precision matrix, $\mathbf{Z} = \left(\frac{\kappa_0}{2\pi}\right)^{\frac{d}{2}}|\boldsymbol{\Psi_0}|^{\frac{\kappa_0}{2}}2^{\frac{d\kappa}{2}}\Gamma_d(\frac{\kappa_0}{2})$, $\Gamma_d(x)$ is the multivariate Gamma function and d is the number of dimensions, i.e. $d = 2$ for a bivariate Normal distribution. The likelihood function for n observations $\mathbf{x}$ is:

$$(2\pi)^{\frac{-dn}{2}}|\boldsymbol{\Lambda}|^{\frac{n}{2}}\exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})\boldsymbol{\Lambda}(\mathbf{x}-\boldsymbol{\mu})^T\right]$$

DeGroot (1970) shows that the posterior is a $\mathcal{NW}(\frac{\kappa_0\boldsymbol{\mu_0}+n\bar{\mathbf{x}}}{\kappa_0+n}, \kappa_0+n, \mathbf{Q}, \nu_0+n)$ distribution, where $\mathbf{Q} = \boldsymbol{\Psi_0^{-1}} + \frac{\kappa_0 n}{\kappa_0+n}(\boldsymbol{\mu_0}-\bar{\mathbf{x}})(\boldsymbol{\mu_0}-\bar{\mathbf{x}})^T + \sum_{i=1}^{n}(\mathbf{x_i}-\bar{\mathbf{x}})(\mathbf{x_i}-\bar{\mathbf{x}})^T$.
We can generate samples from the posterior distribution by sampling $\boldsymbol{\Lambda}$ from a Wishart distribution with parameters $\mathbf{Q}^{-1}$ and ν_0+n and by sampling $\boldsymbol{\mu}$ from a multivariate Normal distribution with mean $\frac{\kappa_0\boldsymbol{\mu_0}+n\bar{\mathbf{x}}}{\kappa_0+n}$ and covariance matrix $((\kappa_0+n)\boldsymbol{\Lambda})^{-1}$.

2.3 Dirichlet Process

In Section 2.2, we showed that the Dirichlet distribution can be used to describe the probability of observing an event when the number of events is finite. In a Bayesian context, we can use the Dirichlet distribution to describe our prior and posterior

beliefs, i.e. uncertainty, about the probabilities associated with a particular event. For example we could use the Dirichlet distribution to describe the variation in people's favourite colours in a finite set of colours (e.g. red, yellow, green and blue). However, if we do not want to restrict the colours to a finite set (e.g. allow for any colour in the infinite RGB colour space) we need a distribution over an infinite sample space. The Dirichlet Process, $DP(\gamma, G_0)$, is a stochastic process that is a distribution over probability measures whose domain is defined by its base measure G_0. As a DP is a stochastic process, it can be used to generate an infinite sequence of random variables, $\boldsymbol{\phi}$. Random variables can be generated from a DP by first drawing a random distribution G from the DP. Next, an infinite sequence of random variables or observations, $\boldsymbol{\phi}$, can be drawn from G. Conditional on G, the variables are independent and identically distributed:

$$\begin{aligned} G &\sim \mathrm{DP}(\gamma, G_0) \\ \boldsymbol{\phi} &\sim G \end{aligned} \tag{2.8}$$

Probability measures G drawn from a Dirichlet Process are discrete and cannot be described using a finite number of parameters. As a consequence, models that are based on DPs are considered to be non-parametric models. The concentration parameter γ is a measure of the likelihood of repeated values in G.

2.3.1 Formal definition

Ferguson (1973) was the first to show the existence of a DP when he introduced it to solve the problem of finding a workable prior distribution which allows Bayesian approaches to be used in non-parametric settings. The DP provides a class of prior distributions which has two desirable properties: it has the same support as G_0 and it leads to a posterior distribution that is manageable analytically. Let G_0 be a probability distribution on a measurable space Φ and let γ be a positive scalar. Consider a finite partition $(A_1, \ldots, A_K)$ of Φ.

$$\bigcup_{k=1}^{K} A_k = \Phi \qquad A_k \cap A_l = \emptyset \qquad k \neq l$$

A random probability measure G on Φ is drawn from a DP if for every measurable partition $(A_1, \ldots, A_K)$, the random vector $(G(A_1), \ldots, G(A_K))$ follows a Dirichlet distribution:

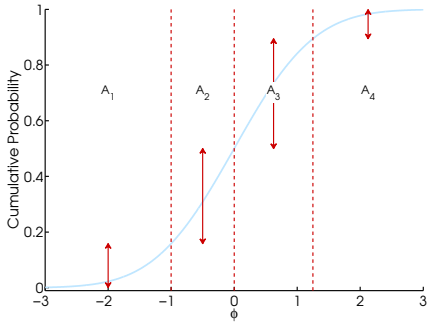
$$(G(A_1), \ldots, G(A_K)) \sim \text{Dirichlet}\,(\gamma G_0(A_1), \ldots, \gamma G_0(A_K)) \tag{2.9}$$

Samples from a DP are discrete with probability one (Sethuraman, 1994). Figure 2.3 provides a graphical overview of a DP with base measure G_0 and concentration parameter γ. Figure 2.3a shows how the parameter space Φ can be split into four parts. The arrows indicate the probability of observing a value in each part A_k. Given γ, we can generate samples of G, shown in Figures 2.3b and 2.3c. The weight that a random measure $G \sim \text{DP}(\gamma, G_0)$ assigns to each part follows a Dirichlet distribution. Note that γ determines the deviation of samples G from G_0: the smaller γ is the larger the variation in G. For large γ the samples of G reflect the probabilities as indicated by the arrows in Figure 2.3a. Figure 2.3d shows that because of the aggregation property of the Dirichlet distribution (see Section 2.2.2.1), all possible partitions are consistent.

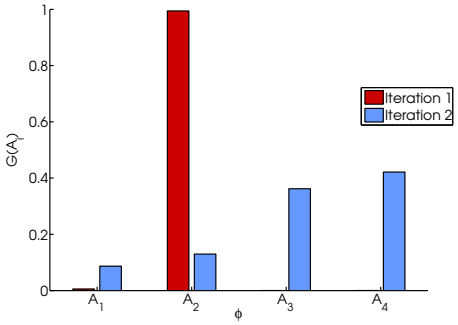
In a Bayesian setting, the concentration parameter can be referred to as a strength parameter, as it determines the strength of the prior distribution when using a DP as a non-parametric prior distribution over distributions. In that setting, its value can be considered as the sample size (or mass) of prior observations. Although a small value for γ implies little strength of the prior distribution it asserts that most of the probability is on a single point. For large γ and as the number of partitions increase, G provides a discrete approximation of G_0, whereas for small γ the uncertainty around G_0 is larger.

Figure 2.3 – *Graphical overview of a DP(γ, G_0) using a finite partition of the parameter space Φ.*

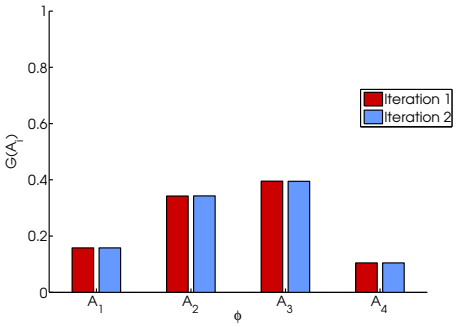
(a) Four partitions of the 1-dimensional parameter space Φ with the cumulative distribution function of example base measure G_0 (blue)



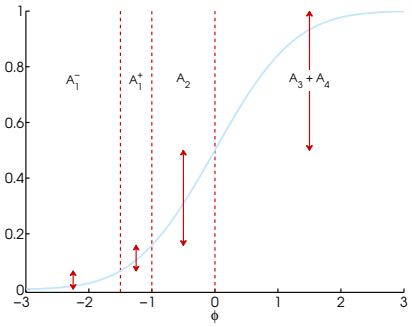
(b) Two samples of G with $\gamma = 1$.



(c) Two samples of G with $\gamma = 1,000,000$.



(d) An alternative partitioning that shows the consistency of G: Partition A_1 is split into two parts and the last two parts (A_3 and A_4) are merged.



2.3.2 Properties of a Dirichlet Process

2.3.2.1 Expected Distribution

The Dirichlet Process is a distribution of distributions and therefore the expectation is a distribution. The expectation of a DP can be obtained using Equations 2.3 (page

59) and 2.9:

$$\begin{aligned}\mathbb{E}[G](A_i) = \mathbb{E}[G(A_i)] &= \frac{\gamma G_0(A_i)}{\sum_{k=1}^{\infty} \gamma G_0(A_k)} \\ &= \frac{\gamma G_0(A_i)}{\gamma} = G_0(A_i)\end{aligned}$$

as $\sum_{k=1}^{\infty} G_0(A_k) = 1$. It is interesting to note that despite the fact that any realisation of G is discrete, the expectation is a continuous distribution if G_0 is continuous.

2.3.2.2 Variance

The variance of a DP can be obtained using Equations 2.4 (page 59) and 2.9:

$$\text{Var}[G(A_i)] = \frac{G_0(A_i)(1 - G_0(A_i))}{(\gamma + 1)}$$

For large values of γ, the variance is small whereas for small values of γ, the variance may be large as shown in Figure 2.3b.

2.3.2.3 Exchangeability

The predictive distribution after observing samples $\boldsymbol{\phi}$ from a DP (see Section 2.3.4.2 for details) is given by:

$$p(\phi_{N+1} \in A_k | \phi_1, \ldots, \phi_N) = \frac{1}{\gamma + N} \left(\gamma G_0(A_k) + \sum_{i=1}^{N} \delta_{\phi_i \in A_k} \right)$$

Using the predictive distribution of a DP we can iteratively draw a sequence $\phi_1, \ldots, \phi_n$. This results in the joint distribution:

$$p(\phi_1, \ldots, \phi_N) = p(\phi_1) \prod_{i=2}^{N} p(\phi_i | \phi_1, \ldots, \phi_{i-1})$$

De Finetti's theorem (De Finetti, 1931) states that a sequence $\phi_1, \ldots, \phi_N$ of random variables is exchangeable if and only if there exists a distribution function P on $[0, 1]$ such that for all i, the joint distribution can be represented as a mixture:

$$p(\phi_1, \ldots, \phi_N) = \int_0^1 \left[\prod_{i=1}^{N} G(\phi_i) \right] dP(G)$$

For a DP, using Equation 2.8 (page 72) we know that the ϕ_i are exchangeable because $P(G) = \text{DP}(\gamma, G_0)$.

2.3.2.4 Discreteness

Another property of a DP that is apparent from the predictive distribution is that it has point masses at previously observed draws $\phi_1, \ldots, \phi_N$. The predictive distribution also shows that with non-zero probability ($\frac{n_k}{N+\gamma}$, where n_k is the number of times a value has been observed in $\boldsymbol{\phi}$ and N is the total number of observations) new draws will take on the same value as previously observed draws, regardless of the distribution G_0. As any value of ϕ will be repeated given a long enough sequence of draws, G is a discrete distribution.

2.3.2.5 Clustering

We can use the predictive distribution after observing samples, $\boldsymbol{\phi}$, from a DP to derive another property of a DP. The fact that values of $\boldsymbol{\phi}$ are repeated implies that DPs have a clustering property that is essential for the use of DPs in infinite mixture models. Infinite mixture models assume that observations come from a mixture of an infinite number of distributions. Note that N observations still come from at most N different distributions, which shows that many components will have no data associated with them. If we draw from a DP mixture model, we would expect a clustering of the $\boldsymbol{\phi}$, i.e. multiple observations y_i are expected to come from the same component. In contrast, if $\boldsymbol{\phi}$ were drawn from a Gaussian distribution, no two values would be the same.

2.3.2.6 Effect of G_0

The base measure G_0 determines the support of the distribution and can be interpreted in a Bayesian context as an expression of one's prior beliefs of the distribution of the variable that is modelled by a DP. In DP mixture models, the selection of G_0 is often determined by mathematical convenience as conjugate distributions will facilitate a simple updating step for the parameters of the mixture components. Both conjugate and non-conjugate base measures have been used extensively (Escobar, 1994; Escobar and West, 1995; MacEachern and Müller, 1998). However, Görür and Rasmussen (2010) suggested that the choice of a conjugate G_0 may affect the num-

ber of components being utilised, so care must be taken when selecting G_0 when using DP mixture models for clustering purposes.

2.3.2.7 Effect of γ

The concentration parameter γ expresses the strength of belief in the base measure G_0. As we observed in Section 2.3.2.2, γ has an effect on the variance of a DP. For small values of γ, samples from a DP are likely to consist of samples that have the same value, whereas for large values of γ, samples from a DP are likely to be distinct, similar to samples from G_0. For each draw ϕ, taken from a $\text{DP}(\gamma, G_0)$, the probability of observing a new, distinct value of ϕ is $\frac{\gamma}{\gamma+\sum_i n_i}$, where n_i indicates the number of times a distinct value ϕ_i has been observed before (see Section 2.3.3.2). The probability of a new draw taking on the same value as previously observed draws, $\phi_{N+1} \in \{\phi_1, \ldots, \phi_N\}$, is $\frac{n_i}{\gamma+\sum_i n_i}$. The expected number of distinct ϕ values, C, after N draws from a $\text{DP}(\gamma, G_0)$, is given by Antoniak (1974):

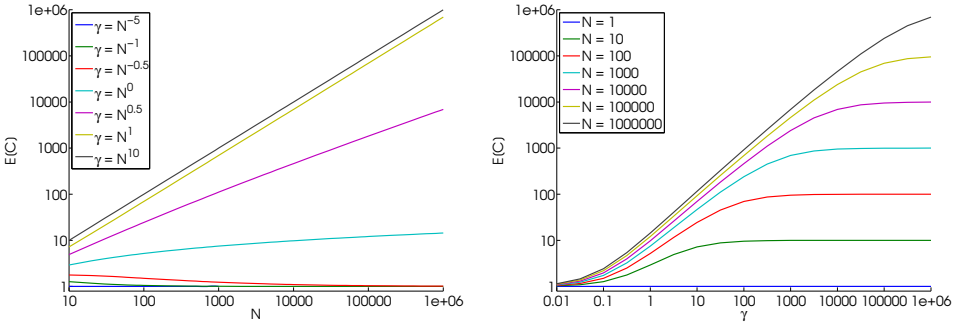
$$\begin{aligned}\mathbb{E}[C(\gamma, N)] &= \sum_{i=1}^{N} \frac{\gamma}{\gamma + i - 1} \\ &\approx \gamma \log\left(1 + \frac{N}{\gamma}\right)\end{aligned}$$

So if we set $\gamma = 10$ and we generate 20,000 values from a DP, $\mathbb{E}[C] = 77$, i.e. we would expect 77 distinct values of ϕ. Figures 2.4a and 2.4b show the expected number of components C as a function of the number of random variables, N, sampled from the DP and γ. C grows logarithmically as a function of the number of samples, which demonstrates the discreteness and clustering properties of DPs. For large γ, $C = N$. If we look at the frequency of components (Figures 2.4c and 2.4d), we notice that this approaches N for small γ and 1 for large γ, indicating that in the latter case, the distribution of $\boldsymbol{\phi}$ will be a discrete approximation of the continuous distribution G_0.

Figure 2.4 – *Expected number of clusters C as a function of the sample size N for various values of γ.*

(a) Number of components as a function of N

(b) Number of components as a function of γ



(c) Frequency of components as a function of N

(d) Frequency of components as a function of γ

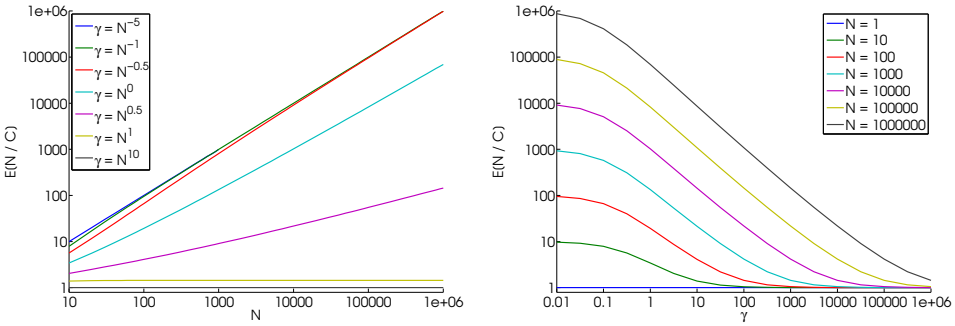


Figure 2.5 shows the effect of γ on samples obtained from a DP(γ, G_0) for various values of γ and $G_0 = \mathcal{N}(0, 1)$. We generated three sets of samples to illustrate that the variation of the DP is large for small γ. As $\gamma \to \infty$, the distributions G become closer to G_0. However this does not mean that $G \to G_0$ as G is discrete. To generate a continuous distribution, we need to extend the DP by convolving G with a continuous distribution $f(\boldsymbol{\phi})$ with latent parameter $\boldsymbol{\phi}$ so that the resulting random distribution is continuous. This is analogous to the finite Dirichlet mixture distributions discussed in Section 2.2.6.1. If we compare the behaviour of a DP with the Dirichlet distribution, we notice that in both cases we would observe repeated values. The Dirichlet distribution is used to assign probabilities to a finite number of categories, e.g. the sides of a die (1, 2, 3, 4, 5, 6). For the DP, the number of

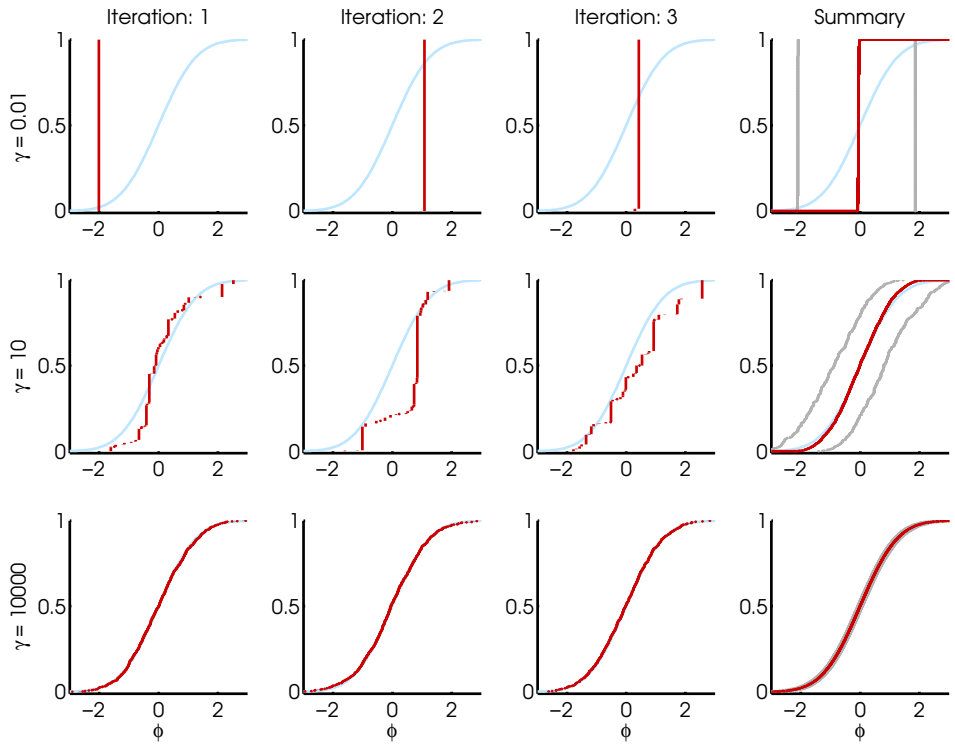


Figure 2.5 – *Effect of γ on samples, G, (red) obtained from a $DP(\gamma, G_0)$ for various values of γ and $G_0 = \mathcal{N}(0,1)$ (blue). The summary graphs show the median cumulative distribution function (red) and a 95% credible interval (grey).*

categories is infinite, so we need to assign probabilities to categories in a different way. In the Dirichlet distribution case, we assign prior weights $(\gamma_1, \ldots, \gamma_6)$ to each category, where $\gamma_0 = \sum_{i=1}^{6} \gamma_i$ is a measure for how certain we are about the relative weights. In the DP case we assign a value to γ whose value is again a measure for how certain we are about the relative probabilities $G(\phi)$ assigned by G_0.

2.3.3 Generating observations from a Dirichlet Process

In this section we discuss various methods to generate observations, ϕ, from a Dirichlet Process using three representations. In Chapters 4 and 5 we make use of the stick-breaking approach to infer the weights of components in a DP mixture model. However, we discuss the other representations of a DP for completeness.

2.3.3.1 Stick-Breaking Approach

In the stick-breaking representation of a DP, the main idea is that a stick of unit length will be broken into pieces. In contrast to the Dirichlet distribution, however, we do not specify how many pieces. This construction of the DP (Sethuraman, 1994) offers a mechanism for sampling from a DP. Sethuraman (1994) provided a constructive definition of the DP that is based on the observation that draws from a DP are composed of a weighted sum of point masses. Let $w = \{w_1, w_2, \dots\}$ be an infinite set of mixture weights that can be derived from the following stick-breaking process with parameter $\gamma > 0$:

$$\begin{aligned} \beta_k &\sim \text{Beta}(1, \gamma) \qquad\qquad i.i.d. \\ w_k &= \beta_k \prod_{j=1}^{k-1} (1 - \beta_j) \end{aligned}$$

Given base measure G_0, Sethuraman (1994) derived that the following random measure guarantees that $G \sim DP(\gamma, G_0)$:

$$\begin{aligned} \phi_k &\sim G_0 \qquad\qquad i.i.d. \\ G(\phi) &= \sum_{k=1}^{\infty} w_k \delta_{\phi_k} \end{aligned}$$

The non-trivial proof in Sethuraman (1994) is based on the observation that the following approaches are equivalent:

$$\begin{array}{ll} G \sim DP(\gamma, G_0) & \boldsymbol{\phi} \sim G_0 \\ \boldsymbol{\phi} | G \sim G & G | \boldsymbol{\phi} \sim DP(\gamma + 1, \frac{\gamma G_0 + \delta_\phi}{\gamma + 1}) \end{array} \quad \Leftrightarrow$$

After we have observed a sample ϕ from G_0, we can partition the parameter space Φ in two: $\{\phi, \Phi \setminus \phi\}$. This leads to:

$$\begin{aligned} (G|\phi(\phi), G|\phi(\Phi \setminus \phi)) &\sim \text{Dirichlet}\left(\gamma G_0(\phi) + 1, \gamma G_0(\Phi \setminus \phi)\right) \\ &\approx \text{Dirichlet}(1, \gamma) \end{aligned}$$

as $G_0(\phi) \approx 0$ and $G_0(\Phi \setminus \phi) \approx 1$. So $G|\phi$ has a point mass w located at ϕ:

$$G|\phi = w\delta_{\phi_1} + (1 - w)G' \quad \text{with } w \sim \text{Beta}(1, \gamma) \tag{2.10}$$

where G' is the renormalised probability measure after removing point mass w. Using the aggregation property, we can partition Φ further into $\{\phi, A_1, \ldots, A_k\}$:

$$(G|\phi(\phi), G|\phi(A_1), \ldots, G|\phi(A_k)) \sim \text{Dirichlet}\,(1, \gamma G_0(A_1), \ldots, \gamma G_0(A_k))$$

We know that:

$$(G|\phi(\phi), G|\phi(\Phi \setminus \phi)) = (w, (1-w)G'_0(A_1), \ldots, (1-w)G'_0(A_k))$$

This leads to:

$$(w, (1-w)G'_0(A_1), \ldots, (1-w)G'_0(A_k)) \sim \text{Dirichlet}\,(1, \gamma G_0(A_1), \ldots, \gamma G_0(A_k))$$

For notational convenience, let $h \sim \text{Gamma}(1, 1)$ and $g_i \sim \text{Gamma}\,(\gamma G_0(A_i), 1)$. We can write the Dirichlet random variable in terms of a normalised set of independent Gamma random variables:

$$\begin{aligned} w &= \frac{h}{h + \sum_{j=1}^{k} g_j} \\ (1-w)G'(A_i) &= \frac{g_i}{h + \sum_{j=1}^{k} g_j} \\ G'(A_i) &= \frac{g_i}{(h + \sum_{j=1}^{k} g_j)(1 - \frac{h}{h + \sum_{j=1}^{k} g_j})} \\ &= \frac{g_i}{\sum_{j=1}^{k} g_j} \end{aligned}$$

So $G'(A_i)$ is a normalised set of independent Gamma random variables which we know to be equal to a Dirichlet distribution:

$$(G'(A_1), \ldots, G'(A_k)) \sim \text{Dirichlet}\,(\gamma G_0(A_1), \ldots, \gamma G_0(A_k))$$

Based on Equation 2.9 (page 73), this implies that:

$$G' \sim \text{DP}(\gamma, G_0)$$

So with $G' \sim \text{DP}(\gamma, G_0)$, Equation 2.10 can be rewritten as:

$$G|\phi = w\delta_\phi + (1-w)DP(\gamma, G_0)$$

Recursively applying this, leads to:

$$\begin{aligned} G|\phi_1 &= w_1\delta_{\phi_1} + (1-w_1)DP(\gamma, G_0) \\ G|\phi_1, \phi_2 &= w_1\delta_{\phi_1} + w_2\delta_{\phi_2} + (1-w_1-w_2)DP(\gamma, G_0) \\ &\vdots \\ G|\phi_1, \ldots, \phi_n, \ldots &= \sum_{k=1}^{\infty} w_k \delta_{\phi_k} \end{aligned}$$

with $w_k = \beta_k \left(1 - \sum_{j=1}^{k-1} w_k\right) = \beta_k \prod_{j=1}^{k-1}(1-\beta_j)$. As $k-1 \to \infty$, this almost surely goes to zero with probability one. To prove this we have $\prod_{j=1}^{k-1}(1-\beta_j) = 0 \iff \sum_{j=1}^{\infty}\beta_j = \infty$ (Folland, 1999). For any constant $\epsilon \in (0,1)$, $\sum_{j=1}^{\infty} P([\beta_j > \epsilon]) = \infty$ so using the second Borel-Cantelli lemma we obtain $P([\beta_j > \epsilon], i.o.) = 1$. In other words, for an infinite sequence of events $[\beta_j > \epsilon]$ for which the sum of the probabilities goes to ∞, the probability of observing infinitely many of the events is one and therefore $\sum_{j=1}^{\infty}\beta_j = \infty$ almost surely. As a result $G|\boldsymbol{\phi} = \sum_{k=1}^{\infty} w_k\delta_{\phi_k}$ is a valid probability measure with $\phi_k \sim G_0$. So we have:

$$\begin{aligned} w_k &= \beta_k \left(1 - \sum_{j=1}^{k-1} w_k\right) \\ &= \beta_k \prod_{j=1}^{k-1}(1-\beta_j) \\ \beta_k &\sim \text{Beta}(1, \gamma) \\ \boldsymbol{\phi} &\sim G_0 \\ G|\boldsymbol{\phi} &= \sum_{k=1}^{\infty} w_k \delta_{\phi_k} \end{aligned}$$

which is the stick-breaking scheme. It is important to emphasise the 'symmetry-breaking' nature of the stick-breaking representation as the weights, w_i, obtained via this approach are size-biased towards large values of w for small i. To overcome the dependency of weights on label i, label-swapping approaches have been proposed which will be discussed in Section 2.3.5.3.

We can calculate the expected weight $\mathbb{E}[w_k]$ of cluster k in the stick-breaking scheme using the fact that $\mathbb{E}[\beta_j] = \frac{1}{1+\gamma}$. Given that the β_j are independent ($\mathbb{E}[XY] =$

$\mathbb{E}[X]\mathbb{E}[Y]$ for two independent random variables X and Y), the expected value for component k is given by:

$$\begin{aligned}\mathbb{E}[w_k] &= \mathbb{E}[\beta_k]\prod_{j=1}^{k-1}\mathbb{E}[(1-\beta_j)] \\ &= \frac{1}{1+\gamma}\prod_{j=1}^{k-1}\frac{\gamma}{1+\gamma} \\ &= \frac{1}{\gamma}\left(\frac{\gamma}{1+\gamma}\right)^k \end{aligned} \tag{2.11}$$

The variance of component weight w_k is given by:

$$\begin{aligned}\text{Var}[w_k] &= E[w_k^2] - E[w_k]^2 \\ &= \frac{2}{\gamma(\gamma+1)}\left(\frac{\gamma}{\gamma+2}\right)^k - \frac{1}{\gamma^2}\left(\frac{\gamma}{1+\gamma}\right)^{2k}\end{aligned}$$

The first term can be derived as follows:

$$\begin{aligned}\mathbb{E}[w_k^2] &= \mathbb{E}\left[\beta_k^2\prod_{j=1}^{k-1}(1-\beta_j)^2\right] \\ &= \mathbb{E}[\beta_k^2]\left\{\mathbb{E}[(1-\beta_k)^2]\right\}^{k-1} \qquad \text{as } \beta_j \text{ are i.i.d.} \\ \beta_j &= \text{Beta}(1,\gamma) \\ 1-\beta_j &= \text{Beta}(\gamma,1) \\ \mathbb{E}[(1-\beta_k)^2] &= \frac{\Gamma(\gamma+1)}{\Gamma(\gamma)}\int_0^1(1-\beta_k)^2(1-\beta_k)^{\gamma-1}d(1-\beta_k) = \frac{\Gamma(\gamma+1)}{\Gamma(\gamma)}\frac{1}{\gamma+2}\end{aligned}$$

Given that $\Gamma(\alpha+1) = \alpha\Gamma(\alpha)$, we get

$$\mathbb{E}[(1-\beta_j)^2] = \frac{\gamma\Gamma(\gamma)}{\Gamma(\gamma)}\frac{1}{(\gamma+2)} = \frac{\gamma}{\gamma+2}$$

For $\mathbb{E}[\beta_k^2]$ we have:

$$\begin{aligned}\mathbb{E}[\beta_k^2] &= \frac{\Gamma(\gamma+1)}{\Gamma(\gamma)}\int_0^1\beta_k^2(1-\beta_k)^{\gamma-1}d\beta_k \\ &= \frac{\Gamma(\gamma+1)}{\Gamma(\gamma)}\frac{\Gamma(3)\Gamma(\gamma)}{\Gamma(\gamma+3)} = \frac{2}{(\gamma+1)(\gamma+2)}\end{aligned}$$

leading to:

$$\mathbb{E}[w_k^2] = \left(\frac{\gamma}{\gamma+2}\right)^{k-1}\frac{2}{(\gamma+2)(\gamma+1)} = \frac{2}{\gamma(\gamma+1)}\left(\frac{\gamma}{\gamma+2}\right)^k$$

The second term follows from Equation 2.11.

2.3.3.2 Pólya Urn Representation

The Pólya urn representation (Blackwell and MacQueen, 1973), sometimes referred to as the Blackwell-MacQueen urn provides an alternative technique for generating samples from a DP with G_0 representing a distribution over an unlimited number of colours. In this scheme, we start with an empty urn and we draw a colour from G_0 and paint ball ϕ_1 with that colour. For each subsequent draw, we either add a ball ϕ_{N+1} with a new colour sampled from G_0 with probability $\frac{\gamma}{\gamma+\sum_{j=1}^{N}\delta_\phi}$ or draw a ball from the urn with probability $\frac{N}{\gamma+N}$ where N is the number of previous draws, paint a new ball with the same colour and put both balls back in the urn. If we have observed n_c balls of colour c, we will draw a ball of that same colour with probability $\frac{\sum_{j=1}^{N}\delta_{\phi_j,c}}{\gamma+N}$ where $\delta_{\phi_j,c}$ is Kronecker's delta, which equals one if ball ϕ_j has a previously observed colour c. It is clear that for large γ the probability of drawing a new colour is larger than drawing a ball of colour c and therefore, we will end up with many different coloured balls in the urn. If γ is small, we see the opposite happening: if we assume that a blue ball is added to the urn in the first draw, the probability of drawing another blue ball from the urn in the next draw is $\frac{1}{\gamma+1} \approx 1$, whereas the probability of picking a new colour is $\frac{\gamma}{\gamma+1} \approx 0$, as $\gamma \to 0$. As a result, the first colour sampled from G_0 will dominate the sample and repeating the whole exercise many times will lead to urns dominated by a single colour. The balls $\boldsymbol{\phi}$ of an infinite sequence of draws from the Pólya urn scheme follow the same distribution as observations obtained from a DP.

Chinese Restaurant Process

The Chinese restaurant representation is very similar to the Pólya urn representation. Consider a Chinese restaurant with an infinite number of tables. For each customer arriving in the restaurant, there are two options: the customer either sits at a new table and selects a meal from distribution G_0 or (s)he joins a table that is already in use and is assigned the same meal as the other customers at that table. If n_i is the number of customers sitting at table i and $N = \sum_i n_i$ is the number of customers already present in the restaurant, the probability of customer $N+1$ joining a previously occupied table i is $\frac{n_i}{\gamma+N}$ and the probability that (s)he sits at

the next unoccupied table is $\frac{\gamma}{\gamma+N}$. Small values for γ indicate that most customers end up at the same table, just as a small value for γ in the Pólya urn representation leads to many balls having the same colour.

2.3.4 Bayesian Inference for a Dirichlet Process

2.3.4.1 Posterior distribution

Let $G \sim \mathrm{DP}(\gamma, G_0)$ and $\phi \sim G$ where $\phi \in A_k$. We can use the conjugacy of the Dirichlet distribution to obtain the posterior distribution:

$$p(G(A_1), \ldots, G(A_K))|\phi \in A_k, \gamma, G_0) = \mathrm{Dirichlet}(\gamma G_0(A_1), \ldots, \gamma G_0(A_k)+1, \ldots, \gamma G_0(A_K))$$

The observation ϕ only affects the parameter of the Dirichlet distribution for part k. For $K \to \infty$, the posterior distribution has a point mass centered on each observation. Extending to multiple observations leads to:

$$\begin{aligned} p(G(A_1), \ldots, G(A_K)|\phi \in A_k, \gamma, G_0) &= \mathrm{Dirichlet}(\gamma G_0(A_1) + n_1, \ldots, \gamma G_0(A_K) + n_K) \\ &= \mathrm{DP}\left(\gamma + N, \frac{1}{\gamma + N}\left(\gamma G_0 + \sum_{i=1}^{N} \delta_{\phi_i \in A_k}\right)\right) \end{aligned} \tag{2.12}$$

So analogously to the Dirichlet distribution, whose posterior after observing Multinomial data is a Dirichlet distribution, the posterior of a DP is a DP itself. DPs are characterised by their neutrality with respect to every partition. This means that the posterior distribution $p(G(A_k)|\phi)$ depends only on the number of observations that fall within A_k, regardless of the locations of ϕ within A_k. Observations near boundaries provide the same amount of information as observations in the centre. The expected value of the posterior distribution is given by:

$$\mathbb{E}[G(A_k)|\phi, \gamma, G_0] = \frac{1}{\gamma + N}\left(\gamma G_0 + \sum_{i=1}^{N} \delta_{\phi_i \in A_k}\right)$$

For finite values of γ this leads to:

$$\lim_{N \to \infty} \mathbb{E}[G(A_k)|\phi, \gamma, G_0] = \frac{\sum_{i=1}^{N} \delta_{\phi_i \in A_k}}{N} = \sum_{k=1}^{\infty} w_k \delta_{\phi_k}$$

where w_k is the limiting frequency of the unique ϕ_k values. This implies that $\mathbb{E}[G(A_k)|\phi, \gamma, G_0]$ is a discrete measure, which in turn implies that $p(G(A_k)|\phi, \gamma, G_0)$ is discrete as well. This alternative representation of a DP was introduced by Theorem 2 in Ferguson (1973).

2.3.4.2 Predictive distribution

The predictive distribution for observations from a DP $p(\phi_{N+1}|\phi_1, \ldots, \phi_N)$ can be obtained using:

$$\begin{aligned} p(\phi_{N+1} \in A_k|\phi_1, \ldots, \phi_N) &= \int_G p(\phi_{N+1}|G)p(G|\phi_1, \ldots, \phi_N)dG \\ &= \int_G G\, p(G|\phi_1, \ldots, \phi_N)dG \\ &= \mathbb{E}[G(A_k)|\phi_1, \ldots, \phi_N] \\ &= \frac{1}{\gamma + N}\left(\gamma G_0(A_k) + \sum_{i=1}^{N} \delta_{\phi_i \in A_k}\right) \end{aligned}$$

as $p(\phi_{N+1}|G) \sim G$. By integrating out G, all the random variables $\boldsymbol{\phi}$ become identically distributed but not independent. Two common metaphors which are used in the interpretations of this result are the Pólya Urn scheme and the Chinese Restaurant Process (see Section 2.3.3.2). This follows from the observation that we draw a new value from G_0 with probability $\frac{\gamma}{\gamma+N}$ and a previously observed value ϕ_i with probability $\frac{n_i}{\gamma+N}$, where n_i indicates the number of times we have observed ϕ_i in the previous N observations. Given that $p(\phi_1, \ldots, \phi_N) = p(\phi_1)p(\phi_2|\phi_1)\ldots p(\phi_N|\phi_1, \ldots, \phi_{N-1})$, this predictive distribution can be used to generate samples from a DP.

2.3.5 Applications of a Dirichlet Process

2.3.5.1 Infinite Mixture Models

The Dirichlet distribution is ideal to model distributions of distributions in finite mixture problems whereas a Dirichlet Process can be used for infinite mixture problems, where the number of components in the mixture is unlimited (Antoniak, 1974;

Escobar and West, 1995; Lo, 1984). Using DP mixture models overcomes the problem that finite Dirichlet mixture models require a specification of the number of components that will be used to model the data.
The DP mixture model is given by:

$$\begin{aligned} y_i|\phi_i &\sim f(\cdot|\phi_i) && \textit{independently} \\ \phi_i|G &\sim G && \textit{i.i.d.} \\ G &\sim DP(\gamma, G_0) \end{aligned}$$

which can also be written as the limit of a finite model (Equation 2.5 on page 67) where the number of components C goes to infinity. A common way of describing a DP mixture model is based on the stick-breaking approach (see Section 2.3.3.1) and the fact that we can use the random variables $\boldsymbol{\theta}$, the set of unique $\boldsymbol{\phi}$ values, as the parameters of a continuous kernel (e.g. a Normal distribution) which is used to smooth out the discrete draws from the DP:

$$\begin{aligned} y_i|\boldsymbol{\theta}, \mathbf{K} &\sim f(y_i|\theta_{K_i}) \\ K_i|\mathbf{w} &\sim \sum_{k=1}^{\infty} w_k \delta_k \\ w_k &= \beta_k \prod_{j=1}^{k-1}(1-\beta_j) \\ \sum_{k=1}^{\infty} w_k &= 1 \\ \beta_k &\sim \text{Beta}(1, \gamma) \\ \theta_k &\sim G_0 \end{aligned} \tag{2.13}$$

2.3.5.2 Generative

We can use the predictive distribution of a DP to generate samples from a DP mixture model as follows. We first sample component parameters $\boldsymbol{\phi}_1$ from G_0 and then generate a sample y_1 from the component distribution f: $y_1 \sim f(\boldsymbol{\phi}_1)$. Subsequent values are sampled from a new component with likelihood $\frac{\gamma}{\gamma+N}$ or from existing component j with probability $\frac{n_j}{\gamma+N}$, where n_j is the number of observations obtained

from component j and $N = \sum_j n_j$. Repeating this, results in a sample $y_1, \ldots, y_N$ from mixing a $\text{DP}(\gamma, G_0)$ with component distribution f.

2.3.5.3 Inference

To fit an infinite mixture model in a Bayesian context, we need to infer the posterior distribution of the component parameters $\boldsymbol{\theta}$, the unique set of values in $\boldsymbol{\phi}$, and weights $\mathbf{w}$. The DP provides a prior distribution for $\boldsymbol{\theta}$ via G_0 and the concentration parameter γ determines the spread of the weight over the components: the smaller γ, the fewer components have non-negligible weights. Exact computation of the posterior distribution $p(\boldsymbol{\theta}, \mathbf{w}|y)$ is not feasible for more than a few observations (Neal, 2000), so MCMC algorithms have been proposed to estimate the posterior distribution (Escobar, 1994; Escobar and West, 1995; Liu, 1996; MacEachern and Müller, 1998; MacEachern et al., 1999; Neal, 2000; Green and Richardson, 2001; Fearnhead, 2004; Jain and Neal, 2004; Blei and Jordan, 2006; Walker, 2007; Papaspiliopoulos and Roberts, 2008; Papaspiliopoulos, 2008). The reason why exact computation is practically impossible is that direct simulation from the posterior distribution is difficult due to the intractability of the normalising constant which involves a summation over an infinite number of terms.

Ishwaran and Zarepour (2000) proposed a sampling approach, based on the stick-breaking algorithm, which was further developed in Ishwaran and James (2001) and Ishwaran and James (2003). The method allows inference for the latent random measure G of the DP and does not rely on being able to integrate out components of the hierarchical model analytically, thereby making it more flexible. We use this method when sampling from a DP mixture model in Chapter 4. As the stick-breaking algorithm requires the imputation of the infinite-dimensional vectors $\mathbf{w}$ and $\boldsymbol{\theta}$, and the computation of an infinite sum of random terms, $\sum_{k=1}^{\infty} w_k f(y|\theta_k)$, Ishwaran and Zarepour (2000) suggested using a C-dimensional approximation of the DP, using a truncation of the stick-breaking algorithm (Equation 2.13) by fixing C and letting $\beta_C = 1$. This implies that $w_k = 0$ for $k > C$, overcoming the issues caused by the infinite-dimensional variables $\mathbf{w}$ and $\boldsymbol{\theta}$ in the stick-breaking representation of a

DP. We can update the parameters $\mathbf{K}$, $\mathbf{w}$, $\boldsymbol{\theta}$ and γ using a blocked collapsed Gibbs sampler:

$$p(\mathbf{K}|\mathbf{w}, \boldsymbol{\theta}, \mathbf{y})$$
$$p(\mathbf{w}|\mathbf{K}, \mathbf{y})$$
$$p(\boldsymbol{\theta}|\mathbf{K}, \mathbf{y})$$
$$p(\gamma|\boldsymbol{\beta}, \mathbf{y})$$

The truncation allows samples for K_i and $\boldsymbol{\theta}$ to be generated using the distributions derived for the finite mixture model (Equations 2.6 and 2.7 on pages 68 and 68, respectively). To facilitate the update of $\boldsymbol{\theta}$, one could use a conjugate prior G_0 for $p(y|\boldsymbol{\theta})$, but this is not necessary as one could use alternative MCMC approaches.

To update $\mathbf{w}$ and γ we follow Ishwaran and Zarepour (2000):

$\mathbf{p(w|K, y)}$

Let us define m_k as the number of values that are assigned to cluster k, i.e. $m_k = \sum_{i=1}^{N} \delta_{K_i,k}$ where N is the number of observations. We can now update w using:

$$\begin{aligned} w_1 &= \beta_1 \\ w_k &= \beta_k \prod_{j=1}^{k-1}(1-\beta_j) \\ \beta_k &\sim \text{Beta}\left(1 + m_k, \gamma + \sum_{j=k+1}^{C} m_j\right) \qquad \text{for } k = 1, \ldots, C-1 \end{aligned} \tag{2.14}$$

$\mathbf{p(\gamma|\boldsymbol{\beta}, y)}$

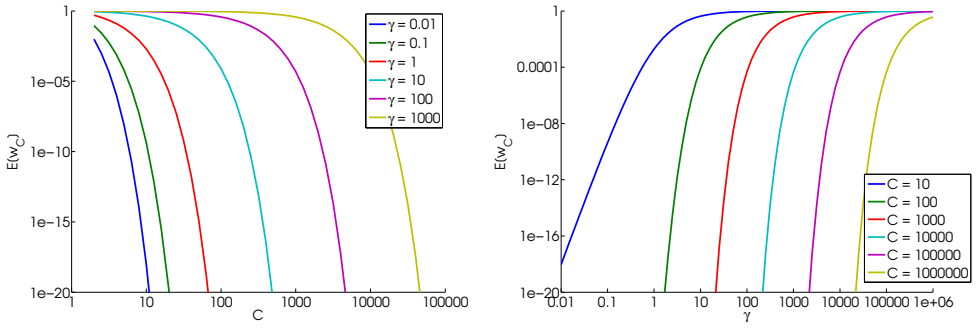
As the number of components with significant posterior probability is sensitive to γ, it has been suggested that a weakly informative prior for γ should be used. Using a conjugate prior distribution, Gamma(ν_1, ν_2) (Ishwaran and Zarepour, 2000), the posterior distribution is given by:

$$\gamma|\boldsymbol{\beta}, \mathbf{y} \sim \text{Gamma}\left(\nu_1 + C - 1, \nu_2 - \sum_{k=1}^{C} \log(1-\beta_k)\right)$$

where β_k comes from Equation 2.14.

Figure 2.6 – *Expected prior probability of cluster C, $\mathbb{E}[w_C]$, as a function of C and γ.*

(a) $\mathbb{E}[w_C]$ as a function of C **(b)** $\mathbb{E}[w_C]$ as a function of γ



Approximation Error

To assess whether the truncation level C is adequate, we can explore the properties of $\sum_{k=C}^{\infty} w_k = 1 - \sum_{k=1}^{C-1} w_k$ for the prior DP mixture model. Ishwaran and Zarepour (2000) suggest that one can test whether or not w_C is small enough by evaluating its mean and variance. As β_C is set to 1 by definition, i.e. the stick length that has not been assigned yet will be assigned to the last component C, we can calculate the expected weight of the final component C as:

$$\mathbb{E}[w_C] = \left(\frac{\gamma}{1+\gamma}\right)^{C-1} \tag{2.15}$$

The variance is given by:

$$\mathrm{Var}[w_C] = \left(\frac{\gamma}{\gamma+2}\right)^{C-1} - \left(\frac{\gamma}{\gamma+1}\right)^{2(C-1)}$$

To assess the impact of truncation, one could either compare Equations 2.11 (page 83) and 2.15, showing that there is a factor $(\gamma+1)^{-1}$ difference, or one could simply assess the expected probability of the final component C using Equation 2.15. Figure 2.6 shows how $\mathbb{E}[w_C]$ depends on the choice of C and γ. If $\mathbb{E}[w_C]$ is large, this implies that one should consider increasing C or reducing γ.

Ishwaran and James (2001) provide an alternative estimate of the approximation error, defined as the distance between the marginal distributions of a truncated

prior DP, $p_C(y)$, and a DP with an infinite number of components, $p_\infty(y)$:

$$\int |p_C(y) - p_\infty(y)| dX \approx 4N \exp\left[-\frac{C-1}{\gamma}\right]$$

where N is the number of observations.

To assess whether the truncation level C is reasonable for the posterior distribution of the DP mixture model we would like to calculate the expected tail probability. However as this is not possible our only option is to assess the tail probability post-analysis, as we do in Chapter 4. If the tail probabilities are too high, the analysis can be redone, either with smaller values of γ (unless that causes issues with the smoothness of the curve) or with a higher truncation level. This can be adjusted after some training runs to suit each application.
Gelfand and Kottas (2002) provide an alternative approximate sampling approach that limits the number of components. They make use of the posterior distribution (Equation 2.12 on page 85) and the expected weight of the final component (see Equation 2.15) to find an acceptable level of approximation error.

Label-swapping

In the stick-breaking representation, the weights assigned to clusters depend on the cluster labelling, i. For components j and k with $j < k$, $\mathbb{E}[w_j] > \mathbb{E}[w_k]$, although there is a non-zero probability that $w_j < w_k$, particularly if $|j-k|$ is small. The posterior distributions of w are multimodal which might lead to poor mixing in Gibbs sampling algorithms as the sampler has to visit all the different modes. Label-swapping moves have been introduced (Porteous et al., 2006; Papaspiliopoulos and Roberts, 2008) to improve the performance of the algorithm. Without label-swapping the Gibbs sampler for the w_i distributions is likely to remain in one of the modes. For example, for w_1 it is likely that the sampler remains in the upper tail of its distribution as the stick-breaking algorithm results in w_1 being on average higher than the other w_i. The problem is likely to arise when there are two (or more) clusters of data which are separated, e.g. n values sampled from $\mathcal{N}(\mu_1, \sigma^2)$ and n values sampled from $\mathcal{N}(\mu_2, \sigma^2)$ where $|\mu_1 - \mu_2| > 5\sigma$. The stick-breaking algorithm

is likely to assign higher weights to cluster j, associated with the data around μ_1, than cluster k, associated with the data around μ_2 if $j < k$, despite the fact that for both clusters the same number of data were observed. If there was an additional observation, $y_{2n+1} = \frac{\mu_1+\mu_2}{2}$, i.e. exactly in between the two clusters, this observation is more likely to be assigned to cluster j than to k, increasing the likelihood that we sample w_j to be larger than w_k. To ensure that the weights w_j and w_k, for clusters of equal size, are similar, the labels should be swapped regularly. Swapping labels encourages the model to move around the sample spaces for w more efficiently. Without label-swapping moves, many iterations would be needed to overcome the problem that the sampler is not efficiently exploring the whole sampling spaces for each of the w_i and high thinning factors would be necessary to reduce the correlation between samples. Three types of label-swapping steps are introduced to overcome this problem:

1. Swap two randomly chosen pairs.

2. Swap adjacent pairs in order.

3. Swap adjacent pairs in random order.

These steps require Metropolis-Hastings steps for which we will derive the algorithms next.

1. Swap two randomly chosen pairs

The first label-swap involves swapping the data assignments and component parameters θ_k of two randomly chosen labels j and l. By doing so, we keep the data assigned to the components that they were assigned to before the swap but we propose to remove the link to the weights w_k. If we want to swap the labels of two randomly chosen components, C_1 and C_2, the proposal ratio equals 1 because (i) the transition from old values to proposal values and proposal values to old values is the same (as we swap the same pair of labels) $P(C_1 = j \cap C_2 = l) = P(C_1 = l \cap C_2 = j)$ and (ii) the mechanism for choosing the pair is independent of the state of the chain.

The posterior distribution is given by:

$$p(\boldsymbol{\theta}, \mathbf{K}, \mathbf{w}|\mathbf{y}) \propto \prod_{i=1}^{N} p(y_i|\theta_{K_i}, K_i, w_{K_i}) p(\mathbf{K}|\boldsymbol{\theta}, \mathbf{w}) p(\theta_{K_i}) p(\mathbf{w})$$

where $\mathbf{K}$ is a vector of allocations K_i, the component to which data value i is allocated, $\boldsymbol{\theta}$ is the set of component parameters and $\mathbf{w}$ is the set of component weights. The probability that an observation y_i will be assigned to a component k is:

$$p(K_i = k|w_k, \theta_k) \propto \frac{w_k p(y_i|\theta_k)}{\sum_{q=1}^{C} w_q p(y_i|\theta_q)}$$

where C is the total number of components. The target ratio is given by the probability that the n_l data, $\mathbf{y}^{(l)}$, assigned to l will now be assigned to component j and the n_j data, $\mathbf{y}^{(j)}$, assigned to j will now be assigned to component l:

$$\begin{aligned} \text{Target Ratio } &= \frac{\prod_{i=1}^{n_l} \left[w_j p(y_i^{(l)}|\theta_l) \right] \prod_{i=1}^{n_j} \left[w_l p(y_i^{(j)}|\theta_j) \right]}{\prod_{i=1}^{n_j} \left[w_j p(y_i^{(j)}|\theta_j) \right] \prod_{i=1}^{n_l} \left[w_l p(y_i^{(l)}|\theta_l) \right]} \\ &= \left(\frac{w_j}{w_l} \right)^{n_l - n_j} \end{aligned}$$

where $y_i^{(c)}$ is the i^{th} data value assigned to component c.

2. Swap adjacent pairs in order

Swapping random pairs works well if the two components have similar weights, but does not work well for very unequal components (with respect to the number of data allocated and the weights). To swap labels for unequal components, an alternative label-swapping step is introduced. Like the previous type of label-swaps the proposal ratio equals 1, so we can focus on the target ratio. Swapping the labels j and $j+1$ of two neighbouring components together with the unit stick-breaking lengths β_j and β_{j+1} means that data, y_i, associated with component j with probability $w_j^{(t)} p\left(y_i|\theta_j^{(t)}\right)$ is proposed to be allocated to component $j+1$:

$$\text{Target Ratio} = \frac{\prod_{i=1}^{n_j} w_{j+1}^{*} p\left(y_i|\theta_{j+1}^{*}\right) \prod_{i=1}^{n_{j+1}} w_j^{*} p\left(y_i|\theta_j^{*}\right)}{\prod_{i=1}^{n_j} w_j^{(t)} p\left(y_i|\theta_j^{(t)}\right) \prod_{i=1}^{n_{j+1}} w_{j+1}^{(t)} p\left(y_i|\theta_{j+1}^{(t)}\right)}$$

where:

$$\begin{aligned} w_{j+1}^{(t+1)} &= (1-\beta_1)(1-\beta_2)\dots(1-\beta_{j-1})(1-\beta_{j+1})\beta_j \\ w_j^{(t)} &= (1-\beta_1)(1-\beta_2)\dots(1-\beta_{j-1})\beta_j \\ w_j^{(t+1)} &= (1-\beta_1)(1-\beta_2)\dots(1-\beta_{j-1})\beta_{j+1} \\ w_{j+1}^{(t)} &= (1-\beta_1)(1-\beta_2)\dots(1-\beta_{j-1})(1-\beta_j)\beta_{j+1} \end{aligned}$$

The factors $(1-\beta_1)\dots(1-\beta_{j-1})$ cancel out and with $\theta_j^{(t)} = \theta_{j+1}^*$ and $\theta_{j+1}^{(t)} = \theta_j^*$ the Normal density functions cancel out as well. As a result, the stick breaking lengths β_j and β_{j+1} are the only terms left in the target ratio:

$$\begin{aligned} \text{Target Ratio} &= \frac{\prod_{i=1}^{n_j} w_{j+1}^* \prod_{i=1}^{n_{j+1}} w_j^*}{\prod_{i=1}^{n_j} w_j^{(t)} \prod_{i=1}^{n_{j+1}} w_{j+1}^{(t)}} \\ &= \prod_{i=1}^{n_j} \frac{(1-\beta_{j+1})\beta_j}{\beta_j} \prod_{i=1}^{n_{j+1}} \frac{\beta_{j+1}}{(1-\beta_j)\beta_{j+1}} \\ &= \frac{(1-\beta_{j+1})^{n_j}}{(1-\beta_j)^{n_{j+1}}} \end{aligned}$$

If the swap is accepted, we calculate new weights w using the reordered β_j:

$$\begin{aligned} w_1 &= \beta_1 \\ w_k &= \beta_k \prod_{j=1}^{k-1}(1-\beta_j) \qquad k \geq 2 \end{aligned}$$

3. Swapping adjacent pairs in random order

In addition to swapping the adjacent pairs in a fixed order we also propose a randomisation step which allows us to swap adjacent pairs in random order. This is easily done by random permutation of the $C-1$ adjacent pairs. The acceptance ratio is equal to the ratio derived in the previous section:

$$\text{Acceptance Ratio} = \frac{(1-\beta_{j+1})^{n_j}}{(1-\beta_j)^{n_{j+1}}}$$

2.4 Conclusion

In this chapter we have provided an introduction to several of the mathematical concepts that will be used throughout this thesis. This included an introduction to Bayesian inference, MCMC, the Dirichlet distribution and Dirichlet Processes. The Dirichlet distribution plays an important role in the model in Chapter 3 whilst Dirichlet Process mixture models are used in Chapter 4.

Chapter 3

Multivariate modelling of pesticide residues

3.1 Introduction

The use of pesticides to protect crops from pests and diseases may result in pesticide residues on agricultural produce. Farmers may treat a crop with multiple pesticides for various reasons including managing various types of pests, using up old stock as part of a tank mix, creating a commercial product using two generic, cheaper products or reducing the risk of resistance by using a range of pesticides which have different modes of action. When multiple pesticides are applied to a crop, either at the same time in a tank mix or at different growth stages, residue levels of multiple pesticides may occur on individual food items. Therefore the variation in these residue levels should be modelled using multivariate techniques to account for any correlations in residue levels. These techniques can then be used in a cumulative risk assessment to assess whether dietary exposure from eating products that are treated with multiple pesticides are below the level of concern.

This chapter introduces two novel approaches to model pesticide log-residues in composite samples which are able to combine information on pesticide usage with data on residue levels from monitoring programmes. They make use of the GB Pes-

ticide Usage Survey (PUS) data to inform the models on the proportion of composite samples that have been treated with pesticides. One of the approaches also offers a solution to model both presence/absence of multiple pesticide residues and the correlation between log-residue amounts when multiple pesticides have been applied to a single crop.

In this chapter, we will begin with a discussion of the available data (Section 3.2) and currently proposed approaches for modelling co-occurrence of pesticides (Section 3.3). In Section 3.4 we will illustrate that it is important to develop models that account for correlations between pesticide log-residue levels. Section 3.5 introduces the new approaches which we will validate using synthetic data sets and compare with current approaches in Section 3.6. Finally, the new approaches will be demonstrated in a case study (Section 3.7).

3.2 Data

3.2.1 Pesticide Usage Survey Data

In Great Britain, Pesticide Usage Survey (PUS) data are collected for a number of purposes including informing the pesticide approval process, assessing the economic and/or environmental implications of introducing new active substances and informing the targeting of monitoring programmes for residues in food and the environment (Fera, 2011). For produce grown in GB, these data can be used to identify patterns in absence/presence of pesticides on different raw agricultural commodities for use in dietary risk assessment if we assume that the proportion of fields with a given treatment equals the probability that a composite monitoring sample has received a certain treatment. When using the PUS data we need to account for the fact that the survey only samples a proportion of total British production, often for broad classes of crops. As a result, it is possible that other combinations of pesticides were used but not sampled, in which case not all possibilities are represented in the survey. We also need to account for the fact that pesticide application may result in a higher yield. As a consequence, it may be the case that the proportion of fields treated with a pesticide is not equal to the proportion of composite samples with

that treatment.

To overcome these issues when using PUS data in dietary risk assessment, the PUS data should be treated as an uncertain estimate of the proportion of composite samples having received a certain treatment. The advantage of including PUS data in a dietary risk assessment is that it provides more information about possible crop treatment histories which is important when we have composite samples with high levels of censoring.

In this chapter we present a case study for British carrots and therefore use GB PUS data. However for produce not grown in GB, other pesticide usage information would need to be identified and used together with residue level information from the country of origin.

3.2.2 Monitoring Data

As described in Section 1.3.1.1, samples of raw agricultural commodities (RACs) are routinely collected and residue levels are measured in composite samples which are derived from multiple units of the commodity. However, little is known about the origin of the units in a composite sample as products that are collected may come from various sources (e.g. different fields with different treatments). As discussed in Chapter 1, the fact that monitoring programmes are primarily aimed at assessing compliance with MRLs introduces various problems when using them in dietary risk assessments. Firstly, monitoring data are a mixture of samples, obtained using some random sampling and an unknown degree of targeted sampling based on e.g. the violation rate in previous years. However, in the absence of other data, we follow existing dietary models by treating these data as though they were a random sample. Secondly, residue levels on composite samples obtained as part of monitoring programmes are often left-censored (see Section 1.4.1.1). Therefore methods describing the variation in residue levels need to model censored data appropriately. Bayesian methods can be used but high levels of censoring increase the influence of the prior distribution on the posterior distribution. Therefore, it is important to obtain prior distributions that are supported by independent data, e.g. PUS data, or expert knowledge.

3.3 Current Approaches for Cumulative Risk Assessment

When modelling residue levels from multiple pesticides two questions need to be answered: (1) what is the likelihood that combinations of pesticides occur (i.e. presence) and (2) given that pesticides co-occur, how can we model the dependency in residue levels (i.e. amounts). Several approaches have recently been proposed for cumulative dietary risk assessments when multiple pesticides need to be considered (EFSA, 2009; Van Klaveren et al., 2009; EFSA, 2012).

One approach (pairwise empirical sampling) resamples observed residue level vectors from a number of composite samples. For each of n composite samples, residue levels are measured for m pesticides and reported in a $n \times m$ matrix. In this approach, residue levels will be obtained by sampling rows from this matrix to account for dependencies between pesticide residues. Pesticide residue data sets may come from multiple sources and samples may be tested for different pesticides. Therefore when combining different data sets, there will be missing values for those pesticides that were not measured in a particular data set. In this approach only observed values can be resampled and non-detects, i.e. values below the limit of determination (LOD), and missing values are set to zero. An implication of the first feature is that residue levels other than the measured ones cannot occur and thus residue concentrations cannot be higher than the highest value observed in the data. EFSA (2010a) reported that in 2008, in 29 countries approximately 53 carrots were sampled per country on average. Given this small sample size, it is unlikely that the observed concentrations provide a representative sample of the whole spectrum of residue levels on carrots. The second feature of setting <LOD and missing values to zero may underestimate the true residue levels. To account for the uncertainty in residue levels bootstrap approaches (Section 1.4.2) have been suggested.

Another approach for cumulative risk assessment ignores dependencies in residue concentrations by modelling the residue levels for each pesticide separately. We will explore two implementations of this approach, the first is based on sampling the

data for each chemical independently assuming an empirical distribution and setting <LOD values to zero. To account for uncertainty a bootstrap approach was used. A more sophisticated implementation made use of a Bayesian mixture model (Paulo et al., 2005) that accounts for the fact that a censored observation may be either a positive, undetectable residue level or the result of untreated food items. Both implementations only use data from those laboratories that have measured the pesticide of interest, thus missing data are no longer part of the model. If pesticide residue levels are not correlated, this approach may be appropriate.

EFSA (2012) suggested two approaches for cumulative dietary risk assessment depending on the exposure scenario and whether the available data included missing values. The first scenario, termed the optimistic approach, assumes that all <LOD and missing values are zeros and uses pairwise bootstrap approaches to account for uncertainty in residue levels. The second scenario, termed the pessimistic approach, assumes that all <LOD are equal to the LOD and then fits a Lognormal distribution to the positive residue data. Missing values are dealt with by imputation from the fitted Lognormal distribution. As imputing missing values independently for each chemical may affect the dependencies in residue levels, EFSA (2012) proposed an approach in which missing data will be dealt with in such a way that a conservative estimate of residue levels is obtained.

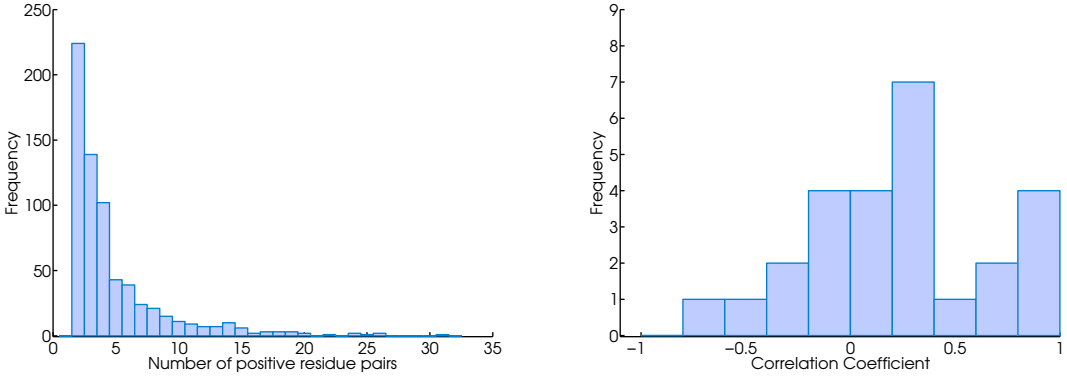
3.4 Correlations in log-residue levels

To assess the correlations in log-residue levels of different pesticides in composite samples, we analysed the 2010 UK monitoring data. Composite samples for 20 crops, which had at least 30 samples with detectable residue levels, were selected from the 2010 surveys (PRC, 2010; PRC, 2011a; PRC, 2011b; PRC, 2011c) and Pearson correlation coefficients were calculated for all pesticide combinations for each crop. The calculated correlation coefficients were only based on residue levels above the limit of determination (LOD). As composite monitoring data tend to have high proportions of data below the LOD, very few composite samples are generally

Figure 3.1 *– Analysis of 2010 UK monitoring data.*

(a) Frequency of the number of composite samples available to calculate the correlation coefficients (i.e. samples containing detectable residues for any pair of possible pesticides) for 20 crops from the 2010 UK monitoring data.

(b) Correlations between log pesticide residue levels in composite samples from 7 crops in 2010 UK monitoring data. Only correlation coefficients based on more than 15 data values are included.



available to determine the correlations in residue levels. This is shown in Figure 3.1a where for each of the 20 crops we count how many of the available composite samples had detectable residues for each possible pair of pesticides. For example, for apples there are 143 composite samples available in which 36 chemicals have been measured. We count the number of cases where only n composite samples of apples contain detectable residues of any pair of pesticides, where n is any integer between 2 and 143. We find that there are 41 cases where only 2 composite samples contained detectable residues of any pair of pesticides, 14 cases where only 3 of the composite samples contained detectable residues of any pair of pesticides, etc. Repeating this process for the other 19 crops leads to the frequencies shown in Figure 3.1a. As we can see, there are more likely to only be 3 composite samples available with detected pesticide residues (frequency of 139) to calculate the pairwise correlation coefficients than 20 composite samples (frequency of 2).

Figure 3.1b shows the distribution of correlation coefficients for those cases where at least 15 residue levels were detected for both pesticides (26 correlation coefficients from 7 crops). This number was a pragmatic choice to ensure there were enough cor-

relation coefficients to observe the variation in correlations. However, we are aware that small sample sizes will have an effect on the estimation of the correlation coefficient. Figure 3.1b indicates that non-zero correlations do occur and that therefore correlations in log-residue levels should be modelled when conducting cumulative dietary risk assessments.

Given that the monitoring samples may contain units from different fields from one or more countries and the pesticides which are authorised may vary between countries, the treatment that food products will have received is likely to vary. Therefore, we need a model which can describe the variation in residue levels and deal with the fact that residue levels that are reported to be below the LOD can either be zero (i.e. product was not treated with a particular pesticide) or somewhere between zero and the LOD (i.e. product was treated but levels are too low to quantify). The model will also need to account for any correlations suggested by available evidence e.g. from monitoring or PUS data. The final requirement for a residue model is that it should be able to account for the uncertainty about the model parameters caused by the small number of observations available to estimate them.

3.5 Model Descriptions

In this section we discuss two approaches: the independent mixture model which can be used when we are interested in either a single pesticide or multiple pesticides where log-residue levels are thought to be independent; and a bivariate mixture model that aims to model correlations in log-residue data.

3.5.1 Independent Mixture Model

Paulo et al. (2005) introduced a univariate mixture model to model the variation in residue levels for a single pesticide. Their model assigned log-residue levels above the limit of determination to a Normal distribution. Data below the limit of determination could either be a zero (i.e. not treated with pesticide) or a value between zero and the limit of determination. In their model, residue data, y, is therefore

described using the following mixture distribution:

$$f(y) = p_0\delta_{y,0} + (1 - p_0)\mathcal{LN}(y; \mu, \sigma^2)$$

where p_0 is the probability of a residue level being zero, $\delta_{y,0}$ is the Kronecker delta function and $\mathcal{LN}(\mu, \sigma^2)$ indicates a Lognormal density with parameters μ and σ^2. From this, it is clear that the probability that 'a datum is less than the LOD and greater than zero' is:

$$(1 - p_0)\Phi\left(\log(\text{LOD}); \mu, \sigma^2\right)$$

This follows from the fact that the Normal cumulative distribution function, $\Phi\left(\log(\text{LOD}); \mu, \sigma^2\right)$, provides the probability of observing a value less than log(LOD). To infer the parameters of this model, Paulo et al. (2005) used a MCMC algorithm:

1. Sample the number of samples below the LOD with zero residue level, n_0.
2. Sample latent residue values for the number of data that are between zero and the LOD: $n_{<\text{LOD}} - n_0$.
3. Sample p_0, σ and μ.
4. Go to step 1.

The conditional probability that the residue level for a single sample <LOD is zero is:

$$p_z = \frac{p_0}{p_0 + (1 - p_0)\Phi(\log(\text{LOD}); \mu, \sigma^2)}$$

If we know p_z, we can sample how many of our observations below the limit of determination are zero using:

$$n_0 \sim \text{Binomial}(n_{<\text{LOD}}, p_z)$$

where $n_{<\text{LOD}}$ is the number of values less than the LOD. Since we do not know the model parameters p_0, μ and σ, we have to learn these from the data by sampling latent values for the censored data to update p_0, μ and σ. Paulo et al. (2005) use a non-informative prior, $\pi(\mu, \sigma) \propto \frac{1}{\sigma}$, for the Normal distribution parameters and

$\pi(p_0) = \text{Beta}(p_0; 1, 1)$ as a prior for p_0. This allows for a simple Bayesian update for p_0, μ and σ:

$$\begin{aligned} p_0|\mathbf{y} &\sim \text{Beta}(1 + n_0, 1 + n - n_0) \\ \frac{(n_+ - 1)s^2}{\sigma^2}|\mathbf{y} &\sim \chi^2_{n-1} \\ \mu|\sigma, \mathbf{y} &\sim \mathcal{N}\left(m, \frac{\sigma^2}{n_+}\right) \end{aligned}$$

where n is the total number of log-residue observations, $\mathbf{y}$, $n_+ = n - n_0$ and m and s are the mean and standard deviation, respectively, of the n_+ non-zero residues including the latent ones.

We propose a generalisation of the Paulo et al. (2005) model by using information on the proportion of untreated field area from the GB PUS data. Although one could argue that we can replace p_0 by this fixed number, we propose the following model:

$$p_0 \sim \text{Beta}\left(w \times \text{PUS}_0, w \times (1 - \text{PUS}_0)\right)$$

where PUS_0 is the proportion of untreated field area and w is the prior sample size, a factor that can be used to express our belief that the PUS data is relevant for the composite monitoring data. Note that $\lim_{w\to\infty} p_0 = \text{PUS}_0$, which means that for large w we believe that we know the proportion of untreated field area and therefore the proportion of untreated samples. For smaller values of w, we assert that p_0 will be more uncertain.

This model can be used in cumulative risk assessment if we assume that two or more pesticide residue distributions are independent.

3.5.2 Bivariate Mixture Model

We now describe a bivariate mixture model for cumulative risk assessments which accounts for the correlations between log-residue concentrations.

Model Specification

Let us assume that two chemicals, X and Y, were measured in a composite sample. The results of the analysis will fit into one of four categories: $\{x, y\}$, $\{x, < \text{LOD}_Y\}$, $\{< \text{LOD}_X, Y\}$ and $\{< \text{LOD}_X, < \text{LOD}_Y\}$, where x and y indicate a measured residue level above the limits of determination (LOD_X and LOD_Y, respectively), $< \text{LOD}_X$ indicates that the composite sample either did not contain residue levels of X (i.e. $x = 0$) or that the levels were too low to be determined ($0 < x < \text{LOD}_X$) and $< \text{LOD}_Y$ similarly indicates that $y = 0$ or the residue levels of Y were too low to be determined. Let us now define the observable indicator functions M_X and M_Y to distinguish the cases where X and Y are above the LOD from the cases where they are below the LOD.

$$M_X = \begin{cases} 1 & \text{if } X \geq \text{LOD}_X \\ 0 & \text{if } X < \text{LOD}_X \end{cases} \qquad M_Y = \begin{cases} 1 & \text{if } Y \geq \text{LOD}_Y \\ 0 & \text{if } Y < \text{LOD}_Y \end{cases}$$

As we do not know whether <LOD values are true zeros or positive residues which are <LOD, we also need to define the latent indicator functions Z_X and Z_Y:

$$Z_X = \begin{cases} 1 & \text{if } X = 0 \\ 0 & \text{if } X > 0 \end{cases} \qquad Z_Y = \begin{cases} 1 & \text{if } Y = 0 \\ 0 & \text{if } Y > 0 \end{cases}$$

The probability that a residue sample comes from each of the four mixture components mentioned above is:

$$\alpha_0 = p(Z_X = 1, Z_Y = 1)$$
$$\alpha_X = p(Z_X = 0, Z_Y = 1)$$
$$\alpha_Y = p(Z_X = 1, Z_Y = 0)$$
$$\alpha_{XY} = p(Z_X = 0, Z_Y = 0)$$

Now let us define a mixture density, f, that can be used to describe the observed composite samples:

$$f(x, y) = \alpha_0 \delta_{x,0} \delta_{y,0} + \alpha_X \delta_{y,0} f_X(x) + \alpha_Y \delta_{x,0} f_Y(y) + \alpha_{XY} f_{XY}(x, y)$$

where f_X is the probability density function (pdf) of X given $Z_X = 0$ and $Z_Y = 1$, f_Y is the pdf of Y given $Z_X = 1$ and $Z_Y = 0$, f_{XY} is the joint pdf of (X, Y) given

$Z_X = 0$ and $Z_Y = 0$ and $\delta_{k,0}$ is the Kronecker delta function. For the remainder of this chapter, we will assume that residue levels can be described with Lognormal distributions (f_X and f_Y) and a bivariate Lognormal distribution (f_{XY}) as assuming Lognormal distributions is common practice in current dietary risk assessments. However, the approach presented here can be applied to any other distribution shape as well. We now have:

$$\begin{aligned} f_X(x) &= \mathcal{LN}(x; \mu_X, \sigma_X^2) \\ f_Y(y) &= \mathcal{LN}(y; \mu_Y, \sigma_Y^2) \\ f_{XY}(\{x, y\}) &= \mathcal{LN}_2(\{x, y\}; \mu^{XY}, \Sigma^{XY}) \end{aligned}$$

where μ_X and μ_Y are the mean log-residue levels for X and Y respectively, σ_X and σ_Y are the standard deviations of the log-residue levels and the mean and covariance matrix of the bivariate Normal distribution are:

$$\mu^{XY} = \begin{bmatrix} \mu_X^{XY} \\ \mu_Y^{XY} \end{bmatrix}$$

$$\Sigma^{XY} = \begin{bmatrix} (\sigma_X^{XY})^2 & \sigma_X^{XY}\sigma_Y^{XY}\rho_{XY} \\ \sigma_X^{XY}\sigma_Y^{XY}\rho_{XY} & (\sigma_Y^{XY})^2 \end{bmatrix}$$

Inference

To infer the parameters of the model we use an MCMC algorithm. The residue samples need to be assigned to each distribution to infer the probabilities α_0, α_X, α_Y and α_{XY} and the distribution parameters of f_X, f_Y and f_{XY}. For this purpose we define K to be a latent variable which indicates which distribution an observation i is assigned to, where $i = 1, \ldots, n$ and n is the sample size:

$$K_i = \begin{cases} 1 & \text{if } Z_X = 1 \wedge Z_Y = 1 \\ 2 & \text{if } Z_X = 0 \wedge Z_Y = 1 \\ 3 & \text{if } Z_X = 1 \wedge Z_Y = 0 \\ 4 & \text{if } Z_X = 0 \wedge Z_Y = 0 \end{cases}$$

For each composite sample we now have the following conditional allocation probabilities given the four possible combinations of the indicator functions, M_X and M_Y,

$\{0,0\}$, $\{1,0\}$, $\{0,1\}$ and $\{1,1\}$:

$$p(K=j|M_X,M_Y)=\frac{p(M_X,M_Y|K=j)p(K=j)}{p(M_X,M_Y)}$$

where $p(M_X,M_Y)=\sum_j p(M_X,M_Y|K=j)p(K=j)$ for $j=1,\ldots,4$ where applicable (e.g. $j\in\{1,2,3,4\}$ for $\{0,0\}$ and for $\{1,1\}$ we know $j=4$).

For the $\{0,0\}$ case, we can calculate the probabilities as follows:

$$\begin{aligned}
p(K=1|M_X=0,M_Y=0)&\propto\alpha_0\\
p(K=2|M_X=0,M_Y=0)&\propto\alpha_X\Phi\left(\frac{\log(\mathrm{LOD}_X)-\mu_X}{\sigma_X}\right)\\
p(K=3|M_X=0,M_Y=0)&\propto\alpha_Y\Phi\left(\frac{\log(\mathrm{LOD}_Y)-\mu_Y}{\sigma_Y}\right)\\
p(K=4|M_X=0,M_Y=0)&\propto\alpha_{XY}\Phi_2\left(\{\log(\mathrm{LOD}_X),\log(\mathrm{LOD}_Y)\};\mu^{XY},\Sigma^{XY}\right)
\end{aligned}$$

where Φ is the standard Normal cumulative distribution function and Φ_2 is the bivariate Normal cumulative distribution function.

For $\{1,0\}$, we can calculate the probabilities as follows:

$$\begin{aligned}
p(K=2|M_X=1,M_Y=0)&\propto\alpha_X\phi\left(\frac{\log(x)-\mu_X}{\sigma_X}\right)/\sigma_X\\
p(K=4|M_X=1,M_Y=0)&\propto\alpha_{XY}p_X(x)P_Y(Y<\mathrm{LOD}_Y|X=x)
\end{aligned}$$

where ϕ is the standard Normal density function, $p_X(x)$ is the marginal distribution of X obtained from f_{XY} and $P_Y(Y<\mathrm{LOD}_Y|X=x)$ is the conditional probability $Y|X$:

$$p_X(x)=\phi\left(\frac{\log(x)-\mu_X^{XY}}{\sigma_X^{XY}}\right)/\sigma_X^{XY}$$

$$P_Y[Y<\mathrm{LOD}_Y|X=x]=\Phi\left(\log(\mathrm{LOD}_Y)|\mu_Y^{XY}+\frac{\sigma_Y^{XY}}{\sigma_X^{XY}}\rho\{\log(x)-\mu_X^{XY}\},(1-\rho^2)\sigma_Y^{2,XY}\right)$$

The other probabilities can be calculated in a similar way.

We will assume the following prior distribution, based on the GB PUS data, for the weights $\boldsymbol{\alpha}$:

$$\boldsymbol{\alpha}\sim\mathrm{Dirichlet}(\mathrm{PUS}_0\times w,\mathrm{PUS}_X\times w,\mathrm{PUS}_Y\times w,\mathrm{PUS}_{XY}\times w)$$

where $\text{PUS}_{L(j)}$ are the proportions of field area in each treatment combination, where $L \in \{0, X, Y, XY\}$ and $j \in \{1, 2, 3, 4\}$. As in the independent mixture model, we multiply these proportions by a weight w to indicate how certain we are that the PUS data is representative of the probability that a composite monitoring sample is treated with one of these combinations. If w goes to infinity, then the posterior proportions will match the prior proportions (i.e. the PUS proportions) and if w goes to zero the posterior proportions are determined by the monitoring data.
The Gibbs sampler MCMC algorithm can be summarised by the following steps:

1. Sample the latent allocation variable K_i for each composite monitoring sample $\{x, y\}$ using the probabilities above.

2. Sample weights $\alpha_{L(j)}$ given allocations: $\alpha_{L(j)} \sim \text{Dirichlet}(\text{PUS}_0 \times w + n_0, \text{PUS}_X \times w + n_X, \text{PUS}_Y \times w + n_Y, \text{PUS}_{XY} \times w + n_{XY})$, where $n_{L(j)}$ is the number of data assigned to distribution $f_{L(j)}$: $n_{L(j)} = \sum_{i=1}^{n} \delta_{K_i,j}$.

3. Sample residue values for <LOD data allocated to $Z_X = 0$ or $Z_Y = 0$ from distributions to which they were assigned.

4. Sample distribution parameters given allocations and positive residues.

5. Store distribution parameters and weights and go to Step 1.

Step 4 is a standard Bayesian parameter update based on conjugate distributions (a Normal-Gamma distribution for the univariate distributions and a Normal-Wishart distribution for the bivariate Normal distribution, see Section 2.2.6 for more details).

3.5.3 Extending to higher dimensions

The model can in theory be extended to more dimensions but the number of distribution parameters that will have to be estimated increases considerably. For n pesticides, the number of parameters P is given by:

$$P = 2^n - 1 + \sum_{k=1}^{n} \binom{n}{k} \times \left(\frac{k(k+3)}{2} \right)$$

For one chemical, we would have three parameters (mean, standard deviation and weight), for a mixture of two chemicals, we would have 12 parameters (two means

and two standard deviations for the univariate distributions, two means for the bivariate distribution, two standard deviations for the bivariate distribution, the correlation coefficient and three weights), for three chemicals we would have 37 parameters and for four chemicals we would have 103 parameters.

Given that composite monitoring data sets consist of very few samples with residue levels above the LOD (see Figure 3.1a), it is unrealistic to expect that a multi-dimensional model can be fitted adequately unless prior information is available on all of the model parameters. However, one option would be to reduce the number of parameters by assuming that the location and scale parameters of the bivariate distribution are equal to the parameters of the univariate distributions. This would reduce the number of parameters to $2^n - 1 + \frac{n(n+3)}{2}$ which in the two chemical case equals eight (μ_X, σ_X, μ_Y, σ_Y, correlation coefficient ρ and three weights). The PUS data for carrots, used in the case study in Section 3.7, might support this approach as the median application rate for treatment with Difenoconazole only is the same as the median application rate for Difenoconazole if both Difenoconazole and Tebuconazole were applied (0.125 kg/ha for both). Analogously the median application rate for Tebuconazole when only Tebuconazole was applied (0.18 kg/ha) was similar to the median application rate for Tebuconazole when both Difenoconazole and Tebuconazole were applied (0.17 kg/ha). Even though it is unknown whether these results can be extrapolated to other crops and pesticides, one could assess from application rates provided in the PUS data whether it is reasonable to use a simpler model.

3.6 Validation Studies

To assess the performance of the models, we use three validation data sets to test whether the models are able to determine the true, underlying distribution from which the log-residue data set was sampled. Since monitoring data sets typically consist of between 50 and 150 composite samples, we present three validation studies based on a sample size of 100. For each validation data set, we run the independent

mixture model and the bivariate mixture model and compare the results with the target distribution.

3.6.1 Design of Validation Studies

Prior distributions used in validation studies

For the independent mixture model we use the same non-informative prior distribution for μ and σ that was used in Paulo et al. (2005). For the bivariate mixture model the parameters of the prior distributions are given in Table 3.1. For all validation data sets we run the bivariate mixture model with the non-informative prior distributions. However, for validation data set C, we also show an example where we use weakly informative prior distributions. These prior distributions were based on simulated composite samples derived from unit market survey data (see Appendix A). The parameters κ, ν_{XY} and κ_{XY} were all set to 10 to add more weight to the prior distributions than in the non-informative case. For all validation studies, we used $w = 100$, indicating that the PUS data provide as much information about the true treatment proportions as the 100 log-residue data values.

Table 3.1 – *Prior distribution parameters for univariate Normal and bivariate Normal distributions of log-residue data in the bivariate mixture model. The parameters α, β, κ and μ_0 are the parameters of the Normal-Gamma prior distribution used for the univariate Normal distributions f_X and f_Y and the parameters λ_{XY}, ν_{XY}, κ_{XY} and μ_{XY} are the parameters of the Normal-Wishart prior distribution used for the bivariate Normal distribution f_{XY}.*

Parameter	Non-informative Value	Weakly informative Value
α	1	0.77
β	0.05	0.04
κ	1	10
μ_0	95^{th} percentile of data	-2
λ_{XY}	$\begin{bmatrix} 10^{-3} & 0 \\ 0 & 10^{-3} \end{bmatrix}$	$\begin{bmatrix} 4.95 & 3.72 \\ 3.72 & 4.95 \end{bmatrix}$
ν_{XY}	2	10
κ_{XY}	10^{-3}	10
μ_{XY}	$\begin{bmatrix} -2 \\ -2 \end{bmatrix}$	$\begin{bmatrix} -2 \\ -2 \end{bmatrix}$

Validation Data Set A

Validation data set A was generated by sampling 25 values from each of the following four distributions with relative weights $\boldsymbol{\alpha} = \{0.25, 0.25, 0.25, 0.25\}$: not treated, $\log(X) \sim \mathcal{N}(-1, 0.25^2)$, $\log(Y) \sim \mathcal{N}(-1, 0.25^2)$ and $\{\log(X), \log(Y)\} \sim \mathcal{N}_2\left(\begin{pmatrix}-3\\-3\end{pmatrix}, \begin{pmatrix}0.25^2 & 0.0619\\0.0619 & 0.25^2\end{pmatrix}\right)$, i.e. correlation coefficient $\rho = 0.99$. We assume that $\text{LOD}_X = \text{LOD}_Y = 0$, leading to a data set where 50% of the values had reported residue levels.

Validation Data Set B

In validation data set A, the marginal distributions for $\log(X)$ and $\log(Y)$ are both bimodal due to the choice of the distributions for $\log(X)$, $\log(Y)$ and $\{\log(X), \log(Y)\}$. To include more overlap between the univariate and bivariate distributions, validation data set B of size 100 was generated from the following four distributions with relative weights $\boldsymbol{\alpha} = \{0.4, 0.3, 0.2, 0.1\}$: not treated, $\log(X) \sim \mathcal{N}(-2, 0.25^2)$, $\log(Y) \sim \mathcal{N}(-2, 0.35^2)$, $\{\log(X), \log(Y)\} \sim \mathcal{N}_2\left(\begin{pmatrix}-1.9\\-2.1\end{pmatrix}, \begin{pmatrix}0.25^2 & 0.0469\\0.0469 & 0.25^2\end{pmatrix}\right)$, i.e. with correlation coefficient $\rho = 0.75$. We assume that $\text{LOD}_X = \text{LOD}_Y = 0$, leading to a data set where 35% of the values had reported residue levels.

Validation Data Set C

In reality, many of the monitoring data will have residue levels below or near the limit of determination with unknown proportions of true zeros and censored data. Therefore for our final validation study C we generate samples from the same distributions as in validation study B but we assume a more realistic level of censoring. To do this we calculate the 75^{th} percentile of the observed data and set this to be the limit of determination for X and Y. As a result, 9% of the data had detected residue levels. As the models may struggle when very few data are available to estimate the parameters, we investigate whether using weakly informative prior distributions improves the model performance for the bivariate mixture model.

3.6.1.1 Results of Validation Studies

We ran both the independent mixture model and the bivariate mixture model on all three validation data sets (A, B and C). For validation data set C, we ran the bivariate mixture model twice, once with non-informative prior distributions and once with weakly informative prior distributions. Each model was run in Matlab 2012a on a computer with an Intel i7-860 2.80 Ghz processor and 8GB RAM. Model runs with 1,000,000 iterations took approximately 10 minutes to complete for the independent mixture model and 100 minutes for the bivariate mixture model. Tables 3.2, 3.3, 3.4 and 3.5 provide an overview of the estimated model parameters together with the true parameters for the validation studies described above. The resulting median and 95% credible intervals of the marginal posterior distributions are shown in Figure 3.2.

Table 3.2 – *Comparison of true values and model estimates for validation data set A.*

	True	Independent Mixture Model			Bivariate Mixture Model		
Variable	Value	Mean	Median	95% CI	Mean	Median	95% CI
α_0	0.25	0.25	0.25	(0.20, 0.30)	0.25	0.24	(0.19, 0.31)
α_X	0.25	0.25	0.25	(0.20, 0.30)	0.25	0.25	(0.19, 0.31)
α_Y	0.25	0.25	0.25	(0.20, 0.30)	0.25	0.25	(0.20, 0.32)
α_{XY}	0.25	0.25	0.25	(0.20, 0.30)	0.25	0.25	(0.19, 0.31)
μ_X	-1	-1.98	-1.98	(-2.29, -1.68)	-0.92	-0.92	(-1.01, -0.84)
σ_X	0.25	1.10	1.09	(0.90, 1.36)	0.23	0.22	(0.17, 0.30)
μ_Y	-1	-2.01	-2.01	(-2.30, -1.70)	-0.98	-0.98	(-1.09, -0.87)
σ_Y	0.25	1.08	1.07	(0.88, 1.32)	0.28	0.27	(0.21, 0.37)
$\mu^X_{\mathcal{N}_2}$	-3				-3.03	-3.03	(-3.14, -2.92)
$\mu^Y_{\mathcal{N}_2}$	-3				-3.02	-3.02	(-3.13, -2.92)
$\sigma^X_{\mathcal{N}_2}$	0.25				0.27	0.26	(0.20, 0.36)
$\sigma^Y_{\mathcal{N}_2}$	0.25				0.27	0.26	(0.20, 0.36)
$\rho_{\mathcal{N}_2}$	0.99				0.99	0.99	(0.98, 1.00)

Table 3.3 *– Comparison of true values and model estimates for validation data set B.*

	True	Independent Mixture Model			Bivariate Mixture Model		
Variable	Value	Mean	Median	95% CI	Mean	Median	95% CI
α_0	0.4	0.43	0.43	(0.37, 0.49)	0.40	0.40	(0.33, 0.47)
α_X	0.3	0.27	0.27	(0.22, 0.33)	0.30	0.30	(0.23, 0.36)
α_Y	0.2	0.18	0.18	(0.14, 0.23)	0.21	0.21	(0.15, 0.27)
α_{XY}	0.1	0.12	0.12	(0.09, 0.15)	0.09	0.09	(0.06, 0.14)
μ_X	-2	-1.97	-1.98	(-2.06, -1.89)	-1.98	-1.98	(-2.08, -1.89)
σ_X	0.25	0.27	0.27	(0.22, 0.34)	0.27	0.26	(0.21, 0.34)
μ_Y	-2	-2.04	-2.04	(-2.16, -1.92)	-1.97	-1.97	(-2.13, -1.81)
σ_Y	0.35	0.32	0.31	(0.25, 0.42)	0.36	0.35	(0.27, 0.49)
$\mu^X_{\mathcal{N}_2}$	-1.9				-1.90	-1.90	(-2.10, -1.70)
$\mu^Y_{\mathcal{N}_2}$	-2.1				-2.13	-2.13	(-2.26, -2.00)
$\sigma^X_{\mathcal{N}_2}$	0.25				0.31	0.30	(0.20, 0.48)
$\sigma^Y_{\mathcal{N}_2}$	0.25				0.19	0.18	(0.13, 0.30)
$\rho_{\mathcal{N}_2}$	0.75				0.77	0.80	(0.46, 0.94)

Table 3.4 *– Comparison of true values and model estimates for validation data set C using non-informative prior distributions.*

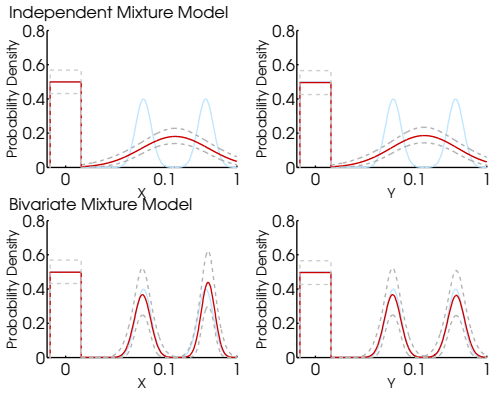
	True	Independent Mixture Model			Bivariate Mixture Model		
Variable	Value	Mean	Median	95% CI	Mean	Median	95% CI
α_0	0.4	0.41	0.41	(0.33, 0.50)	0.42	0.42	(0.32, 0.52)
α_X	0.3	0.28	0.27	(0.20, 0.35)	0.30	0.30	(0.21, 0.40)
α_Y	0.2	0.19	0.19	(0.13, 0.25)	0.20	0.20	(0.13, 0.29)
α_{XY}	0.1	0.12	0.12	(0.08, 0.18)	0.08	0.07	(0.04, 0.12)
μ_X	-2	-2.10	-2.07	(-2.47, -1.88)	-2.12	-2.09	(-2.55, -1.89)
σ_X	0.25	0.36	0.34	(0.21, 0.66)	0.34	0.31	(0.19, 0.64)
μ_Y	-2	-2.35	-2.31	(-2.98, -1.98)	-2.47	-2.40	(-3.46, -1.90)
σ_Y	0.35	0.53	0.48	(0.27, 1.05)	0.65	0.58	(0.31, 1.39)
$\mu^X_{\mathcal{N}_2}$	-1.9				-1.59	-1.56	(-1.92, -1.43)
$\mu^Y_{\mathcal{N}_2}$	-2.1				-1.75	-1.73	(-2.03, -1.66)
$\sigma^X_{\mathcal{N}_2}$	0.25				0.14	0.12	(0.06, 0.37)
$\sigma^Y_{\mathcal{N}_2}$	0.25				0.09	0.07	(0.04, 0.28)
$\rho_{\mathcal{N}_2}$	0.75				0.76	0.82	(0.14, 0.99)

Table 3.5 *– Comparison of true values and model estimates for validation data set C for the bivariate mixture model using weakly informative prior distributions.*

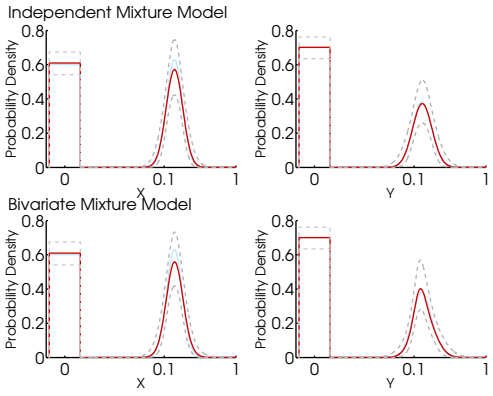
	True	Bivariate Mixture Model		
Variable	Value	Mean	Median	95% CI
α_0	0.4	0.44	0.44	(0.34, 0.53)
α_X	0.3	0.29	0.29	(0.21, 0.38)
α_Y	0.2	0.16	0.16	(0.10, 0.24)
α_{XY}	0.1	0.10	0.10	(0.06, 0.16)
μ_X	-2	-2.03	-2.02	(-2.19, -1.93)
σ_X	0.25	0.23	0.21	(0.14, 0.39)
μ_Y	-2	-2.16	-2.12	(-2.64, -1.89)
σ_Y	0.35	0.50	0.43	(0.23, 1.19)
$\mu^X_{\mathcal{N}_2}$	-1.9	-1.93	-1.93	(-2.10, -1.78)
$\mu^Y_{\mathcal{N}_2}$	-2.1	-1.99	-1.99	(-2.14, -1.84)
$\sigma^X_{\mathcal{N}_2}$	0.25	0.30	0.29	(0.21, 0.45)
$\sigma^Y_{\mathcal{N}_2}$	0.25	0.27	0.26	(0.18, 0.40)
$\rho_{\mathcal{N}_2}$	0.75	0.85	0.86	(0.64, 0.95)

Figure 3.2 – *Median (red line) and 95% credible intervals (grey dashed lines) of the marginal posterior distributions inferred using the independent mixture model and the bivariate mixture model with the target distribution plotted in blue.*

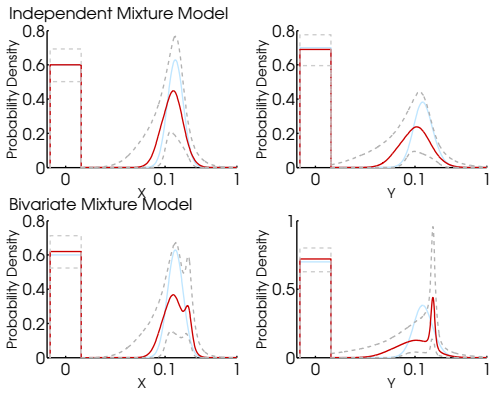
(a) Validation Data Set A



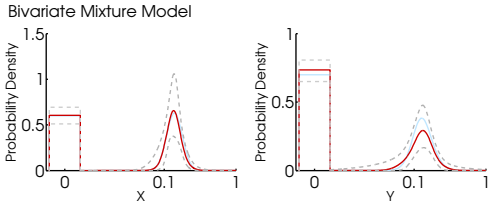
(b) Validation Data Set B



(c) Validation Data Set C - non-informative prior distributions



(d) Validation Data Set C - bivariate mixture model using weakly informative prior distributions



It is clear from Figure 3.2a that the bivariate mixture model results in a good description of the true marginal distributions for data set A, whereas the independent mixture model struggles to handle the bimodal character of the marginal distributions. As the independent mixture model tries to fit a single Normal distribution to the log-residue values, it compensates for the observed bimodality by increasing the variance, leading to a poor fit as seen in Table 3.2. The model also assumes that the log-residue data for X and Y are independent and therefore provides no indication of the correlation. In contrast, the bivariate mixture model provides a good estimate of the correlation between the log-residue levels (Table 3.2).

Figure 3.2b clearly shows that both models provide a good description of the true marginal distributions for validation data set B. Although both models provide good parameter estimates, the independent mixture model results in better estimates for some parameters and narrower credible intervals than the bivariate mixture model (Table 3.3). This is because the bivariate mixture model assigns a proportion of the observed data to four different distributions (untreated, treated with X only, treated with Y only or treated with both X and Y) and therefore fewer data are available to estimate the distribution parameters for each distribution. However the bivariate mixture model provides a good estimate of the correlation between log-residue levels which the independent mixture model ignores. Therefore overall the bivariate mixture model results in a better representation of validation data set B.

For validation data set C, which has a high proportion of values below the LOD, it is clear from Figure 3.2c that the low number of data has a strong influence on both models when non-informative prior distributions are used. For the bivariate mixture model it means that the estimate of the proportion of data that are not treated and the marginal posterior distributions for X and Y are more uncertain than for validation data set B. In addition, the mean and median estimates of the correlation are reasonable but the credible interval is wider than for validation data set B. The independent mixture model performs slightly better than the bivariate mixture model in terms of estimating the distribution parameters (Table 3.4) for

the same reasons explained for validation data set B. Overall the bivariate mixture model results in a better representation of validation data set C because it takes the correlation between the log-residue values into account. However as there are few data available it is interesting to investigate if the bivariate mixture model can be improved by using more informative prior distributions. Figure 3.2d shows that using weakly informative prior distributions improves the estimates of the model parameters and reduces the uncertainty about them. Therefore it would be advisable to incorporate any relevant information available into the prior distributions to improve the model performance.

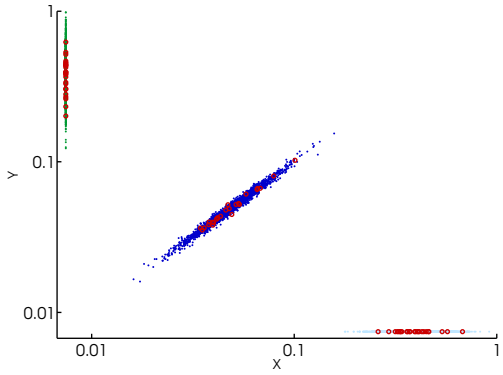
To assess in more detail how well the bivariate mixture model describes the correlation, we generate log-residue level predictions from the posterior distributions. These samples are shown together with the validation data sets A, B and C in Figure 3.3. For data sets A and B, the predictions closely follow the observations. For data set C it is harder to assess the performance as there are fewer data available. However, the model appears to do well using both the non-informative and weakly informative prior distributions.

3.6.2 Comparison with current approaches

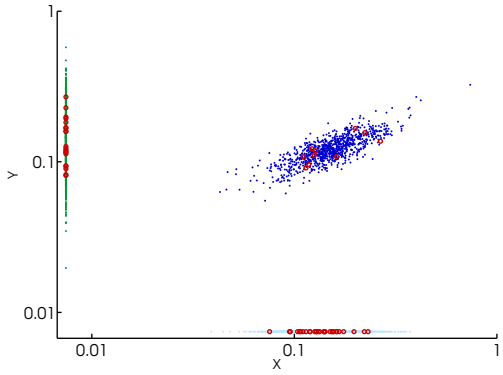
In this section we will compare the new approaches with the current approaches for cumulative risk assessments (see Section 3.3) using the validation data sets A, B and C. We have described the method from Paulo et al. (2005) previously so here we only briefly illustrate how the pairwise and independent empirical sampling approaches are applied to data using validation data set C. Bootstrap approaches are used to account for uncertainty about the log-residue levels. We will now refer to these methods which use bootstrapping of the empirical distribution of the log-residue data as pairwise and independent bootstrap. Figure 3.4 shows the predictive distribution when values below the limit of determination were set to zero (i.e. untreated) for each bootstrap method. The resulting samples do not reflect the information from the validation scenario (see Table 3.6). If we had instead set all values below the LOD to (a proportion of) the LOD (0.01 for both X and Y), all the

Figure 3.3 – *Predictive sample (dark blue dots for predictions 'treated' with both pesticides, light blue dots for predictions 'treated' with X only, green dots for predictions 'treated' with Y only) obtained from applying the mixture model to the validation data sets (red circles). Samples from the univariate components are plotted along the axes. Red dashed lines indicate the LOD where applicable.*

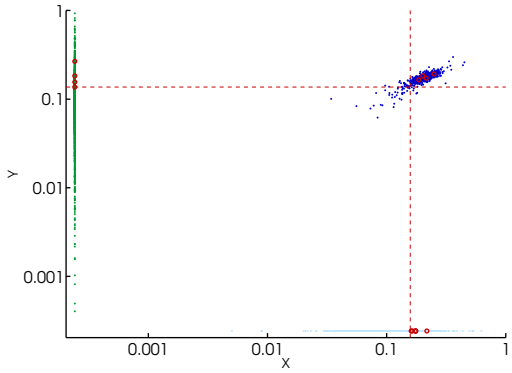
(a) Validation Data Set A



(b) Validation Data Set B



(c) Validation Data Set C using non-informative prior distributions.



(d) Validation Data Set C using weakly informative prior distributions.

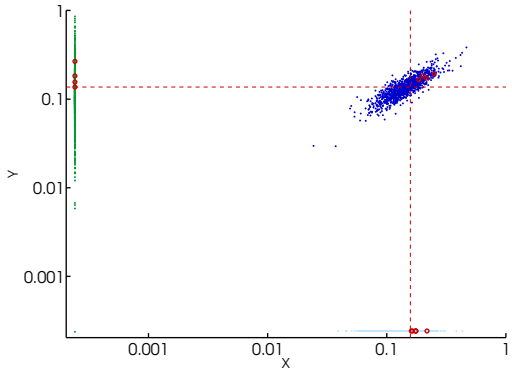


Figure 3.4 – *Predictive empirical bootstrap samples for validation data set C. The percentage next to each sample indicates the percentage of the samples at that residue value. Labels for values which were sampled less frequently than 0.9% are omitted for clarity.*

(a) Pairwise **(b)** Independent

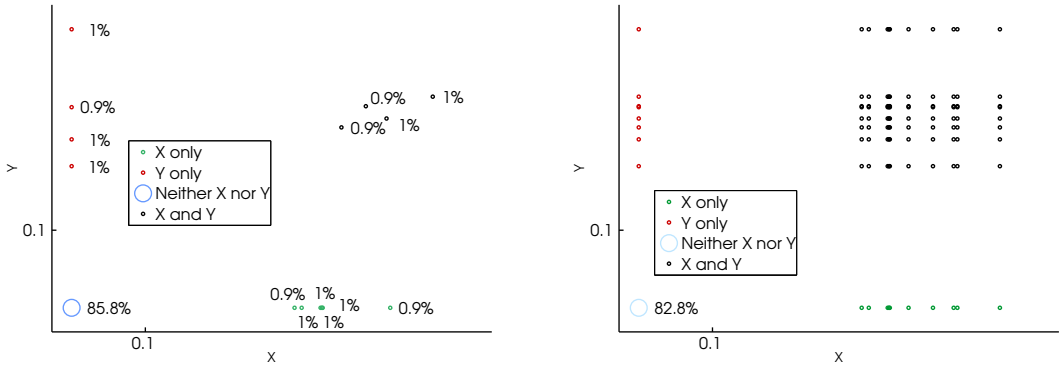


Table 3.6 – *Comparison of true proportions of units having received a certain treatment type and predictions of those proportions using the pairwise and independent bootstrap approaches for validation data set C.*

Treatment	True Proportion	Pairwise Bootstrap	Independent Bootstrap
Untreated	0.4	0.86	0.83
X	0.3	0.06	0.09
Y	0.2	0.04	0.07
X and Y	0.1	0.04	0.01

samples would have been assumed to be treated with both pesticides, which does not reflect the information we have from the validation scenario either (10% (true values) vs 100% (both bootstrap approaches) treated with X and Y). This shows that empirical bootstrap approaches cannot deal with censored data very well as the censored data will have to be set to 0, the LOD or a fraction of the LOD. Another issue with using the empirical distribution is that it is very unlikely that only 15 (pairwise) or 99 (independent) distinct sets of values are possible and that residue levels will not be higher than the highest observation. The independent bootstrap approach cannot be used to model correlations in residue levels. The pairwise boot-

strap approach may provide a good estimate of the correlation for large data sets with little or no censoring but will not be suitable for residue data sets where there are often few positive residue data values. Comparing the results with the predictive distributions from the bivariate mixture model (Figures 3.3c and 3.3d) shows that the bivariate mixture model allows a larger number of distinct values and accounts for the uncertainty in the correlation and distribution parameter estimates caused by the small number of samples and high levels of censoring and therefore provides a better description of the data.

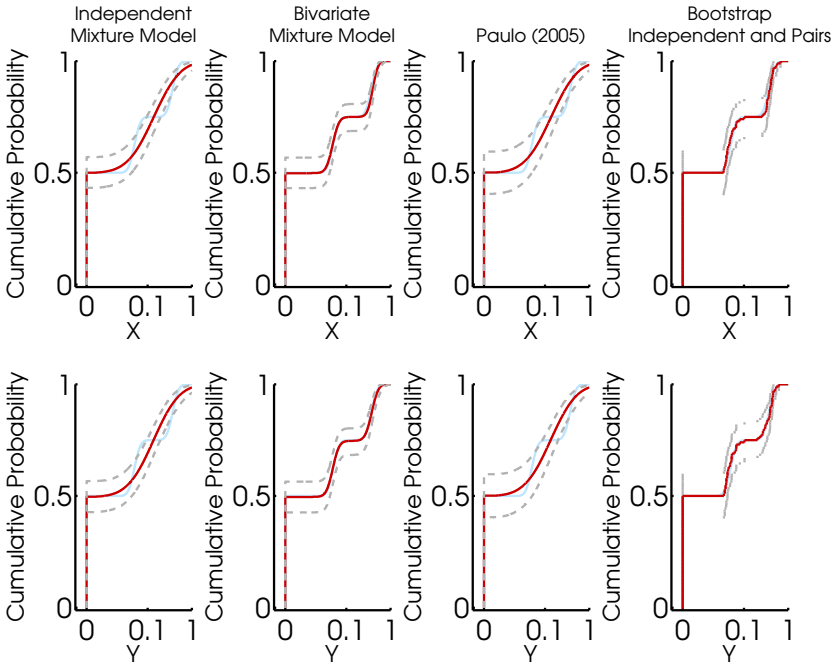
Marginal posterior cumulative distribution functions for all the methods (bivariate mixture model, independent mixture model, Paulo et al. (2005) and the independent and pairwise bootstrap) are shown in Figures 3.5 and 3.6. As the marginal distributions for the pairwise and independent bootstrap are very similar, we only display the independent case. For validation data set A, the independent mixture model and Paulo et al. (2005) both perform poorly because they cannot describe the bimodal nature of the target distribution. The bivariate mixture model and bootstrap approaches appear to do better. However the independent bootstrap assumes that pesticide residue levels are independent and therefore will not provide an estimate of the correlation in residue levels. The pairwise bootstrap will provide a good estimate of the correlation and the uncertainty of the correlation coefficient because there is no censoring. However the bivariate mixture model seems to be the best approach of those tested because it provides an estimate of the uncertainty of the correlation coefficient of log-residue levels and it allows values other than those observed in the data set to be sampled.

The results for validation data set B show that all methods perform well for this data. The bivariate mixture model and independent mixture model are less uncertain than the Paulo et al. (2005) and bootstrap approaches due to the use of the GB PUS data. Again the bivariate mixture model and pairwise bootstrap are the only methods to provide an uncertain estimate of the correlation coefficient.

For the heavily censored validation data set C, the bivariate mixture model provides an acceptable estimate of the true marginal distribution when using non-informative

Figure 3.5 – *Median (red line) and 95% credible intervals (grey dashed lines) of the marginal posterior distributions for validation data sets A and B inferred using various methods with the target distribution plotted in blue.*

(a) Validation Data Set A



(b) Validation Data Set B

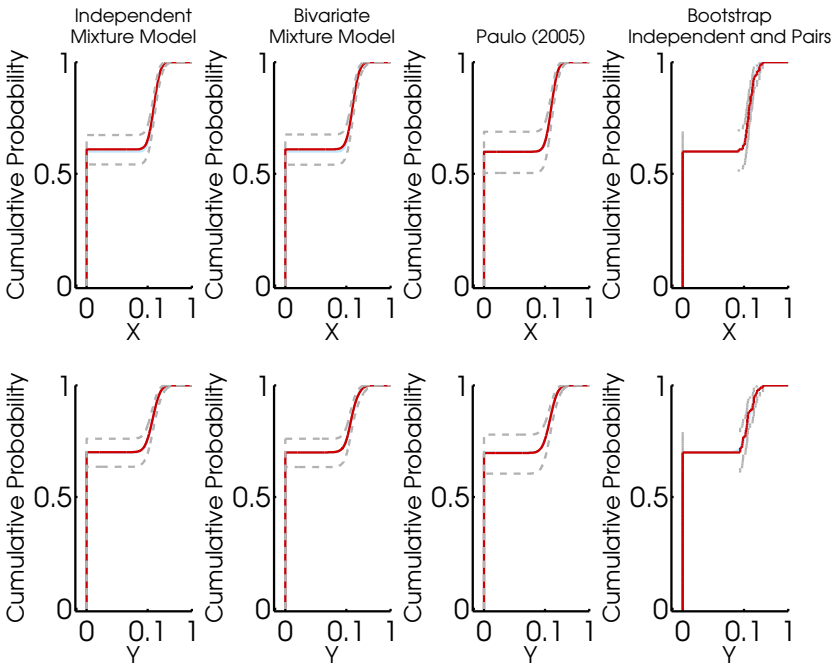
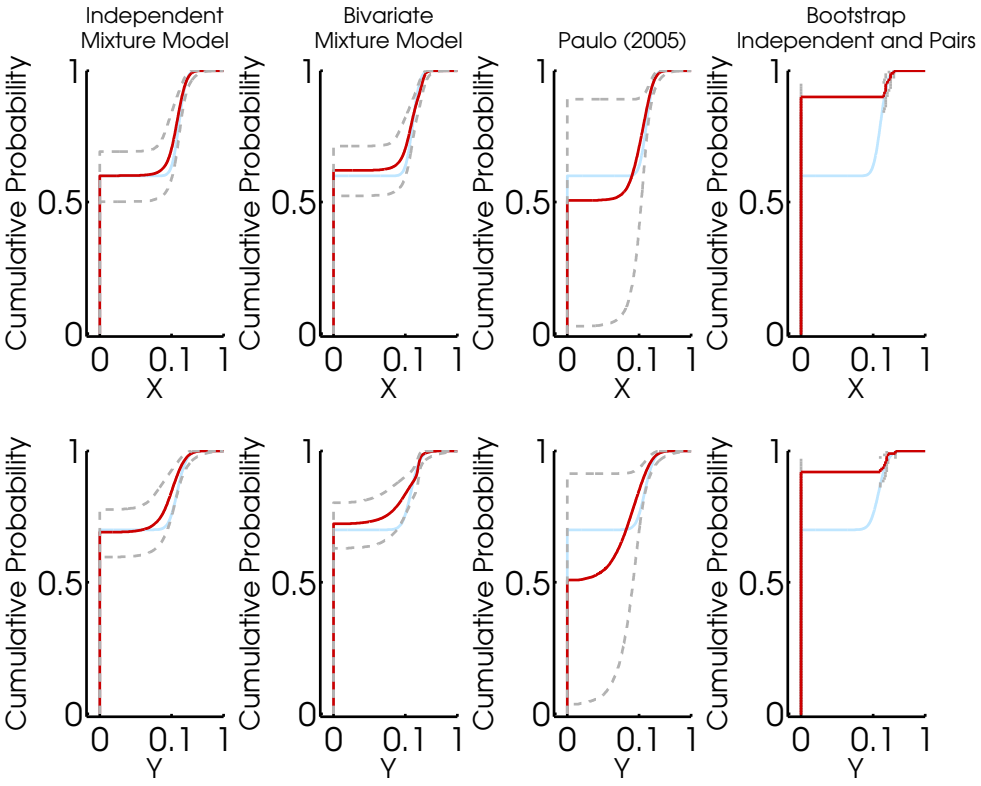
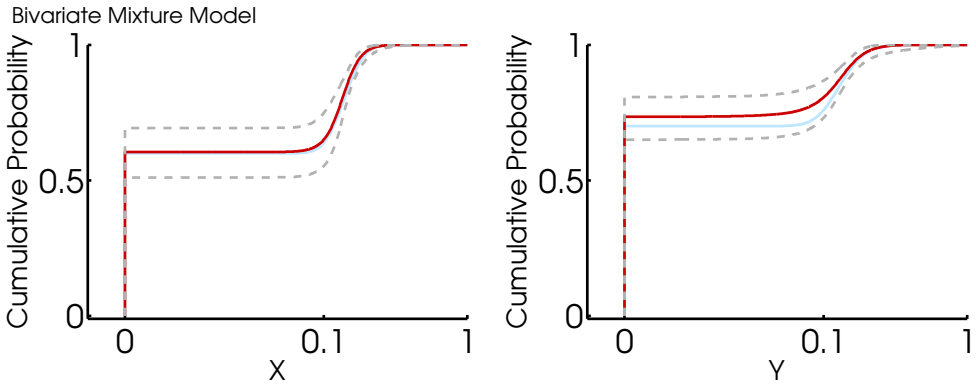


Figure 3.6 – *Median (red line) and 95% credible intervals (grey dashed lines) of the marginal posterior distributions for validation data set C inferred using various methods with the target distribution plotted in blue.*

(a) Validation Data Set C using non-informative prior distributions



(b) Bivariate mixture model with weakly informative prior distributions



prior distributions and an improved estimate when using weakly informative prior distributions. The marginal distributions of the independent mixture model look reasonable but the method incorrectly assumes independence between log-residue levels of X and Y and therefore does not provide a good description of the underlying distributions. As the Paulo et al. (2005) method does not make use of PUS data, it results in very uncertain estimates of the marginal distributions. It also does not account for any correlations between pesticide residues. The performance of the bootstrap approaches is very poor, regardless of whether the censored data are considered to be untreated or set to a proportion of the LOD. As the bootstrap approaches also underestimate the uncertainty in the correlation for validation data set C (see Figure 3.4), other methods may be a better choice when analysing data sets which are heavily censored.

Overall, to describe the variation in log-residue levels in composite monitoring data, the bivariate mixture model seems to provide the best description of the data sets used in this section, particularly if expert knowledge is available to define the prior distributions. The independent mixture model performs well for unimodal distributions and is applicable when there is no evidence of correlations between the log-residue levels. The use of PUS data results in a reduction of the uncertainty so whenever PUS data are available, they should be considered to provide an initial estimate of treatment proportions. Empirical bootstrap approaches seem to provide a reasonable estimate when large data sets are available that do not contain any censored data. As this is unlikely to be the case when modelling log-residue levels, empirical bootstrap approaches may be inappropriate for modelling residue levels.

3.7 Case Study

In this section we show the results of our new approaches for triazoles which are a group of chemicals used as fungicides on carrots. To apply the proposed approaches in a case study, we need monitoring data on residue levels on carrots and PUS data on pesticide treatments for carrots. For this case study, monitoring data on car-

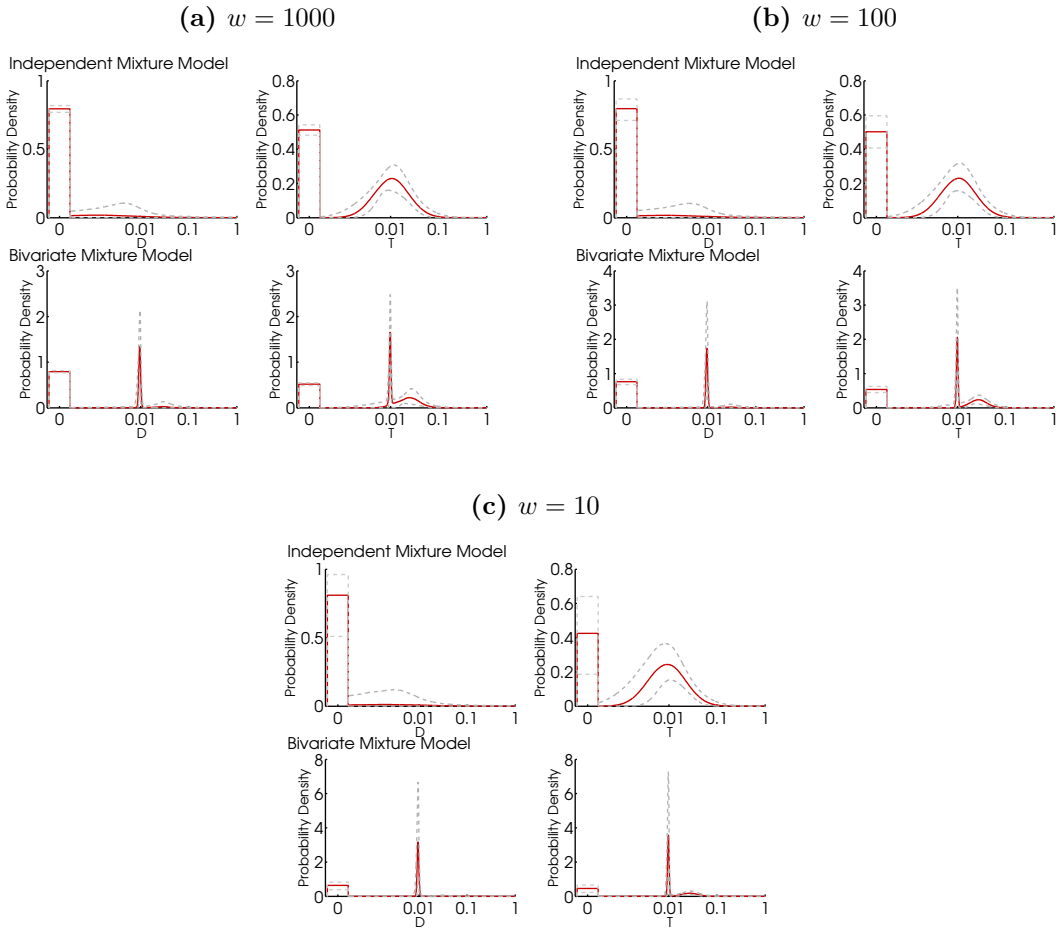
Table 3.7 *– Summary of UK monitoring data for carrots for triazoles Difenoconazole (D) and Tebuconazole (T). Values within brackets provide the number of times a value was observed.*

D	T	Number of samples	Proportion
<LOD	<LOD	68	0.71
0.03	<LOD	1	0.01
<LOD	0.01 (10), 0.02 (9), 0.03 (2), 0.04 (4), 0.06 (1)	26	0.27
0.01	0.01	1	0.01

rots from the UK for the triazoles difenoconazole (D) and tebuconazole (T) were obtained from PRC reports for 2008 (PRC, 2008; PRC, 2009) and are summarised in Table 3.7. Out of 96 values, only 1 sample contained residue levels above the LOD of both D and T and in total only 28 samples had detectable values. The PUS data from 2007 for carrots (Fera, 2011) indicate that 46.1% of carrot fields are not treated with any triazoles, 5.3% are treated with D, 33.4% are treated with T and 15.3% are treated with D and T.

We ran the independent mixture model and the bivariate mixture model on the carrot data using the non-informative prior distributions described in Table 3.1 and with differents weights, w, which reflect our belief in how representative the PUS data are for the residue data set. The marginal posterior distributions for D and T are shown in Figure 3.7 and the posterior distributions for the weights $\boldsymbol{\alpha}$ are shown in Figure 3.8. As only one observation was available with residue levels above the LOD for both D and T, the only information the model has about the variation in residue levels for samples treated with both D and T comes from the choice of prior distribution. As the chosen non-informative prior distribution did not suggest large variation for samples treated with both D and T, the bivariate mixture model has a sharp peak at 0.01 for both D and T. One could argue that this is unreasonable, but a counterargument would be that we have not observed any variation in the monitoring data, so there is no evidence to support large variation in the bivariate Normal distribution. If there was evidence from other sources, this should be in-

Figure 3.7 – *Median (red line) and 95% credible intervals (grey dashed lines) of the marginal posterior distributions of D and T for both the independent and bivariate mixture models applied to the UK carrot data set with different prior weights, w, for the GB PUS data.*

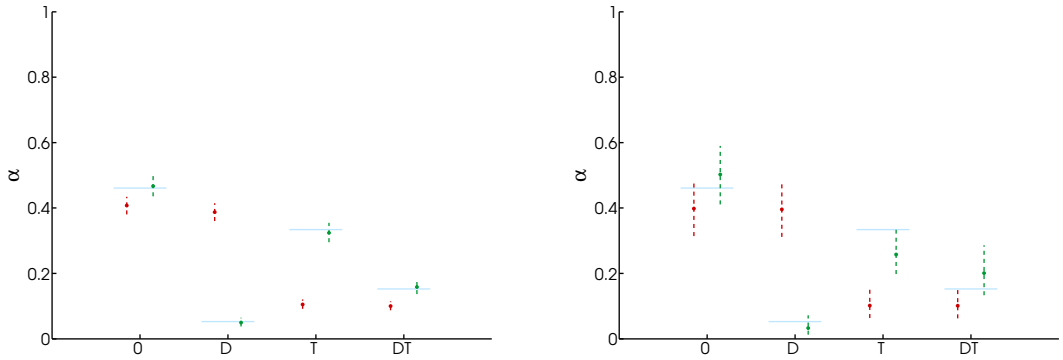


cluded in the analysis by using a different prior distribution. The peaks seen in the bivariate mixture model are not seen in the independent mixture model because in the latter a single distribution is used to describe all the data for D and another distribution is used for T.

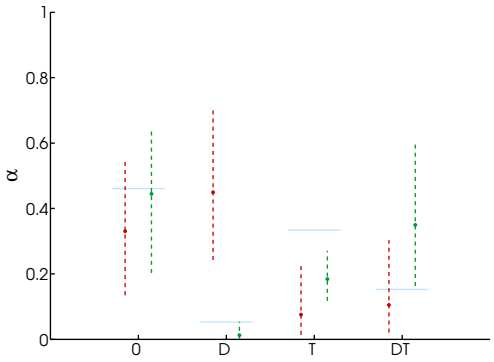
The low number of positive data values means the results depend strongly on the choice of the prior distributions. Therefore it is important that the chosen prior distributions reflect our beliefs. It is clear from Figure 3.8, where we show posterior weights $\boldsymbol{\alpha}$, that the stronger our belief in the treatment probabilities from the GB

Figure 3.8 – *Median (dots) and 95% credible intervals of the posterior weight distributions, $\boldsymbol{\alpha}$, for both the independent (red line) and bivariate (green line) mixture models applied to the UK carrot data set with different prior weights, w, for the GB PUS data (blue line).*

(a) $w = 1000$ **(b)** $w = 100$



(c) $w = 10$



PUS data, i.e. the higher the value of w, the lower the uncertainty about the probabilities $\boldsymbol{\alpha}$. This is because if $w = 1000$, the prior sample size has more influence on the posterior distribution of $\boldsymbol{\alpha}$ than the data sample size of 96. When w is smaller the data drives the model leading to values of $\boldsymbol{\alpha}$ that are less influenced by the evidence provided by the PUS data. This can be seen in the results for the bivariate mixture model where for $w = 1000$ the posterior distributions of $\boldsymbol{\alpha}$ are in agreement with the PUS values. For $w = 10$ and $w = 100$ the posterior estimates of $\boldsymbol{\alpha}$ are influenced by both the PUS data and the log-residue data. Therefore it is important to use a value for w that reflects our belief in how representative the PUS data are of

the proportion of samples receiving a certain treatment. For the independent mixture model, the posterior distribution for $\boldsymbol{\alpha}$ does not follow the PUS data estimates for any value of w as a result of the assumption that the distributions for D and T are independent. Since we do not know what the true values are for any of the model parameters, we cannot assess the model performance and therefore we do not compare the results with current approaches.

3.8 Extension of model to predict unit residue levels

The approaches presented in this chapter offer improved modelling of pesticide residues in composite samples. However, for an acute dietary risk assessment, we need a model that can also simulate correlations in residue levels between units. Little information is available about how composite samples are generated and a composite sample could consist of units that come from fields that have received different treatments. For example, for a composite sample consisting of 5 units, it is unknown whether a sample consists of 5 units from a single field treated with just a single pesticide or a range of pesticides, 5 units from different fields treated with a single (but possibly different) pesticide or 5 units from different fields treated with a range of pesticides.

We first show that information on correlations on residue levels between composite samples may provide little information on the correlations on residue levels in units using an example. In Table 3.8 we consider three simulated composite samples. The unit correlations are 0.01 for the units in composite sample 1, 0.01 for the units in composite sample 2 and 0.02 for the units in composite sample 3. However, the correlation coefficient for the three composite samples 1, 2 and 3 is 0.99. If we sort the columns to induce high correlations between the units, the correlation coefficient for the composite samples will stay the same. This indicates that knowledge of the correlations in the composite samples may not provide any information on unit correlations. Therefore, to model correlations in unit residues, we propose

Table 3.8 *– Simulated unit residue level data used to explain the limited relationship between correlations in unit residue levels and correlations in composite sample residue levels.*

	Composite 1		Composite 2		Composite 3	
Units	X	Y	X	Y	X	Y
1	2.1	3.3	0.9	1.8	5.1	2.1
2	2.5	8.2	1.7	6.1	4.8	11.9
3	3.8	3.5	2.6	2.4	8.0	6.7
4	3.5	5.7	2.5	4.0	5.7	8.1
5	5.5	6.5	3.7	4.9	8.4	5.8
6	5.9	5.0	3.8	3.2	7.9	4.6
7	6.5	7.2	4.7	5.3	9.4	9.6
8	7.9	4.5	5.6	2.8	9.9	6.1
9	8.6	8.7	6.2	6.4	9.2	9.6
10	9.4	2.7	7	1.4	12.5	7.1
Composite Residue Level:	5.6	5.5	3.9	3.8	8.1	7.2
Unit Correlation:	0.01		0.01		0.02	
Composite Correlation:			0.99			

the following solution. Firstly, model correlations in composite samples to generate composite residue levels for dietary modelling and then model unit variation using various scenarios (see Table 3.9). If there is at least some variation between unit residue levels (Scenarios A, B, D, E and F in Table 3.9), we could model the correlations in residue levels using multiple scenarios. This seems to be the only feasible solution when little is known about unit residue levels in units available on the market. Depending on the selected scenario, unit residue levels can be generated by assigning all residues to a single unit (Scenarios A and B), assigning the same residue level to every unit (Scenario C) or by assigning different residue levels to each unit (Scenarios D-F). For Scenarios D-F, the level of heterogeneity, which is a measure of how much variation there is likely to be between units in a composite sample, has to be chosen and unit values will be sampled in each iteration of the Monte Carlo procedure. This could easily be implemented in current dietary risk assessment software by adding a heterogeneity variable. Once the values are sampled,

Table 3.9 *– Scenarios for unit variation modelling based on composite residue values 10 and 20 for pesticide X and Y, respectively, which consist of 5 units all assumed to be of equal weight. The numbers presented here represent a single iteration in a Monte Carlo simulation.*

Residue on a single unit				All units treated							
Scenario A: one unit treated with both pesticides		Scenario B: different units treated with a single pesticide		Scenario C: same concentration on each unit		Varying concentration					
						Scenario D: independent		Scenario E: strong positive correlation		Scenario F: strong negative correlation	
X	*Y*	*X*	*Y*	*X*	*Y*	*X*	*Y*	*X*	*Y*	*X*	*Y*
50	100	50	0	10	20	2.5	2.5	2.5	2.5	2.5	45.5
0	0	0	100	10	20	3.5	9	3.5	9	3.5	28.5
0	0	0	0	10	20	7	45.5	7	14.5	7	14.5
0	0	0	0	10	20	12	28.5	12	28.5	12	9
0	0	0	0	10	20	25	14.5	25	45.5	25	2.5

rank correlations can be induced according to the chosen scenario. If the choice of scenario has a significant impact on the outcome of the dietary risk assessment, one could consider measuring residue levels on units to obtain a better understanding of which of the unit modelling scenarios is most likely to reflect the distribution of residue levels on units. As this is likely to vary between analyses the scenario approach provides a pragmatic way of exploring the possible residue levels on unit food items.

3.9 Discussion

This chapter discussed various techniques to model pesticide residue levels for cumulative exposure assessments and introduced two new approaches that are able to combine information on pesticide usage with data on residue levels from monitoring programmes. One of the new approaches is also able to account for correlations in residue levels in composite samples. The approaches have been tested alongside alternative approaches in a series of validation studies (see Section 3.6.2). The results of these validation studies indicate that the bivariate mixture model offers a more flexible way of describing residue levels of multiple pesticides in food products which can be used in cumulative risk assessments. The bivariate mixture model is the only

method that provides an accurate estimate for the correlation in log-residue levels whilst accounting for uncertainty when there is high censoring and few data values. The independent mixture model provides a good estimate if the log-residue levels are independent. Both models seem to provide a better description of the target distribution in comparison with existing approaches. When there are high levels of censoring, e.g. validation data set C, the use of informative prior distributions seems to provide a transparent approach for predicting residue levels in food items which allows for an assessment of the impact of choosing different prior distributions. For example, the PUS data were used to inform the prior distributions of the proportion of composite samples that have received a certain treatment. This led to both proposed mixture models providing a better description of residue levels in composite samples than existing approaches when censoring levels are high. The PUS data could also be used to assess whether pesticide combinations were applied in a tank mix or not, to help inform the prior distribution on correlations.

The bivariate mixture model presented in this chapter is based on mixtures of univariate and bivariate Normal distributions. As the number of parameters increases considerably for more than two pesticides, it is clear that they cannot be estimated reliably from the limited number of monitoring data that is generally available. As a result, any attempt to model the cumulative exposure for more than two pesticides will have to rely heavily on assumptions, e.g. by eliciting prior distributions from experts, collecting larger data sets than are currently available and/or reducing the number of parameters as discussed in Section 3.5.3.

When modelling the acute dietary risk associated with pesticide exposure, we need to be able to estimate residue levels on individual items. Currently, dietary risk assessments ignore whether units in composite samples originate from the same field or multiple fields because there is little information available. In the absence of data on how units are mixed before they are consumed, we have suggested a scenario-based approach (Table 3.9), but more work is needed to assess which of the proposed scenarios are realistic.

Despite issues related to the residue data available for dietary risk assessment for pesticides (see Section 1.4.1.1), which cause problems for all available methodologies, both approaches presented in this chapter make better use of the available data. Both mixture models performed well in validation studies in comparison with other available techniques and offer promising alternatives which could improve the modelling of cumulative dietary exposure.

Chapter 4

Modelling unit variation in residue data

4.1 Introduction

To assess the risk associated with acute exposure to pesticide residues, we need information on residue levels on unit food items. However, unit residue data are not routinely collected as part of the pesticide registration process. Therefore in current acute dietary risk assessments unit variation is modelled using a variability factor (see Section 1.3.3) which has been derived using data from field trials for other crops and chemicals (EFSA, 2005). For probabilistic risk assessments, a distributional form needs to be selected to describe the variation in unit residue levels. EFSA (2012) recommend the use of a Lognormal distribution, but also provide evidence that the Lognormal distribution may not always be appropriate to describe the variation in unit residue levels.

In this chapter we introduce a novel non-parametric Bayesian approach which provides a distribution of unit residue levels, which may be a better alternative to the commonly used Lognormal distribution. The approach aims to determine the location, scale and shape of log-residue distributions whilst accounting for the uncertainty of these parameters. To overcome the issue with the relatively small size

of unit log-residue data sets, the shape of the distribution will be determined by sharing information between various data sets, by assuming that they share a common shape. The shape of a probability distribution is determined by (a set of) parameter(s) that are neither location or scale parameters or functions thereof. The distribution shape is modelled using a Dirichlet Process mixture model (see Section 2.3.5). After specifying the model, we briefly explain the challenges associated with applying the model to log-residue data. The method is then tested in a wide range of simulation studies to assess the performance for different distributions and sample sizes before being applied to log-residue data. Finally, we compare the method to the current approach for describing the variation in unit log-residue levels and investigate some model refinements.

4.2 Model

In this section, we propose a novel statistical method that uses non-parametric Bayesian techniques to (a) move away from an assumption of Lognormality and (b) share knowledge from multiple data sets to learn about the distribution shape for all the pesticide/crop scenarios under consideration. The new approach is fundamentally based on the observation that populations may share certain characteristics (e.g. shape) whilst others (e.g. location, scale) will be population-specific. Even when sample sizes for the individual populations are considered to be too small to define the shape distribution, we can still use the data to learn about their locations and scales. Subsequently, we can use these characteristics to relocate and rescale (i.e. normalise) the data and pool them to obtain a larger data set from which we may be able to learn other, common characteristics, for example the distribution shape.

For our application this means that instead of analysing each pesticide/crop scenario individually, a common shape can be used for pooled log field trial data sets. This assumption is supported by an analysis of available unit field trial data (Ambrus, 1979; Ambrus, 1995; Ambrus, 2006; Holland and Malcolm, 2002; Kaethner,

2001a; Kaethner, 2001b; Tew, 1993; Valdez-Flores et al., 2002; Xu et al., 2008) which showed that the location (e.g. median) and scale (e.g. range) of log-residue data vary considerably between pesticide/crop scenarios. The analysis also indicated that there may be a common shape that is shared between several scenarios although possibly more than one shape may be needed to describe all scenarios. The advantage of sharing information between pesticides/crops is that more information will be available to estimate the shape of the distribution. As current approaches commonly assume that log pesticide residue distributions share a common Normal shape, an approach in which the common shape is learned from the log-residue data rather than selected for pragmatic reasons seems to be an improvement.

The model developed in this chapter uses a blocked Gibbs sampler which alternates sampling location and scale parameters for each data set with sampling a common shape distribution for the pooled, normalised log-residue data. Figure 4.1 shows an overview of one iteration of this blocked Gibbs sampler for three fictitious pesticide log-residue data sets. After taking logs of the unit residue data, we sample a location and scale parameter from the posterior distribution of each of the three pesticides in each iteration (given the log-residue data and current realisation of the shape distribution). These will then be used to normalise the log-residue data. After pooling this normalised data, a new realisation of the shape is sampled from the posterior shape distribution (given the log-residue data and current location and scale parameters). This Bayesian approach will account for the uncertainty of the distribution parameters caused by the limited size of the data sets. When these steps are repeated we will ultimately obtain an uncertain distribution over distributions of pesticide log-residues.

We use a Dirichlet Process Mixture of Normal distributions (DPMN; see Section 2.3.5) to learn about the shape of the log-residue distribution. This may provide a better way to describe pesticide log-residue levels on unit food items than simply assuming a Normal distribution.

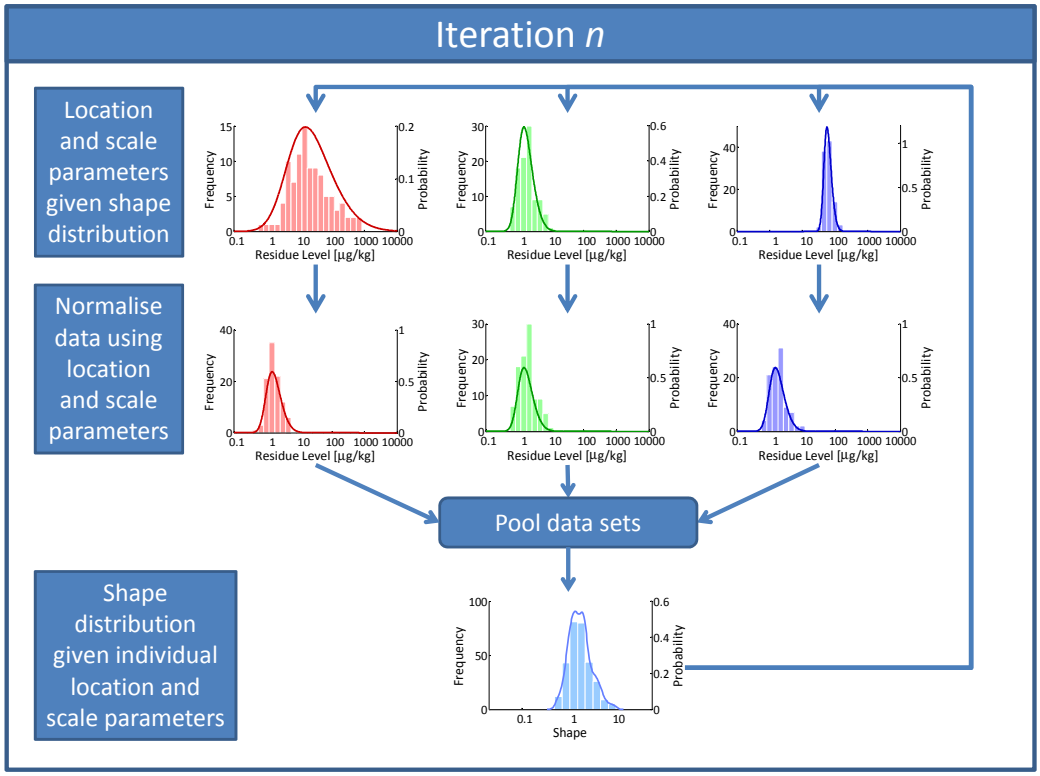


Figure 4.1 – *Graphical overview of the proposed blocked Gibbs sampler to describe variation in unit log-residue levels. In each iteration, the fictitious pesticide log-residue data sets (3 in this example) will be normalised using a sample from the posterior distribution of their location and scale parameters given the current shape distribution. Subsequently the normalised data will be pooled to obtain a single shape distribution given the current location and scale parameters. After n iterations, we will obtain n samples from the posterior distributions of the location, scale and shape parameters.*

The shape, location and scale parameters will together define the distribution of pesticide log-residue levels on units for existing pesticide/crop data sets. This distribution can either be used to infer a more realistic estimate of the variability factor or to model unit variation if a second model is available to model the distribution of field means, as explored in Chapter 5.

4.2.1 Inference for the distribution shape

In this section, we discuss how to model the distribution shape of log-residues with a DPMN model. DPMNs offer a flexible approach to distribution fitting which assumes that the observed data set is a random sample from a population distribution that consists of a mixture of an infinite number $k = \{1, 2, \dots\}$ of Normal distributions, $\mathcal{N}(\theta_k, \sigma_c^2)$, each with relative weight w_k, where θ_k is the location parameter of component k and σ_c^2 is the fixed variance for all of the components. We select Normal distributions for the components because this leads to a simple conjugate Bayesian update of the distribution parameters. DPMN models have mostly been used to describe the population distribution given an observed data set (Escobar and West, 1995; Ishwaran and Zarepour, 2000; Neal, 2000; Papaspiliopoulos and Roberts, 2008). However, in the model presented here, we want to use the DPMN model to describe the shape of the unit log-residue distribution. As a consequence, we want the location and scale parameters of the DPMN to be zero and one respectively. This will have an impact on both the DPMN itself and the prior distribution, G_0, which we will discuss in the next two sections.

4.2.1.1 Relation between location, scale and shape parameters

In the approach presented here, the data $\tilde{\mathbf{y}}$ will be the pooled normalised unit log-residue data for J pesticide/crop combinations. This pooled data, $\tilde{\mathbf{y}}$, will be used to infer the distribution shape using a DPMN model. Applying a standard DPMN model would not restrict the location and scale parameters so we instead use a different approach to ensure that the prior shape distribution has location and scale parameters of approximately zero and one, respectively. To do this we split the Dirichlet Process into three separate processes on the intervals $(-\infty, -\phi)$, $(-\phi, \phi)$ and (ϕ, ∞), where the first and last intervals have probability p and the middle interval has probability $1 - 2p$. If we define ϕ as a quantile of the distribution, we automatically obtain p. A convenient choice would be to map $-\phi$ onto the lower tertile and ϕ onto the upper tertile. This way, each of the three Dirichlet Processes has a probability of a third. The choice of tertiles seems logical because the shape

distribution can be divided into three parts with equal probability and tertiles are relatively robust statistics to estimate, i.e. they are not affected as much by outliers as more extreme quantiles, while also being able to capture the scale of the data. Now we need to define the location and scale parameters, μ_j and σ_j, of the distribution for data set j as a function of ϕ in such a way that the normalised log-residue data will have location zero and scale one. If $Q_j^{\frac{i}{3}}$ are the i^{th} tertiles of the j^{th} log-residue data set we can ensure that the scale for the j^{th} data set is one using:

$$\sigma_j = \frac{Q_j^{\frac{2}{3}} - Q_j^{\frac{1}{3}}}{2}$$

To centre the normalised shape distribution around 0 we use:

$$\mu_j = \frac{Q_j^{\frac{2}{3}} + Q_j^{\frac{1}{3}}}{2}$$

Thus $Q_j^{\frac{2}{3}} = \mu_j + \sigma_j$ for each log-residue data set, the first tertile of each of the normalised log-residue data sets is -1 and the second tertile is 1.

4.2.1.2 Prior distribution G_0 and consequent calculations

Now that the location and scale parameters, μ_j and σ_j, are defined we need to infer the distribution shape using the DPMN model. If n_j defines the number of data for each pesticide/crop population j, the first step is to normalise the log-residue data y_{ij} with $i = 1, ..., n_j$ for the j^{th} pesticide/crop data set:

$$\tilde{y}_{ij} = \frac{y_{ij} - \mu_j}{\sigma_j} \tag{4.1}$$

Then we can pool the normalised log-residue data $\tilde{\mathbf{y}} = \{\tilde{y}_{11}, \ldots, \tilde{y}_{n_J J}\}$ where J is the number of pesticide/crop populations for which we have data. We need to select a prior distribution G_0, where $G_0(\theta) = F(\theta; \mu_0, \sigma_0)$, for the location parameters of the Normal components, θ_k. We will discuss two possible distribution shapes for F below, but first we focus on the prior location and scale parameters μ_0 and σ_0 and the fixed component variance σ_c^2. The selection of values for μ_0, σ_0 and σ_c^2 is critical for the performance of the model. Given that the log-residue data that will be used in the model are normalised and will approximately have location zero and scale

one, it seems logical to set $\mu_0 = 0$. If we assume F is a $\mathcal{N}(\mu_0, \sigma_0^2)$ distribution, we have:

$$\theta_k \sim \mathcal{N}(\mu_0, \sigma_0^2)$$
$$\tilde{y}_{ij}|K_{ij} = k \sim \mathcal{N}(\theta_k, \sigma_c^2)$$

where K_{ij} is an index parameter, indicating which component $\tilde{y}_{ij}$ is assigned to. If we set $\mu_0 = 0$, we need to select σ_0 in such a way that a-priori the first and second tertiles of $\tilde{\mathbf{y}}$ will be at -1 and 1, respectively. If we assume that $\sigma_c = 0$, i.e. all the observed variation in $\tilde{\mathbf{y}}$ is a result of variation in the location parameters of the Normal components, we can define σ_0 in such a way that the following condition is true:

$$\frac{\Phi^{-1}(2/3; 0, \sigma_0^2) - \Phi^{-1}(1/3; 0, \sigma_0^2)}{2} = 1$$

where $\Phi^{-1}(p; \mu, \sigma^2)$ is the inverse of the cumulative Normal distribution. Due to symmetry, this can be rearranged to result in:

$$\sigma_0 = \frac{1}{\Phi^{-1}(2/3; 0, 1)} \approx 2.32$$

Instead of assigning this variance to the prior distribution, G_0, we can use this value as the total observed variance. Then, to define the variance, σ_0^2, of G_0 assuming that the component variance parameters, σ_c^2, are fixed but non-zero, we can make use of the fact that the observed variance is the sum of the variance of the component locations, σ_0^2, and the component variance, σ_c^2:

$$\sigma_{observed}^2 = \sigma_0^2 + \sigma_c^2$$

We now define κ as a factor indicating the proportion of $\sigma_{observed}$ that is assigned to the Normal components:

$$\begin{aligned}\sigma_c &= \kappa \sigma_{observed} \\ &= \frac{\kappa}{\Phi^{-1}(2/3, 0, 1)}\end{aligned}$$

and the proportion that is assigned to σ_0:

$$\begin{aligned}\sigma_0 &= \sqrt{\sigma_{observed}^2 - \sigma_c^2} \\ &= \frac{\sqrt{1 - \kappa^2}}{\Phi^{-1}(2/3, 0, 1)}\end{aligned}$$

We can either define κ, σ_c or σ_0 given the predefined tertiles. Of these three, it is easiest to define κ as it is a value on the interval $[0, 1]$. Note that the derivation above is equally valid for other symmetric prior distribution shapes for G_0. When G_0 is not a Normal distribution we need to set the parameters of G_0 in such a way that the normalised log-residue data, $\tilde{\mathbf{y}}$, generated from a $\mathcal{N}(\theta_k, \sigma_c^2)$ distribution with $\theta_k \sim G_0(\mu_0, \sigma_0^2)$ have location parameter zero and scale parameter one. This can easily be achieved by selecting σ_c and using a numerical solver to find the value of σ_0 that leads to tertiles of $\tilde{\mathbf{y}}$ at -1 and 1.

Even though DPMN models allow σ_c to be uncertain (i.e. to be inferred from observations) and to vary between components, the restriction of having a scale parameter of approximately one for the shape distribution, means that we assume σ_c is fixed and known. It is important to realise that the model as defined above will result in some leaching of probability beyond the tertile borders as the infinite tails of the Normal components will stretch beyond them. This will result in the scale of the shape distribution not being precisely equal to one, but this is accounted for in the sampling of the pesticide/crop scale parameters σ_j.

Choice of prior distribution G_0

Several functional forms are available for G_0. In the following sections we will explore the Normal and Student's t distributions as prior distributions for θ.

$\mathcal{N}(\mu_0, \sigma_0^2)$ *prior with known* σ_0

The case with $G_0(\theta) = \mathcal{N}(\theta; \mu_0, \sigma_0^2)$ has already been mentioned above, but will be discussed here in more detail for completeness. Let us define a Normal prior distribution, $\pi(\theta_k)$, for each location parameter, $\theta_1, \ldots, \theta_C$, with mean, μ_0, and variance, σ_0^2:

$$\pi(\theta_k) = \mathcal{N}\left(\theta_k; \mu_0, \sigma_0^2\right)$$

After assigning the normalised log-residue data $\tilde{\mathbf{y}} = \{\tilde{y}_{11}, \ldots, \tilde{y}_{n_j J}\}$ to one of the Normal components $k = 1, \ldots, C$ with location parameter, θ_k, and known standard

deviation, σ_c, the likelihood function becomes:

$$p(\tilde{\mathbf{y}}|\boldsymbol{\theta}, \mathbf{K}) \propto \prod_{j=1}^{J} \prod_{i=1}^{n_j} \sigma_c^{-1} \exp\left[-\frac{1}{2\sigma_c^2}(\tilde{y}_{ij} - \theta_{K_{ij}})^2\right] \tag{4.2}$$

As this factorises, we can regard each component location, θ_k, in turn and focus on a subset of the data $\tilde{\mathbf{y}}_{[k]}|\mathbf{K}$ where $\tilde{\mathbf{y}}_{[k]}$ are those values of $\tilde{\mathbf{y}}$ that are currently allocated to component k. The likelihood function is given by:

$$p(\tilde{\mathbf{y}}_{[k]}|\boldsymbol{\theta}, \mathbf{K}) \propto \sigma_c^{-m_k} \exp\left[-\frac{1}{2\sigma_c^2} \sum_{j=1}^{J} \sum_{i=1}^{n_j} \delta_{K_{ij},k}(\tilde{y}_{ij} - \theta_k)^2\right]$$

where $\delta_{i,j}$ is the Kronecker delta function and the number of data allocated to component k, m_k, is given by :

$$m_k = \sum_{j=1}^{J} \sum_{i=1}^{n_j} \delta_{K_{ij},k} \tag{4.3}$$

The posterior distribution is now given by:

$$p(\theta_k|\tilde{\mathbf{y}}, \mathbf{K}) = \mathcal{N}\left(\theta_k; \frac{\mu_0\sigma_c^2 + \sigma_0^2 \sum_{j=1}^{J} \sum_{i=1}^{n_j} \delta_{K_{ij},k}\tilde{y}_{ij}}{\sigma_c^2 + m_k\sigma_0^2}, \frac{\sigma_0^2\sigma_c^2}{\sigma_c^2 + m_k\sigma_0^2}\right) \tag{4.4}$$

Thus using a Normal prior distribution for θ_k leads to a simple conjugate update once the normalised data have been allocated to the components. When many data are allocated to component k, the component becomes approximately fixed in that location. For components to which no data are assigned, i.e. $m_k = 0$, the locations follow the prior G_0, $\theta_k \sim \mathcal{N}(\mu_0, \sigma_0^2)$.

Student's t_ν prior

When there is evidence to suggest that the shape distribution has longer tails than a Normal distribution, it may be better to use a Student's t distribution with ν degrees of freedom. The Student's t distribution can be represented as a mixture of a Normal distribution and a Gamma distribution:

$$\begin{aligned} p(\theta_k|\nu, \mu_0, \sigma_0^2) &= \int_0^\infty \mathcal{N}\left(\theta_k; \mu_0, \frac{\sigma_0^2}{\lambda}\right) \text{Gamma}\left(\lambda; \frac{\nu}{2}, \frac{\nu}{2}\right) d\lambda \\ &= \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sigma_0\sqrt{\nu\pi}\Gamma\left(\frac{\nu}{2}\right)} \left\{\frac{(\theta_k - \mu_o)^2}{\nu\sigma_0^2} + 1\right\}^{-\frac{\nu+1}{2}} \end{aligned}$$

Note that σ_0 can be selected to ensure that tertiles, which are affected by both the component variance σ_c^2 and the scale parameter σ_0^2 of G_0, are at -1 and 1 as described previously. If we assume a Student's t prior distribution for the location parameter, θ_k, of each component k, the standard approach to generate samples from the posterior distribution of θ_k would be to use:

$$p(\theta_k|\tilde{\mathbf{y}}) \propto p(\tilde{\mathbf{y}}|\theta_k) \int_0^\infty p(\theta_k|\lambda_k)p(\lambda_k)d\lambda_k$$

where λ_k is the prior precision parameter for each component. Let us assume that $\theta_k|\lambda_k, \mu_0, \sigma_0 \sim \mathcal{N}\left(\mu_0, \frac{\sigma_0^2}{\lambda_k}\right)$, resulting in the following prior distribution for C components:

$$\begin{aligned}\pi(\boldsymbol{\theta}, \boldsymbol{\lambda}) &= \pi(\theta_1, \ldots, \theta_C, \lambda_1, \ldots, \lambda_C) \\ &\propto \prod_{k=1}^{C} \mathcal{N}\left(\theta_k; \mu_0, \frac{\sigma_0^2}{\lambda_k}\right) \text{Gamma}\left(\lambda_k; \frac{\nu}{2}, \frac{\nu}{2}\right)\end{aligned}$$

Using the likelihood function (Equation 4.2) we obtain the posterior distribution:

$$p(\boldsymbol{\theta}, \boldsymbol{\lambda}|\tilde{\mathbf{y}}, \mathbf{K}) \propto \prod_{k=1}^{C} \mathcal{N}\left(\theta_k; \mu_0, \frac{\sigma_0^2}{\lambda_k}\right) \text{Gamma}\left(\lambda_k; \frac{\nu}{2}, \frac{\nu}{2}\right) \prod_{j=1}^{J} \prod_{i=1}^{n_j} \sigma_c^{-1} \exp\left[-\frac{(\tilde{y}_{ij} - \theta_{K_{ij}})^2}{2\sigma_c^2}\right]$$

This factorises, so it is easier to focus on each of the C components individually:

$$p(\theta_k, \lambda_k|\tilde{\mathbf{y}}, \mathbf{K}) \propto \mathcal{N}\left(\theta_k; \mu_0, \frac{\sigma_0^2}{\lambda_k}\right) \text{Gamma}\left(\lambda_k; \frac{\nu}{2}, \frac{\nu}{2}\right) \exp\left[-\frac{\sum_{j=1}^{J}\sum_{i=1}^{n_j} \delta_{K_{ij},k}(\tilde{y}_{ij} - \theta_{K_{ij}})^2}{2\sigma_c^2}\right] \tag{4.5}$$

The posterior of θ_k can be obtained by integrating Equation 4.5 over λ_k, but an easier solution is to retain the auxiliary variable λ_k and sample from $p(\lambda_k|\theta_k, \tilde{\mathbf{y}}, \mathbf{K})$ and $p(\theta_k|\lambda_k, \tilde{\mathbf{y}}, \mathbf{K})$ using the following Gibbs sampler:

$$p(\lambda_k|\theta_k, \tilde{\mathbf{y}}, \mathbf{K}) = \text{Gamma}\left(\lambda_k; \frac{\nu+1}{2}, \frac{\nu + \left(\frac{\theta_k - \mu_0}{\sigma_0}\right)^2}{2}\right)$$

$$p(\theta_k|\lambda_k, \tilde{\mathbf{y}}, \mathbf{K}) = \mathcal{N}\left(\theta_k; \frac{\mu_0\lambda_k\sigma_c^2 + \sigma_0^2\sum_{j=1}^{J}\sum_{i=1}^{n_j}\delta_{K_{ij},k}\tilde{y}_{ij}}{m_k\sigma_0^2 + \lambda_k\sigma_c^2}, \frac{\sigma_0^2\sigma_c^2}{m_k\sigma_0^2 + \lambda_k\sigma_c^2}\right)$$

If the number of data assigned to a component is very large, the posterior distribution of θ_k will approach $\mathcal{N}\left(\frac{\sum_{j=1}^{J}\sum_{i=1}^{n_j}\delta_{K_{ij},k}\tilde{y}_{ij}}{m_k}, \frac{\sigma_c^2}{m_k}\right)$ for small ν.

4.2.1.3 Controlling smoothness

Effect of γ

Using a DPMN model requires the selection of a concentration parameter γ. This can be considered as a prior sample size which controls the extent to which samples from a Dirichlet Process reflect the prior distribution G_0. It is therefore important to compare the value of γ to the pooled sample size $\sum_{j=1}^{J} n_j$. As the approach presented in this chapter consists of a DPMN model for each tertile, the model uses a prior sample size of $\gamma/3$ for each tertile.

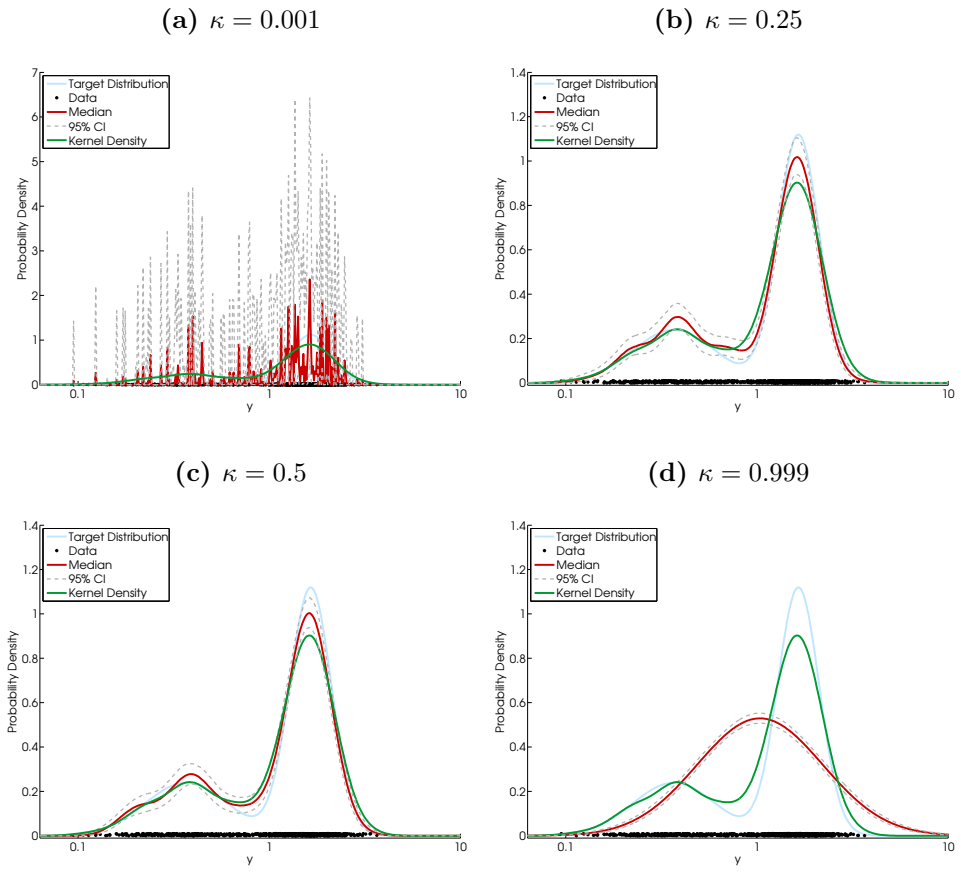
We explained in Section 2.3.2.7 how γ affects the behaviour of the Dirichlet Process: the smaller γ, the more weight will be given to a few components in the Normal mixture. For $\gamma \to 0$, the posterior distribution resembles the data, almost becoming a step function. For $\gamma \approx \sum_{j=1}^{J} n_j$ the posterior distribution is a mixture of the prior distribution G_0 and the population distribution from which the sample was taken and for $\gamma >> \sum_{j=1}^{J} n_j$, the posterior distribution resembles G_0. The reason for this is that for larger γ, more of the components in the Normal mixture distribution will have non-zero weights. In addition, the few data that are assigned to each component k have a minimal effect on the weight w_k as $\gamma + m_k \approx \gamma$.

Effect of κ

The smoothness of the DPMN is also affected by κ, a parameter on the interval $[0, 1]$, which determines the standard deviation of the Normal components in the mixture distribution (see Section 4.2.1.2). If κ is large then all the component locations, θ_k, will tend to be close to μ_0 and the shape distribution will tend to be close to the prior distribution, $\mathcal{N}(\mu_0, \sigma_0^2)$. If κ is small then the components, k, will have a small variance, σ_c^2, and as a result, the shape distribution will not be as smooth. Figure 4.2 shows the results of simulation studies where 100 values were sampled from a mixture of two Normal distributions, $p(y) = p \times \mathcal{N}(y; -1, 0.5^2) + (1-p) \times \mathcal{N}(y; 0.5, 0.25^2)$ with $p = 0.3$ and a DPMN model was used with $G_0(\theta) = \mathcal{N}(\theta; 0, \sigma_0^2)$, $\gamma = 10$ and κ equals 0.001, 0.25, 0.5 and 0.999, respectively. The larger the value of κ, the smoother the distribution is, however, if κ is too large, e.g. $\kappa = 0.999$, the model

will not be able to describe peaks. Therefore the choice of κ will depend on the application.

Figure 4.2 – *Results of simulations using 1000 samples from a Normal mixture distribution* ($p(y) = p \times \mathcal{N}(y; -1, 0.5^2) + (1-p) \times \mathcal{N}(y; 0.5, 0.25^2)$ *with* $p = 0.3$*) with* $\gamma = 10$ *and varying* κ*. The population distribution is displayed as a blue line and a kernel density estimate is represented by a green line. The red line represents the median estimate of the population distribution and the dashed grey lines show the 95% credible interval.*



In many applications, one wants a relatively smooth distribution that can still account for any existing non-smooth areas of the population distribution. In these applications, gaps in the data are considered to be a result of the sampling procedure, for example, because few data were collected or because data were reported as rounded figures. However, in applications where one would only expect certain

values to appear and gaps in the data are likely to be real, smoothing the posterior distribution would result in incorrect inferences. We explore the effect of κ in multiple simulation studies in Section 4.3. It is clear from these studies that κ has a clear effect on the smoothness of the distribution and on the fit. In many applications, we will not have much information on the expected smoothness of the population distribution and there will be no 'true' population distribution that we can use to compare results against. Therefore, the analyst will have to decide what level of smoothness is considered reasonable. The results of the simulation studies in this chapter provide some guidance on the effect of κ on the smoothness of the posterior distribution for our model and which values for κ could be considered appropriate for our application of log-residue data.

Effect of γ and κ together

Now that we have determined that both κ and γ can act as smoothing parameters, we need to assess how they work together. If $\kappa \to 1$, the influence of γ becomes limited as it does not matter whether the shape distribution consists of a mixture of $\mathcal{N}(\mu_0, \sigma_0^2)$ distributions where one distribution has the vast majority of the weight or a mixture of many $\mathcal{N}(\mu_0, \sigma_0^2)$ distributions where the weights are spread more equally across the distributions. Small γ will result in one (or a few) component(s) in the mixture having most of the weight. Large γ will spread the weight over multiple components. Either way, the posterior distribution will barely be influenced by the data. If $\kappa \to 0$, γ will have a larger influence: for small γ the posterior distribution will essentially be an empirical step function at the data values. For large γ many components in the mixture will have non-zero weights which are hardly influenced by the data, resulting in a posterior distribution that is similar to G_0.

4.2.1.4 Computation of the shape distribution

The inference challenge for the shape distribution is to learn the component location parameters, $\boldsymbol{\theta}$, and weights, $\mathbf{w}$, given the normalised log-residue data $\tilde{\mathbf{y}}$. For this purpose, we propose a Markov Chain Monte Carlo algorithm, which consists of the following steps:

1. Allocate each data point $\tilde{y}_{ij}$ to one of the Normal components $\mathcal{N}(\theta_k, \sigma_c^2)$ of the mixture distribution using allocation parameter K_{ij}.

2. Update weights for each Normal component $w_k|m_k$ where $m_k = \sum_{j=1}^{J} \sum_{i=1}^{n_j} \delta_{K_{ij},k}$, i.e. the number of data assigned to component k.

3. Update location parameters $\boldsymbol{\theta}|\tilde{\mathbf{y}}, \mathbf{K}$

These steps are explained in Figure 4.3 and result in a sample from the posterior shape distribution.

We use the truncated stick-breaking representation of the DPMN (Ishwaran and Zarepour, 2000) as it leads to a simple step to allocate data to the Normal components and a straightforward conjugate update of the weights. One issue with the truncated stick-breaking representation is that unassigned weights will be assigned to the last component in the mixture. If this value is high, the approximation of the DPMN will be poor and the truncation level, defined by the number of components C, should be increased. This is discussed in detail in Section 2.3.5.3. Sampling from a posterior DPMN model requires a fine balance between computational efficiency and finding an approximation that meets the required quality criteria. This can be achieved by selecting a large number of components and by monitoring the tail probabilities. To improve the mixing of the Markov chain, we make use of label-swapping moves (see Section 2.3.5.3 for details).

Allocations K_{ij}

Given that we use a truncated approximation of the DPMN model, the allocation of data $\tilde{y}_{ij}$ is given by:

$$p(K_{ij} = k|\tilde{y}_{ij}) = \frac{w_k \mathcal{N}(\tilde{y}_{ij}; \theta_k, \sigma_c^2)}{\sum_{k=1}^{C} w_k \mathcal{N}(\tilde{y}_{ij}; \theta_k, \sigma_c^2)} \tag{4.6}$$

With $u_{ij} \sim \text{Uniform}(0, 1)$, we set $K_{ij} = k$ if and only if

$$\sum_{l=0}^{k-1} w_l \mathcal{N}(\tilde{y}_{ij}; \theta_l, \sigma_c^2) < u_{ij} \leq \sum_{l=1}^{k} w_l \mathcal{N}(\tilde{y}_{ij}; \theta_l, \sigma_c^2)$$

where $w_0 = 0$.

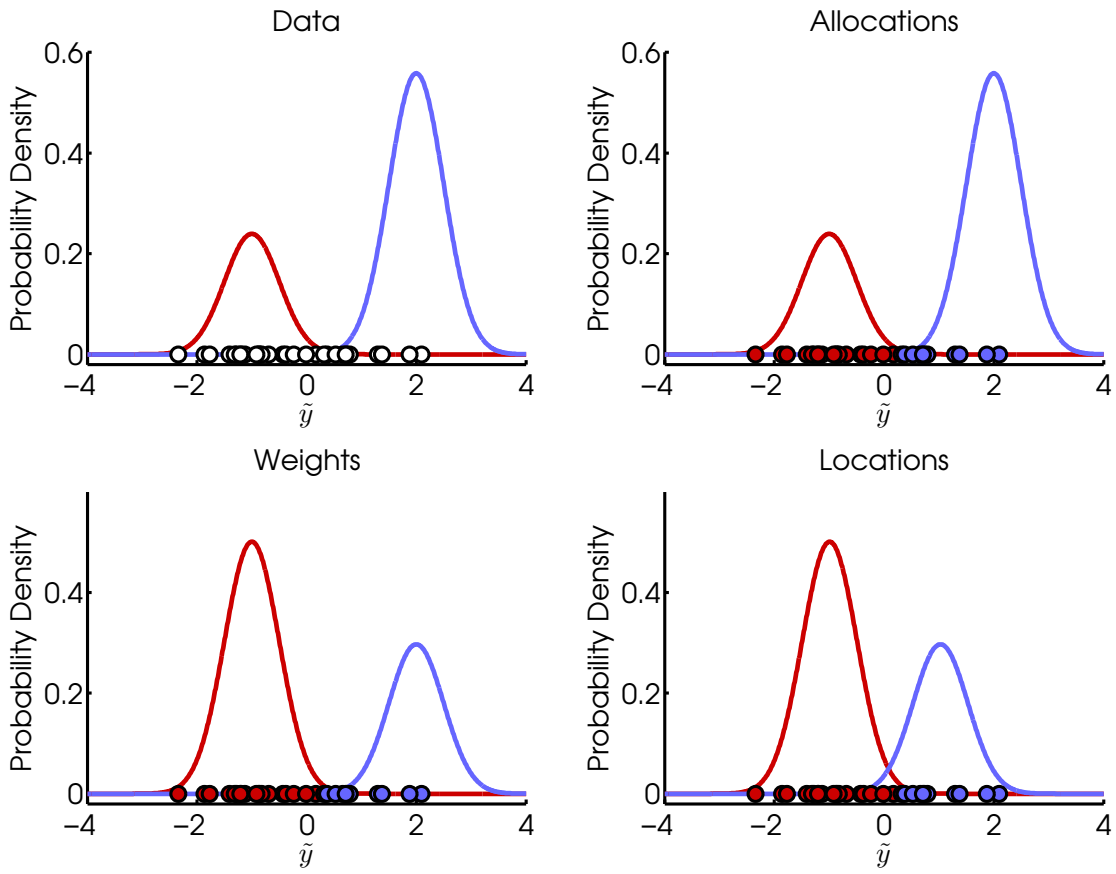


Figure 4.3 – *Graphical overview of the shape model using a mixture distribution with two Normal components. Given initial values of the weights* $\mathbf{w}$ *and locations* $\boldsymbol{\theta}$*, data will be assigned to one of the Normal components (upper right pane) based on the likelihood. Given the allocations, we can update the weights* $\mathbf{w}$ *(bottom left pane) and locations* $\boldsymbol{\theta}$ *(bottom right pane). These three steps will be alternated with updates of the location and scale parameters and the subsequent normalisation step.*

Weights w_k

Given the allocations $\mathbf{K}$ and $m_k = \sum_{j=1}^{J} \sum_{i=1}^{n_j} \delta_{K_{ij},k}$, the posterior distribution of $\mathbf{w}$ is given by:

$$\begin{aligned} w_1 &= \beta_1 \\ w_k &= \beta_k \prod_{l=1}^{k-1}(1-\beta_l) \qquad \text{for } k > 1 \\ \beta_k|\mathbf{K} &\sim \text{Beta}\left(1 + m_k, \gamma + \sum_{l=k+1}^{C} m_l\right) \end{aligned} \tag{4.7}$$

The prior expected tail probability plots in Figure 2.6 (page 90) indicate that a value of $\gamma = 10$ results in a low mean tail probability. Therefore, we use this value in simulation studies in cases where γ is a fixed parameter. In Section 4.6 we allow γ to be learned from the data to see how this effects the shape distribution for the log-residue data. To assess whether the number of components is sufficient, the tail probability can be monitored post-analysis.

Location parameters $\boldsymbol{\theta}$

The selection of a conjugate prior $G_0(\theta) = \mathcal{N}(\theta; \mu_0, \sigma_0^2)$ for θ leads to the posterior distribution, $p(\theta_k|\tilde{\mathbf{y}}, \mathbf{K})$, given in Equation 4.4. We then need to sample the location parameters from truncated posterior distributions to make sure that they are within the ranges of each of the three tertiles. If there are C components in the mixture, the first $C/3$ component location parameters will have to be in the range $[-\infty, -1]$, the second $C/3$ component location parameters in the range $(-1, 1]$ and the third set in $(1, \infty]$.

4.2.2 Estimating the location and scale parameters

Before we can infer the distribution shape, we need to normalise the log-residue data sets from the various populations. As explained before, the model is based on the assumption that we have samples from multiple populations which share a distribution shape but each of which has different location and scale parameters. Normalising the log-residue data will allow us to infer the distribution shape by

sharing information between data sets. The normalisation requires the definition of the location and scale parameters, μ_j and σ_j, of the j^{th} population of unit log-residue data from which we have obtained a sample of size n_j from a unit field trial. Both μ_j and σ_j were defined in Section 4.2.1.1 as the mean of the two tertiles and half the intertertile range, respectively. Since we do not know what μ_j and σ_j are, we need to use a Bayesian framework to learn them from the data. We will update the values of μ_j and σ_j, given the shape distribution, which is defined by the locations, $\boldsymbol{\theta}$, and weights, $\mathbf{w}$. Given $\boldsymbol{\theta}$ and $\mathbf{w}$, the shape distribution is given by:

$$p(\tilde{y}|\boldsymbol{\theta}, \mathbf{w}) \propto \sum_{k=1}^{C} w_k \mathcal{N}\left(\tilde{y}; \theta_k, \sigma_c^2\right)$$

To obtain a new $\boldsymbol{\mu} = \{\mu_1, ..., \mu_J\}$ and $\boldsymbol{\sigma} = \{\sigma_1, ..., \sigma_J\}$ we need to sample from the conditional distribution:

$$\begin{aligned} p(\boldsymbol{\mu}, \boldsymbol{\sigma}|\boldsymbol{\theta}, \mathbf{w}, \mathbf{y}) &\propto \pi(\boldsymbol{\mu}, \boldsymbol{\sigma}) p(\mathbf{y}|\boldsymbol{\theta}, \mathbf{w}, \boldsymbol{\mu}, \boldsymbol{\sigma}) \\ &\propto \pi(\boldsymbol{\mu}, \boldsymbol{\sigma}) \prod_{j=1}^{J} \prod_{i=1}^{n_j} \frac{1}{\sigma_j} p(\tilde{y}_{ij}|\boldsymbol{\theta}, \mathbf{w}) \end{aligned}$$

where $\pi(\boldsymbol{\mu}, \boldsymbol{\sigma})$ is the joint prior distribution for $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$. We propose to use independent Jeffreys priors, $\pi(\boldsymbol{\mu}, \boldsymbol{\sigma}) \propto \prod_{j=1}^{J} \sigma_j^{-1}$, given the reasonably large sample sizes. This results in the following posterior distribution:

$$p(\mu_j, \sigma_j|\boldsymbol{\theta}, \mathbf{w}, \mathbf{y}) \propto \sigma_j^{-n_j-1} \prod_{i=1}^{n_j} p(\tilde{y}_{ij}|\boldsymbol{\theta}, \mathbf{w}) \tag{4.8}$$

As we cannot easily sample from Equation 4.8, we will use a Metropolis-Hastings step within Gibbs.

Let $\mu_j^{(t)}$ and $\sigma_j^{(t)}$ be the values for the location and scale parameters of population j at the t^{th} iteration of the Markov Chain Monte Carlo simulation. In a Metropolis-Hastings step, we propose new values μ_j^* and σ_j^* which we will either accept or reject using the Metropolis-Hastings rule applied to Equation 4.8. Firstly, we need to define proposal distributions for μ_j and σ_j. If we look at σ_j first, we have the following characteristics of the distribution of the sample variance s^2 for moderate to large n_j: $\mathbb{E}[s^2] = \sigma^2$ and $\text{Var}[s^2] = \frac{2\sigma^4}{n-1}$. From this we can derive that the expected value

and variance of the random variable $\frac{s^2}{\sigma^2}$ is given by:

$$\mathbb{E}\left[\frac{s^2}{\sigma^2}\right] = 1$$
$$\text{Var}\left[\frac{s^2}{\sigma^2}\right] = \frac{1}{\sigma^4}\text{Var}\left[s^2\right] = \frac{2}{n-1}$$

Using the central limit theorem, we can approximate the distribution of $\frac{s^2}{\sigma^2}$ with a Normal distribution.

$$\frac{s^2}{\sigma^2} \sim 1 + z\sqrt{\frac{2}{n-1}}$$

where $z \sim \mathcal{N}(0, 1^2)$. Taking logs and using the Taylor series for $\log(1+q)$ for small values of q, i.e. $\log(1+q) \approx q$, leads to:

$$2\left(\log(s) - \log(\sigma)\right) \approx z\sqrt{\frac{2}{n-1}}$$

for large n. This results in:

$$\log s \sim \mathcal{N}\left(\log\sigma, \frac{1}{2(n-1)}\right)$$

Using a proposal distribution for $\log(\sigma)$ that makes steps of size proportional to $\frac{1}{\sqrt{n-1}}$ results in reasonable acceptance rates, so we suggest the following proposal distribution for $\log(\sigma)$:

$$q(\log\sigma_j^* | \log\sigma_j^{(t)}, \mathbf{y}_j) = \mathcal{N}\left(\log\sigma_j^*; \log\sigma_j^{(t)}, \frac{1}{n_j - 1}\right)$$

If we focus on μ_j, using a standard random walk could be considered as the proposal distribution in a Metropolis-Hastings step:

$$q(\mu^* | \mu^{(t)}, \sigma^*) = \mathcal{N}\left(\mu^*; \mu^{(t)}, \frac{(\sigma^*)^2}{n}\right)$$

The disadvantage of this is that if σ^* is much smaller than $\sigma^{(t)}$, we will frequently find that $\mu^{(t)}$ is far into the tail of $p(\mu^{(t)}; \mu^*, \sigma^*)$ and the random walk will result in many rejections. To overcome this, we could limit the proposal step size for σ_j, but this would also lead to a need for more thinning. Therefore we instead propose:

$$q(\mu_j^* | \mu_j^{(t)}, \sigma_j^{(t)}, \sigma_j^*, \mathbf{y}_j) = \mathcal{N}\left(\mu_j^*; \bar{y}_j + \frac{\sigma_j^*(\mu_j^{(t)} - \bar{y}_j)}{\sigma_j^{(t)}}, \frac{(\sigma_j^*)^2}{n_j}\right)$$

Fundamentally, we expect the variance of the location parameter given the scale parameter to be roughly proportional to the scale parameter divided by the square

root of the sample size. We also expect the 'relative uncertainty' about the scale parameter to be related to $1/\sqrt{n_j}$. This proposal distribution firstly normalises the current location with respect to the sample location and the current scale and then returns to the original scale using $\bar{y}_j$ and the proposed scale. As a result, our proposal distribution aims to keep the proposed location at the same percentile of the conditional distribution of the location parameter for both the proposed and current scale values. We then propose the location parameter using a random step based on the proposed scale parameter divided by the square root of the sample size.

So let the proposal distributions for μ_j^* and σ_j^* be:

$$\begin{aligned} q(\log \sigma_j^* | \log \sigma_j^{(t)}, \mathbf{y}_j) &= \mathcal{N}\left(\log \sigma_j^*; \log \sigma_j^{(t)}, \frac{1}{n_j - 1}\right) \\ q(\mu_j^* | \mu_j^{(t)}, \sigma_j^{(t)}, \sigma_j^*, \mathbf{y}_j) &= \mathcal{N}\left(\mu_j^*; \bar{y}_j + \frac{\sigma_j^*(\mu_j^{(t)} - \bar{y}_j)}{\sigma_j^{(t)}}, \frac{(\sigma_j^*)^2}{n_j}\right) \end{aligned} \tag{4.9}$$

where $\bar{y}_j = \frac{Q_j^{\frac{2}{3}} + Q_j^{\frac{1}{3}}}{2}$. Let $u_j \sim \text{Uniform}(0, 1)$. We accept the proposed values if:

$$u_j \leq \frac{p(\mu_j^*, \sigma_j^* | \boldsymbol{\theta}, \mathbf{w}, \mathbf{y})}{p(\mu_j^{(t)}, \sigma_j^{(t)} | \boldsymbol{\theta}, \mathbf{w}, \mathbf{y})} \frac{q\left(\mu_j^{(t)} | \mu_j^*, \sigma_j^*, \sigma_j^{(t)}, \mathbf{y}_j\right)}{q\left(\mu_j^* | \mu_j^{(t)}, \sigma_j^{(t)}, \sigma_j^*, \mathbf{y}_j\right)} \frac{q\left(\sigma_j^{(t)} | \sigma_j^*, \mathbf{y}_j\right)}{q\left(\sigma_j^* | \sigma_j^{(t)}, \mathbf{y}_j\right)}$$

The first fraction, the target ratio, can be calculated directly using Equation 4.8. The second fraction, the proposal ratio for μ_j, will lead to:

$$\frac{q(\mu_j^{(t)} | \mu_j^*, \sigma_j^*, \sigma_j^{(t)}, \mathbf{y}_j)}{q(\mu_j^* | \mu_j^{(t)}, \sigma_j^{(t)}, \sigma_j^*, \mathbf{y}_j)} = \frac{\frac{\sqrt{n_j}}{\sigma_j^{(t)}\sqrt{2\pi}}}{\frac{\sqrt{n_j}}{\sigma_j^*\sqrt{2\pi}}} = \frac{\sigma_j^*}{\sigma_j^{(t)}}$$

and the proposal ratio for σ_j leads to:

$$\frac{q(\sigma_j^{(t)} | \sigma_j^*, \mathbf{y}_j)}{q(\sigma_j^* | \sigma_j^{(t)}, \mathbf{y}_j)} = \frac{\sigma_j^*}{\sigma_j^{(t)}}$$

As a result, the acceptance ratio becomes:

$$u_j \leq \left(\frac{\sigma_j^*}{\sigma_j^{(t)}}\right)^{-(n_j - 1)} \frac{\prod_{i=1}^{n_j} p(\tilde{y}_{ij}^* | \boldsymbol{\theta}, \mathbf{w})}{\prod_{i=1}^{n_j} p(\tilde{y}_{ij}^{(t)} | \boldsymbol{\theta}, \mathbf{w})} \tag{4.10}$$

where $\tilde{y}_{ij}^*$ is the normalised log residue data given the proposed values mu_j^* and σ_j^*. We have now discussed the technical aspects of the model. To apply the model to pesticide log-residue data two additional model refinements are necessary which will be discussed in the next section.

4.2.3 Handling censored and rounded data

The model introduced above describes an approach that can be used to learn the location, scale and shape distribution for a data set consisting of samples from multiple populations which share a common shape but each of which has their own location and scale parameters. This section will discuss two application-specific issues that are important to address before the model can be applied to unit log-residue data sets.

4.2.3.1 Censoring

Residue levels in food items are often lower than concentrations that can be measured reliably, i.e. the observed response cannot be distinguished from the response observed when analysing a blank sample (see Section 1.4.1.1). For field trial data we know that the field was treated with the pesticide under consideration and therefore, if we ignore measurement uncertainty as suggested by EFSA (2012), the residue level will be somewhere between zero and the reported limit of determination (LOD). Therefore, if we have observed a data set x with values reported as <LOD, we can use a simple data augmentation procedure to account for the limited amount of information provided by the <LOD values. Given a distribution form $f(x;\omega)$ with cumulative distribution function $F(x;\omega)$ and given a prior distribution and initial values for the parameter(s) ω, repeat the following steps:

1. For each of the $q = 1, \ldots, Q$ values that are reported as <LOD, sample a new value using the following steps:

   (a) Calculate $u_{\max} = F(LOD|\omega)$

   (b) $u \sim \text{Uniform}(0, u_{\max})$

   (c) $x_q = F^{-1}(u|\omega)$

2. Update ω given the observations $x_{x>LOD}$ and imputed values $x_{x<LOD}$

We could use a similar approach for dealing with <LOD data in the DPMN model. However, calculation of $u_{\max}$ and in particular $F^{-1}(u|\omega)$ is not very efficient so we

instead propose an approach which has the additional benefit that it will implicitly lead to allocating the censored value to one of the components of the mixture distribution. Let y_{ij} be a censored log-residue value, <log(LOD), originating from data set j. We impute values using the following algorithm:

Algorithm 1

1. *Normalise the* $\log(LOD)$*:* $\tilde{y}_{ij} = \frac{\log(LOD)-\mu_j}{\sigma_j}$
2. *Calculate* $p_k = w_k \Phi\left(\frac{\tilde{y}_{ij}-\theta_k}{\sigma_c}\right)$ *where* Φ *is the standard Normal cumulative distribution function.*
3. *Sample allocation* $K_{ij} = Multinomial\left(\frac{p_k}{\sum_{k=1}^{C} p_k}\right)$
4. *Given* K_{ij}*, it is easy to calculate* $r_{\max} = \Phi\left(\frac{\log(LOD)-\theta_{K_{ij}}}{\sigma_c}\right)$
5. $r \sim Uniform(0, r_{\max})$
6. $\tilde{y}_{ij} = \theta_{K_{ij}} + \Phi^{-1}(r)\,\sigma_c$

We can then update the weights and parameters of the Normal components of the shape distribution and the location and scale parameters for each log-residue data set.

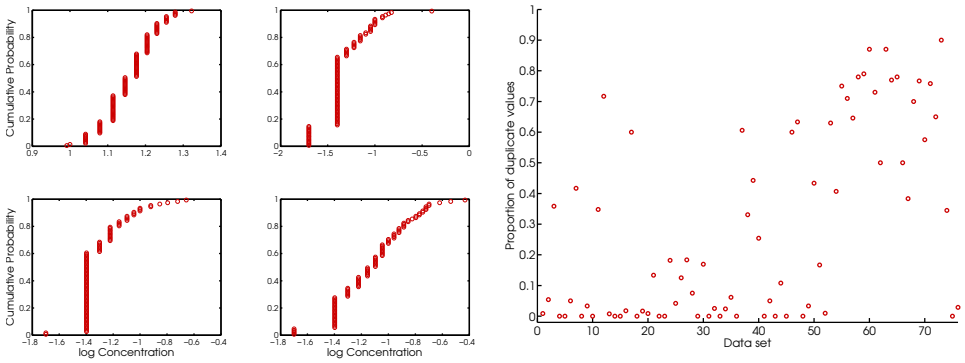
4.2.3.2 Uncertainty in reported values

The second issue with residue data is that they are often reported after rounding to n_d decimal places or n_s significant figures. As a result, many of the values in a data set are repeated, which could suggest that the population distribution is discrete. Figure 4.4a shows cumulative empirical distribution functions of the four field trial data sets with the highest proportion of repeated values. Figure 4.4b shows that repeated values occur frequently in field trial residue data. As we expect residue level distributions to be continuous, we add some random noise around each reported value. The approach to do this depends on which rounding rules were used when reporting the values.

Figure 4.4 *– Level of reporting uncertainty in residue data sets.*

(a) Cumulative empirical distribution functions of the four data sets with the highest proportions of repeated values.

(b) Proportion of repeated values for each field trial data set, defined as $\frac{n_j - n_j^u}{n_j}$, where n_j^u is the number of unique values in field j.



Two common rounding rules are:

1. **Decimal places**: If the data were rounded to n decimal places, we know that the value before rounding was in the interval: $[x_{reported} - \delta, x_{reported} + \delta)$, where $\delta = 1/2 \times 10^{-n}$. For example, if $n = 2$, 18.90 could be the result from rounding any observation in the range $[18.895, 18.905)$ and a value of 0.02 could be the result from any observation in the range $[0.015, 0.025)$.

2. **Significant Figures**: If n significant figures were used, we know that the value before rounding was in the interval: $[x_{reported} - \delta, x_{reported} + \delta)$, where $\delta = 1/2 \times 10^{\lfloor \log_{10} |x| \rfloor - n + 1}$, $\lfloor q \rfloor$ is the largest integer not greater than q and $|x|$ is the absolute value of x. For example, if $n = 2$, a value of 0.17 could be the result from rounding any observation in the range $[0.165, 0.175)$ and a value of 15 could be the result from any observation in the range $[14.5, 15.5)$.

The following algorithm provides an approach for dealing with reporting uncertainty in each iteration of the MCMC simulation. First decide which rounding rule (e.g. number of decimal places or significant figures) was used for the data set x. Then given a distribution form $f(x; \omega)$, with cumulative distribution function $F(x; \omega)$ and given a prior distribution and initial values for the parameter(s) ω, repeat the following steps:

1. For each value $x_{reported}$ that was subject to rounding, sample a new value:

   (a) Calculate δ using the following equations and the selected rounding method:

      i. **Decimal Places**: $\delta = 1/2 \times 10^{-n}$

      ii. **Significant Figures**: $\delta = 1/2 \times 10^{\lfloor \log_{10} |x| \rfloor - n + 1}$

   (b) Calculate $r_{\min} = F(x_{reported} - \delta|\omega)$ and $r_{\max} = F(x_{reported} + \delta|\omega)$

   (c) Sample $r = \text{Uniform}(r_{\min}, r_{\max})$

   (d) $x = F^{-1}(r|\omega)$

2. Update ω given the observations x

Analogously to the censored data approach, several adjustments are necessary to model rounded unit residue data with the DPMN model. The first reason for this is that the analysis is conducted on log-residue data, whereas the data are rounded before they are log-transformed. Secondly, as in the censored data approach, it is easier to sample from a Normal component after the data are allocated to components than to sample from F directly. For the log-residue data, the algorithm is given by:

Algorithm 2

1. *Allocate the normalised log-residue data to one of the Normal components as described before. If $\tilde{y}_{ij}$ is assigned to component k, $K_{ij} = k$.*

2. *For each data set, we determine whether the reported residue levels are rounded using one of the two approaches (decimal places or significant figures). With y_{ij} being observation i from field j of the log-transformed residue data, calculate δ:*

   (a) **Decimal Places***: $\delta = 1/2 \times 10^{-n}$*

   (b) **Significant Figures***: $\delta = 1/2 \times 10^{\lfloor \log_{10} |\exp[y_{ij}]| \rfloor - n + 1}$*

3. *Calculate $L_{\min} = \frac{\log(\exp(y_{ij}) - \delta) - \mu_j}{\sigma_j}$ and $L_{\max} = \frac{\log(\exp(y_{ij}) + \delta) - \mu_j}{\sigma_j}$*

(a) *The situation for residue levels is slightly more complex because we are analysing pesticide residue levels after log-transforming the data. As residue levels are bounded below at zero, we have to put in an additional condition:* $\exp(y_{ij}) - \delta > 0$.

4. *Calculate* $u_{\min} = \mathcal{N}(L_{\min}|\theta_{K_{ij}}, \sigma_c^2)$ *and* $u_{\max} = \mathcal{N}(L_{\max}|\theta_{K_{ij}}, \sigma_c^2)$

5. *Sample* $u = Uniform(u_{\min}, u_{\max})$

6. $\tilde{y}_{ij} = \Phi^{-1}(u|\theta_{K_{ij}}, \sigma_c^2)$

7. *Update* $\boldsymbol{\theta}$, $\mathbf{w}$, $\boldsymbol{\mu}$ *and* $\boldsymbol{\sigma}$ *given* $\tilde{\mathbf{y}}$

4.2.3.3 Summary of MCMC Algorithm

Let us define:

j	Data set index, $j \in \{1, \ldots, J\}$.
n_j	Total number of observations in data set j.
i	Unit index for an observation from a data set j, $i \in \{1, \ldots, n_j\}$.
y_{ij}	Log-residue data observation i from data set j.
$\tilde{y}_{ij}$	Normalised log-residue data observation i from data set j.
γ	Concentration parameter of the DP.
G_0	The base measure of the DP, which will act as the prior distribution for θ, here defined as $\mathcal{N}(\mu_0, \sigma_0^2)$.
μ_0	Mean of Normal prior distribution G_0.
σ_0	Standard deviation of Normal prior distribution G_0.
κ	Proportion of observed variance that is assigned to the Normal components.
n_{it}	Number of MCMC iterations.
$n_{burn\text{-}in}$	Number of burn-in iterations.
n_{thin}	Thinning factor.
μ_j	Location parameter for log-residue data set j.
σ_j	Scale parameter for log-residue data set j.
w_k	Weight assigned to Normal component k in the mixture distribution.
θ_k	Location parameter of Normal component k in the mixture distribution.
σ_c	Standard deviation of all Normal components in the mixture distribution.
K_{ij}	Allocation indicator for normalised log-residue observation i from data set j.
m_k	Number of data allocated to a component k.

The algorithm for the DPMN model for log-residue data which can be used to generate samples from the posterior distribution $p(\boldsymbol{\theta}, \mathbf{w}, \mathbf{K}, \boldsymbol{\mu}, \boldsymbol{\sigma}|\mathbf{y}, G_0, \gamma, \sigma_c^2)$ can be summarised as follows:

1. Select γ, G_0, κ, μ_0, n_{it}, $n_{burn\text{-}in}$ and n_{thin}.
2. Calculate the component standard deviations σ_c and the scale parameter of the prior σ_0.
    (a) Alternatively, γ can be considered to be a model parameter that needs to be learned from the data. In that case, distribution parameters v_1 and v_2 (see Equation 4.11 on page 178) will have to be defined for the prior Gamma(ν_1, ν_2) distribution of γ and an initial value needs to be assigned to γ.
3. Set initial values for $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ by calculating the tertiles of data sets.
4. Set initial values for the component locations $\boldsymbol{\theta}$ and the component weights $\mathbf{w}$.
5. Repeat the following steps for $(n_{it} + n_{burn\text{-}in}) \times n_{thin}$ iterations:
    (a) Normalise data (Equation 4.1 on page 138).
    (b) Update data allocations K_{ij} (Equation 4.6 on page 146).
    (c) Calculate m_k (Equation 4.3 on page 141).
    (d) Account for censored data (Algorithm 1 on page 153) if necessary.
    (e) Account for rounding error (Algorithm 2 on page 155) if necessary.
    (f) Update distribution shape:
        i. Swap component labels (see Section 2.3.5.3).
        ii. Update weights (Equation 4.7 on page 148).
        iii. If γ is considered to be a model parameter (see Section 4.6), update γ (Equation 4.11 on page 178).
        iv. Update locations (Equation 4.4 on page 141).

(g) Update location and scale for each data set using a Metropolis-Hastings step:

i. Propose μ_j and σ_j (Equation 4.9 on page 151).

ii. Calculate acceptance probability (Equation 4.10 on page 151) and accept or reject proposed values.

(h) Store values of $\boldsymbol{\theta}$, $\mathbf{w}$, $\boldsymbol{\mu}$, $\boldsymbol{\sigma}$ and (if appropriate) γ if $t - n_{burn\text{-}in} > 0$ and the remainder after division, $\text{rem}(t - n_{burn\text{-}in}, n_{thin}) = 0$, where t is the iteration index.

4.3 Validation Studies

Before applying the non-parametric Bayesian method to a case study, we want to assess whether the method is capable of recovering a common shape distribution. To assess the performance of our DPMN approach, we cannot apply the data to real residue data as we do not know what the true underlying distribution is and therefore we will not be able to determine whether the resulting shape distribution provides a good estimate of the population shape distribution. Therefore, one way to test the method is to compare it with data generated from a distribution or set of distributions for which we know the shape. In this section, we present the results of several simulation studies to assess whether the new approach is capable of determining the shape of a selection of distributions. The validation simulations focus on two aspects:

- Determination of the distribution shape using samples obtained from various populations with a shared shape.
- Determination of the distribution shape for large sample sizes ($n = 1000$) for a wide range of distributions.

For all validation studies, we ran the model with 1000 iterations after 1000 burn-in samples using a thinning factor of 25 and fixed the number of components in the mixture to 201. Each study was run in Matlab 2012a on a computer with an Intel i7-860 2.80 Ghz processor and 8GB RAM and took approximately 70 minutes. The first

type of study is based on an overall sample size of $n = 1000$, but instead of taking 1000 samples from a single distribution, $\frac{1000}{p}$ samples are taken from p populations, each with different location and scale parameters. We begin by explaining the general approach for the validation studies for a sample generated from a mixture of two Normal distributions. To test the performance of the approach, we generated $n = 1000$ samples from the distribution:

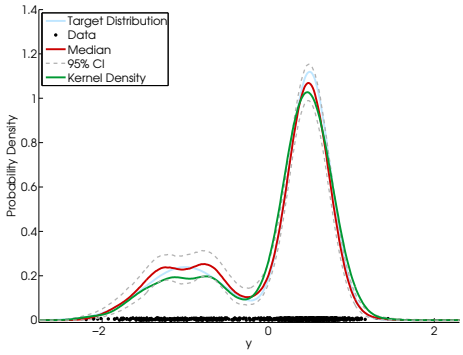
$$p(y) = p \times \mathcal{N}\left(y; -1, 0.5^2\right) + (1 - p) \times \mathcal{N}\left(y; 0.5, 0.25^2\right)$$

where $p = 0.3$. The aim of the simulation study is to estimate the shape of the distribution using these samples. Figure 4.5a shows the results from applying our DPMN model to these data, with $\gamma = 10$ and $\kappa = 0.3$. Clearly, the estimated shape is very close to the true shape of the distribution, i.e. the DPMN model is capable of estimating the shape of the distribution, indicating that DPMN models can be used to fit distributions to data sets. One of the reasons why the DPMN model was capable of estimating the shape may be the large sample size that was used in this test. For many applications, the number of samples taken from a population will be much lower. However, the application that we will be working on will have samples from multiple populations. The challenge will be to use these multiple samples to estimate both the common shape and the location and scale parameters of each population. For that purpose, we sampled $n = 100$ values from 10 populations with a common shape, each with their own location and scale parameters. Applying our DPMN model to these data sets to learn a common shape resulted in the shape distribution in Figure 4.5b. The results indicate that even when the 1000 samples come from 10 populations with different location and scale parameters and we have to learn both the common shape and the population-specific location and scale parameters, the DPMN model is able to describe the shape, location and scale parameters.

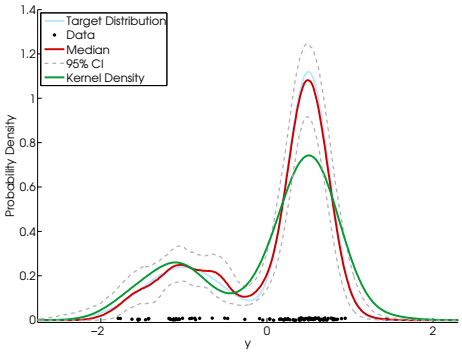
To assess the performance of the method for smaller sample sizes, we took 10 samples from 100 distributions, again with different location and scale parameters but a common shape. Figure 4.5c shows that even for such small sample sizes, the performance of the DPMN model is acceptable, although the uncertainty in the shape distribution and location and scale parameters, indicated by a 95% credible interval,

Figure 4.5 – *Results of validation exercise for a mixture of two Normal distributions,* $p(y) = p \times \mathcal{N}(y; -1, 0.5^2) + (1-p) \times \mathcal{N}(y; 0.5, 0.25^2)$*, with* $p = 0.3$*,* $\kappa = 0.3$ *and* $\gamma = 10$*. The population distribution is displayed as a blue line and a kernel density estimate is represented by a green line. The red line represents the median estimate of the population distribution and the dashed grey lines show the 95% credible interval. For b - d, we only illustrate the results for one of the populations.*

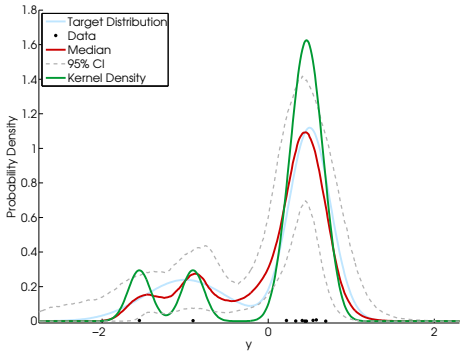
(a) Results of simulation using 1000 samples from a Normal mixture distribution.



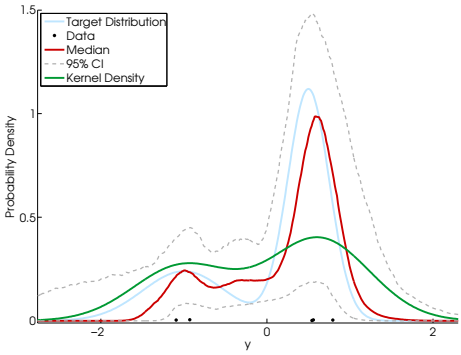
(b) Results of DPMN simulation using 100 samples from 10 populations.



(c) Results of DPMN simulation using 10 samples from 100 populations.



(d) Results of DPMN simulation using 5 samples from 200 populations.



becomes more pronounced. This is even more evident in the final study (Figure 4.5d) in which 5 values were sampled from 200 population distributions. These analyses show that if populations do share a common shape, even a few samples from each population should be sufficient to learn the shape of the distribution. However, the problem is that when only a few samples are available per population there is very little information available to estimate the population location and scale parameters. As a result, even though pooling the data in the DPMN model may lead to a good estimation of the shape distribution, the location and scale parameter estimates are very uncertain. Even if the distribution shape was known, small sample sizes would often result in poor estimates of the location and scale parameters and in those cases, a hierarchical model may have to be considered. However our method is an improvement on current approaches which make assumptions about the distribution shape and are likely to result in considerable parameter uncertainty as a result of analysing each data set individually.

In the following sections we will summarise the results for other target distributions, highlighting strengths and weaknesses of the DPMN approach. We start by exploring how well the approach applies to different distribution shapes using various values of κ. In addition, we will explore the effect of sample size.

4.3.1 Performance for various distributions

In this section we will discuss the results of simulation studies for a range of distributions to assess the robustness of our approach for heavy tailed and/or highly skewed distributions. For all simulation studies γ was set to be 10 and 1000 samples were taken from a single population distribution, i.e. the target distribution, unless indicated otherwise. We ran the model with 4 different values for κ: 0.1, 0.2, 0.4 and 0.8. The posterior probability density functions for each simulation study are supplied in Appendix B.

4.3.1.1 Normal Distribution

The first distribution for which we assess the performance of the DPMN model is the Normal distribution. For larger values of κ the DPMN model provides a reasonable fit (Figures B.1c and B.1d). The posterior probability density function for lower values of κ is clearly less smooth and shows the importance of selecting an appropriate value for κ.

4.3.1.2 Student's t Distribution

The second distribution function that is used in the simulation studies is Student's t distribution with density function $p(x|\nu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{x^2}{\nu}\right)^{-(\nu+1)/2}$ where ν is the number of degrees of freedom ($\nu > 0$). We sample data from Student's t distributions with $\nu = 3$, 4 and 5. Student's t distributions have relatively long tails, particularly for small values of ν. The simulation studies (Figures B.2 to B.4) indicate that the DPMN model struggles to 'learn' the Student's t shape from the samples. The reason for this is that to describe the central peak, κ needs to be small, but to describe the long tails κ needs to be large. The results indicate that the method needs to be refined to describe data with long tails or the data may need to be transformed before applying the model.

4.3.1.3 Skew-Normal Distribution

The next family of distributions considered is the Skew-Normal distribution with density $p(x|\lambda) = 2\phi(x)\Phi(\lambda x)$ where $\phi(x)$ is the standard Normal probability density function and $\Phi(x)$ is the standard Normal cumulative distribution function. Note that for $\lambda = 0$, the Skew-Normal distribution is the standard Normal distribution function. The simulations (Figures B.5 to B.14) show that for various values of λ, the DPMN is capable of describing the Skew-Normal distribution. Again, the quality of the description depends to some extent on the value of κ: for small values of κ the distribution is too jagged and for large values of κ, the DPMN overestimates the upper tail for left-skewed distributions and the lower tail for right-skewed ones.

4.3.1.4 Exponential Power Distribution

The Exponential Power distribution, also known as the generalised Normal distribution is defined as $p(x|\mu,\sigma,\lambda) = \frac{\lambda}{2\sigma\Gamma(1/\lambda)} \exp\left[-\left|\frac{(x-\mu)^2}{\sigma}\right|^{\lambda}\right]$. For $\lambda = 2$, the Exponential Power distribution becomes the Normal distribution. This symmetric family of distributions allow for heavier tails than the Normal distribution for $\lambda < 2$ and for lighter tails than the Normal distribution for $\lambda > 2$. Note that if $\lambda \to \infty$, the distribution resembles a symmetric Uniform distribution, centred at μ. In the simulation studies presented in Figures B.15 to B.19, we set $\mu = 0$ and $\sigma = 1$. When fitting a DPMN model to a sample obtained from the Exponential Power distribution, it is clear that for $\lambda < 2$, as with the simulations for the Student's t distribution, the DPMN model struggles to capture the central peak and the heavy tails. Small values for κ are needed to capture the narrow peak but this generally leads to jagged distributions. For $\lambda > 2$, the performance of the DPMN model seems better, particularly for larger values of κ. However, the clear plateau for the $\lambda = 5$ simulation is not matched.

4.3.1.5 Beta Distribution

The Beta distribution is a family of continuous distributions, generally defined by 2 shape parameters α and β, on the interval (0,1). The two parameter probability density function is $p(x|\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$. For $\alpha = \beta = 1$ the Beta distribution is the Uniform(0,1) distribution and for $\alpha < 1$ and $\beta < 1$ it is U-shaped. The Beta distribution can be extended to the interval (p,q) by introducing two additional variables p and q: $p(x|\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)(q-p)^{\alpha+\beta-1}}(x-p)^{\alpha-1}(q-x)^{\beta-1}$. In the simulations, we used $p = -2$ and $q = 2$. It is clear from the simulations (Figures B.20 to B.22) that for the selected range, the DPMN model worked best in terms of smoothness for large values of κ. However, large values of κ also meant that the modelled tails extended beyond the limits of the Beta distribution.

4.3.1.6 Normal Mixture Distribution

The final family of distributions that we use in the simulation studies is a mixture distribution of two Normal distributions: $p(y|p, \mu_1, \sigma_1, \mu_2, \sigma_2) = p \times \mathcal{N}(y; \mu_1, \sigma_1^2) + (1-p) \times \mathcal{N}(y; \mu_2, \sigma_2^2)$. Here, we used a symmetrical setup with $\mu_1 = -\mu_2 = 2$ and $\sigma_1 = \sigma_2 = 1$ and we varied p. The simulation studies (Figures B.23 to B.27) show that the DPMN model performs well depending on the value of κ. For $p = 0.5$ the DPMN model works best for $\kappa = 0.2$. Smaller values for κ are not very smooth and larger values for κ lead to poorer fits as the DPMN starts to struggle to pick up the bimodality. In contrast, for the $p = 0.1$ and $p = 0.9$ case the DPMN model fits better when larger values for κ are used whereas the $p = 0.75$ simulation works best with $\kappa = 0.4$. The results indicate again the importance of selecting an appropriate value for κ.

4.3.2 Effect of Sample Size

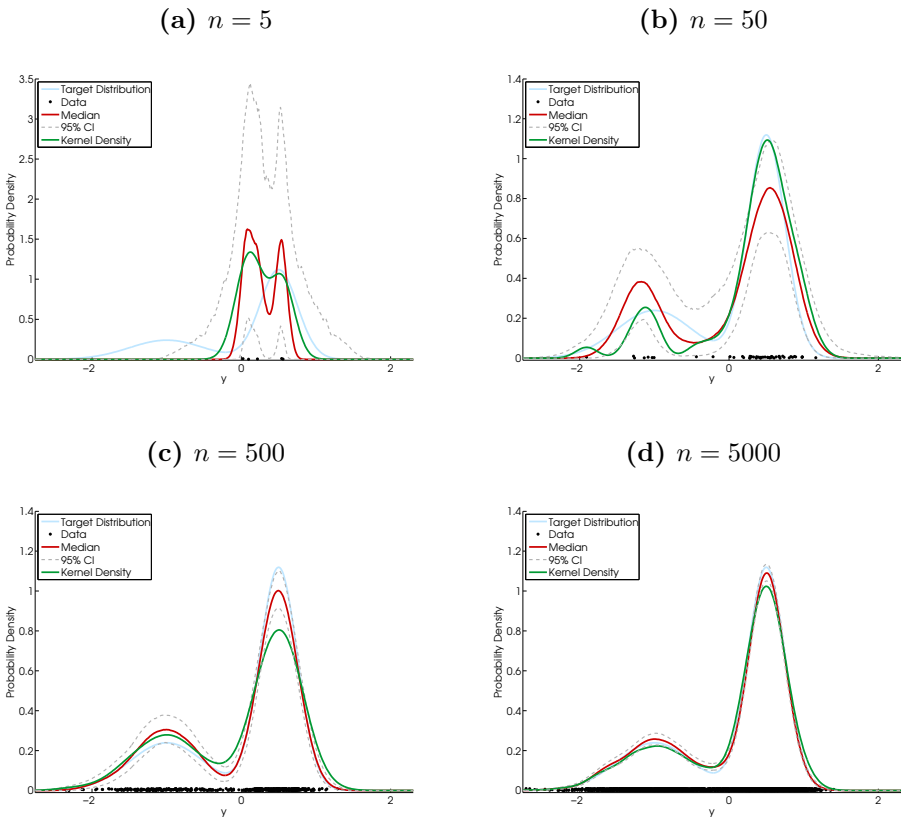
To assess how the DPMN results depend on the overall sample size, we limit our attention to the Normal mixture distribution we used earlier, $p(y) = p \times \mathcal{N}(y; -1, 0.5^2) + (1-p) \times \mathcal{N}(y; 0.5, 0.25^2)$ with $p = 0.3$. Figure 4.6 shows that the DPMN model is capable of describing the shape of the distribution reasonably well for a sample size as small as 50. For $n = 5$, there is only a slight indication of bimodality from the data, whereas it is clearly visible for $n = 50$. For both these simulations, it is clear that uncertainty about the location parameter and the scale parameter is considerable. For larger n, the model is able to learn the distribution shape from the data.

4.3.3 Results of Simulation Studies

The simulation studies above indicate that the DPMN model is capable of determining the shape of a distribution when we have a large sample size, either from a single population or from multiple populations which share the same shape. However, two issues have been identified that need to be addressed. Firstly, the model struggled to fit distributions with long tails such as the Student's t distribution and

the Exponential Power distribution with $\lambda \leq 2$. The problem is caused by the scale of the Normal components which need to be small for components with location parameters close to the mode to capture the central peak and large for components in the tails to capture the longer tails. As a result, choosing small values of κ will provide a better description of the central part of these distributions whereas large values of κ will provide a better description of the tails. Therefore, changing κ cannot provide a solution. However, one could try learning κ from the data, to obtain a compromise between the best fit for the central part of the distribution and the

Figure 4.6 – *Results of simulations determining the effect of sample size on the performance of the DPMN model for a Normal Mixture Distribution,* $p(y) = p \times \mathcal{N}(y; -1, 0.5^2) + (1-p) \times \mathcal{N}(y; 0.5, 0.25^2)$, *with* $p = 0.3$, $\kappa = 0.3$ *and* $\gamma = 1$. *The population distribution is displayed as a blue line and a kernel density estimate is represented by a green line. The red line represents the median estimate of the population distribution and the dashed grey lines show the 95% credible interval.*



best fit for the tails.

If we believe that the population distribution has long tails, we have several options available. The first option is to select a different distribution shape for the components in the mixture (e.g. Student's t distribution). The second is to allow the component variance, σ_k^2, to be variable and uncertain. By learning the variance from the data assigned to each component, we can use Normal distributions with large values for σ_k^2 for components that aim to describe the tails and Normal distributions with small values for σ_k^2 for components that are used to describe the central part of the distribution if this is necessary. These solutions would also work for population distributions whose tails are bounded like the Beta distribution. The main problem with this solution is that it will cause problems when defining the prior shape distribution.

Another option may be to transform the data before we infer the distribution shape. A simple example would be to log-transform data sampled from populations which are expected to have a long upper tail (e.g. pesticide residue data). More complex transformations may also be possible but they will affect the ease of interpretation of the model output. For example, applying the model to log-transformed data means that the location parameter for the log-residue data matches the scale parameter for residue data and that the scale and shape parameters for log-residues become aspects of the shape of the residue distribution.

The second issue is that the model requires the analyst to select κ. In the simulation studies, it is often clear which of the 4 selected values of κ is most appropriate because we know the target distribution. For real applications, however, we do not have this luxury and the analyst needs to decide what level of smoothness is appropriate. The simulation studies indicate that a value for κ between 0.2 and 0.4 seems to work reasonably well for most distributions shown here.

4.4 Application to residue data

In this section, we apply the DPMN model to unit pesticide field trial data. All data analysed in this section are log-transformed before the analysis to account for the fact that residue levels should be larger than zero and that their distributions tend to have long upper tails. Before we can apply the DPMN approach we briefly summarise the available data and the selection process of which data are included in the analysis.

4.4.1 Data

Unit residue data are often collected as part of research programmes. For example, Ambrus (2006) reports a large set of unit residue data that were collected as part of a research programme which was intended to provide a better estimate of variability factors. These research studies are often conducted under different circumstances to supervised field trials, e.g. by applying more than one pesticide in a tank mix or measuring residue levels immediately after the last application to increase the probability of obtaining measurable residue levels. Ambrus (2006) states that the variability of residue levels is not significantly affected by the average residue level, nor by the time interval between the application and sampling, so we assume that the shape of the distribution is not affected by these deviations either.

Unit field trials in which tank mixes are applied result in residue levels which cannot be considered to be independent. An analysis of the data shows that rank correlation coefficients between residue levels of pesticides in a tank mix are often very high, i.e. 66% of rank correlations are larger than 0.75 and 39% are higher than 0.9. To deal with this, we could use average concentrations (EFSA, 2005), but there are two disadvantages of this: firstly, we need to decide how to deal with <LOD data as we cannot calculate the average concentration for units on which one of the pesticides could not be quantified. Secondly, the data may be influenced heavily by measurement imprecision for those residue levels that are close to the limit of determination. To overcome these issues, we apply two somewhat arbitrary rules

to select a single data set for those field trials that were conducted with multiple chemicals:

1. Select the data sets with the highest number of measurements above the LOD.

2. If more than one data set was available after applying the rule above, the data set for the pesticide with the highest average residue level was selected. The argument for this is that we assume that the higher the residue level, the more reliable the residue level estimation will be.

Field trial data were available for 164 pesticide/crop combinations. However, as pesticides were applied in tank mixes, only 75 independent pesticide/crop combinations could be used for the analysis, using the selection process explained above. For those data sets, residue levels were measured on between 66 and 319 crop units, resulting in a data set consisting of 9314 normalised log-residue values (see Appendix A for details).

4.4.2 Results

Figure 4.7a shows the distribution of the common shape for the field trial log-residue data. Due to the number of data available, the uncertainty about the shape is relatively small. As explained in Section 4.2.1.4, we need to explore whether the number of components used in the DPMN model is adequate. The tail probabilities for the three truncated Dirichlet Processes for each of the tertiles, given $\gamma = 10$ and using 201 components in the mixture, are given in Figure 4.7b. Even though these probabilities may appear relatively small at first sight, the question as to whether they are small enough depends on the protection target of probabilistic dietary risk assessments. Assigning a probability of 10^{-6} to a single component rather than multiple components may affect residue levels for food items that are consumed on a regular basis by a large proportion of the population of interest. However, the impact on the analysis will most likely be relatively small as the tail probability may be spread over the whole tertile range and could then be considered as random noise that is almost negligible in comparison with the uncertainty of the shape, location and scale parameters. The shape distribution in Figure 4.7a can be combined with

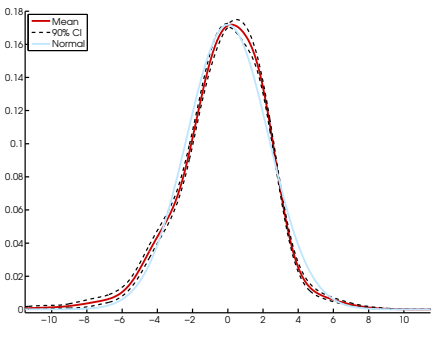
the uncertain location and scale parameters to obtain an estimate of the posterior distribution for an individual data set (Figure 4.8). The approach is compared with a kernel density distribution of the data.

Comparison with current approach

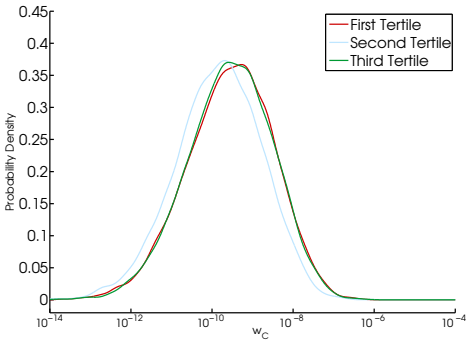
There are no existing models which use unit residue data directly, as most probabilistic risk models are based on a variability factor approach, assuming a Lognormal distribution for unit residue levels. Therefore, it seems appropriate to compare the results of the DPMN model with a Lognormal shape. Figure 4.7a shows that the shape obtained from the DPMN model is similar to the Normal distribution, although the DPMN model has longer tails. However, the Normal shape results in higher estimates for percentiles up to the first tertile and between the second tertile and the 99^{th} percentile. Therefore, the 97.5^{th} percentile of the DPMN shape distribution is generally lower than that of the Normal distribution. This is illustrated in Figure 4.9 where we compare the 97.5^{th} percentile of the log-residue data estimated using the DPMN model with using a Normal distribution. The percentiles have been rescaled by empirical estimates of the 97.5^{th} percentile to aid the comparison. We do not know which method provides a better representation of unit residues

Figure 4.7 – *Results from applying the DPMN model (with* $\kappa = 0.3$ *and* $\gamma = 10$*) to log-transformed field trial data.*

(a) Median (red line) with 95% credible interval (black dashed line) compared with a Normal distribution shape (blue line).



(b) Tail probability w_C for the field trial model runs.



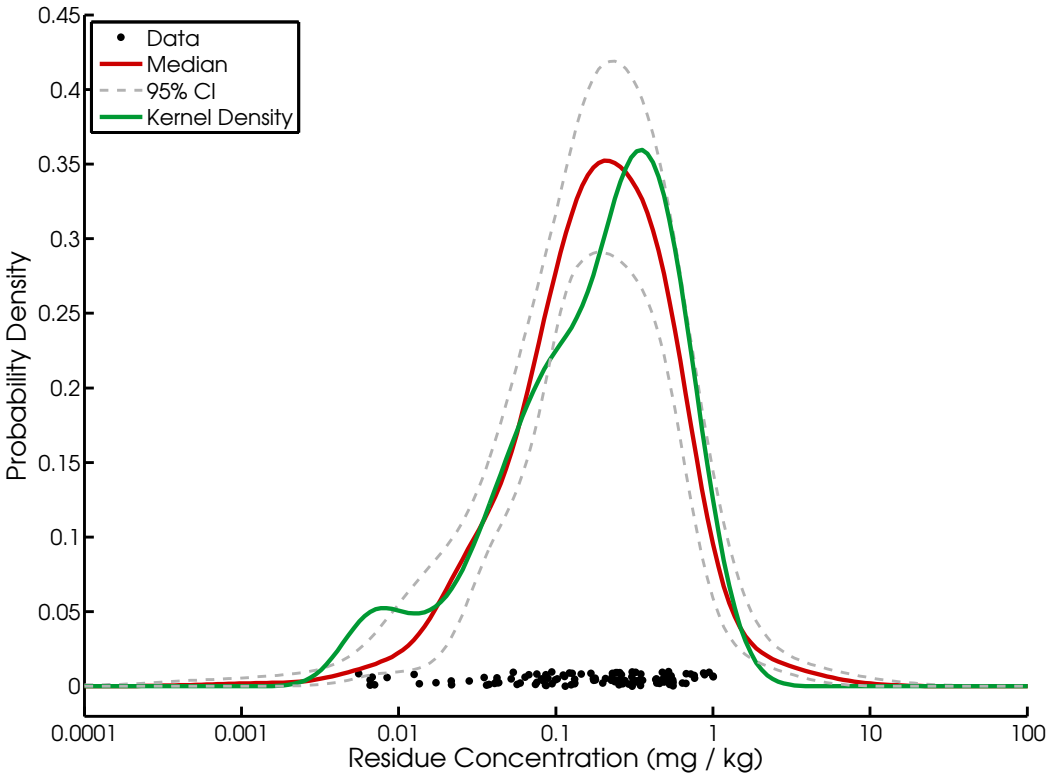


Figure 4.8 – *Posterior distribution for one of the field trial data sets using the DPMN model with* $\kappa = 0.3$ *and* $\gamma = 10$. *Median (red line) with 95% credible interval (grey dashed line) compared with a kernel density estimate (green line).*

because we do not know the true distribution. However it is interesting to note that including the uncertainty about the distribution shape has generally not resulted in wider uncertainty intervals for the 97.5^{th} percentile. If the whole distribution is taken into account, further differences can be seen (Figure 4.7a) which will result in different exposure distributions if both approaches were to be used in a probabilistic dietary risk assessment.

Effect of crop and pesticide

Next we explore whether there are crop and/or pesticide effects which could be used to refine the model. For this purpose, we cannot just look at quantiles of the distributions as they are affected by application rates which may vary between pesticide/crop combinations. Therefore, to account for the effect of application rates,

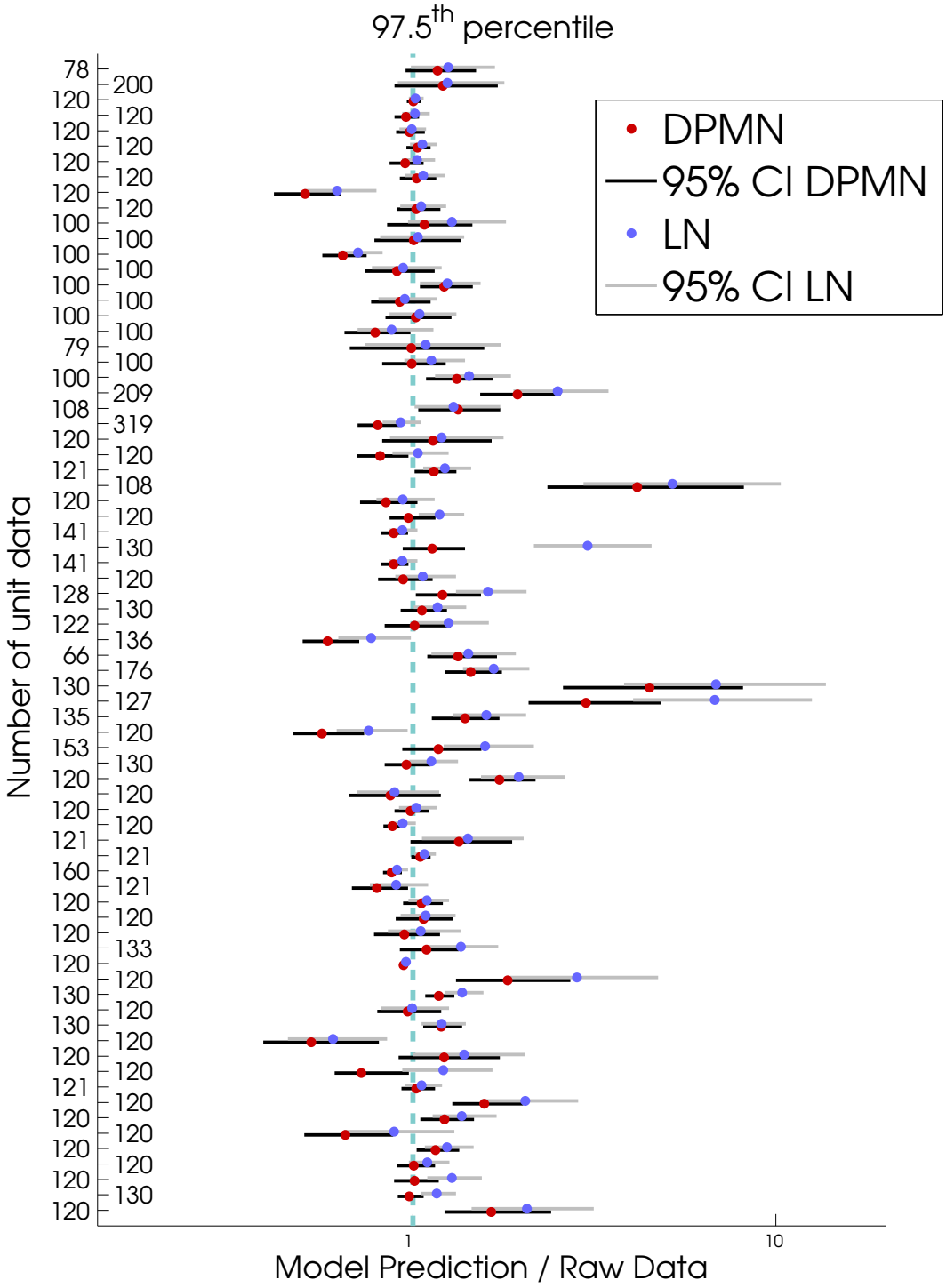


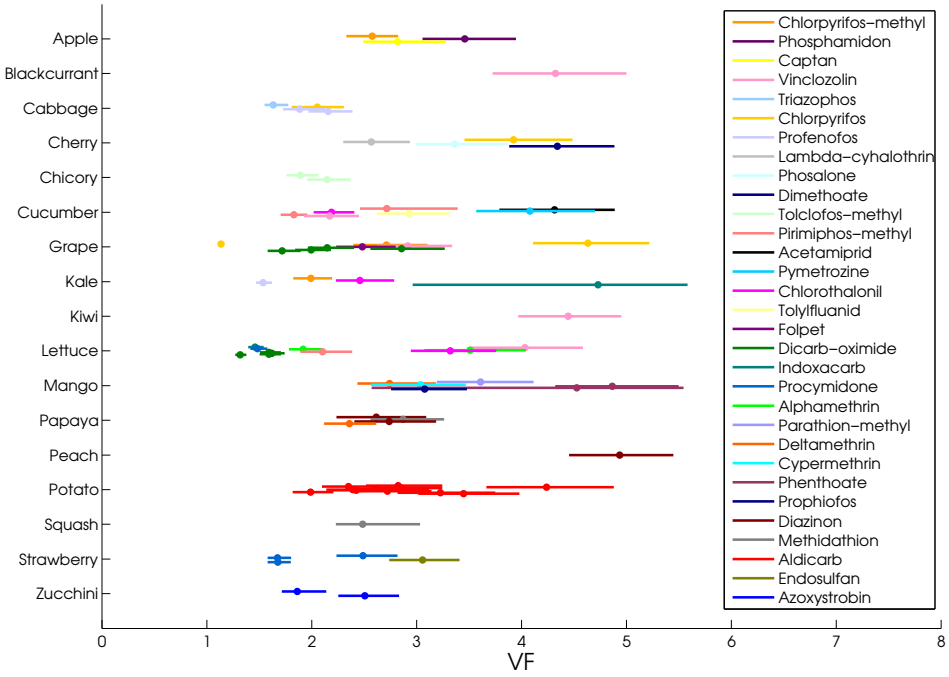
Figure 4.9 – *Comparison of the distribution of the* 97.5^{th} *percentile of field trial data using a Lognormal distribution (median (blue dot) with 95% credible interval (grey lines)) and using a DPMN model with* $\kappa = 0.3$ *and* $\gamma = 10$ *(median (red dot) with 95% credible interval (black lines)). The blue dashed line represents the case where the* 97.5^{th} *percentile of the model predictions is equal to the* 97.5^{th} *percentile of the data.*

a comparison of variability factors calculated using the DPMN shape distribution seems to be more appropriate. A simple graphical comparison (Figure 4.10) was conducted to assess whether any patterns could be explained by either the crop type or the pesticide. As there is limited data available it is not feasible to assess whether there are crop and/or pesticide effects.

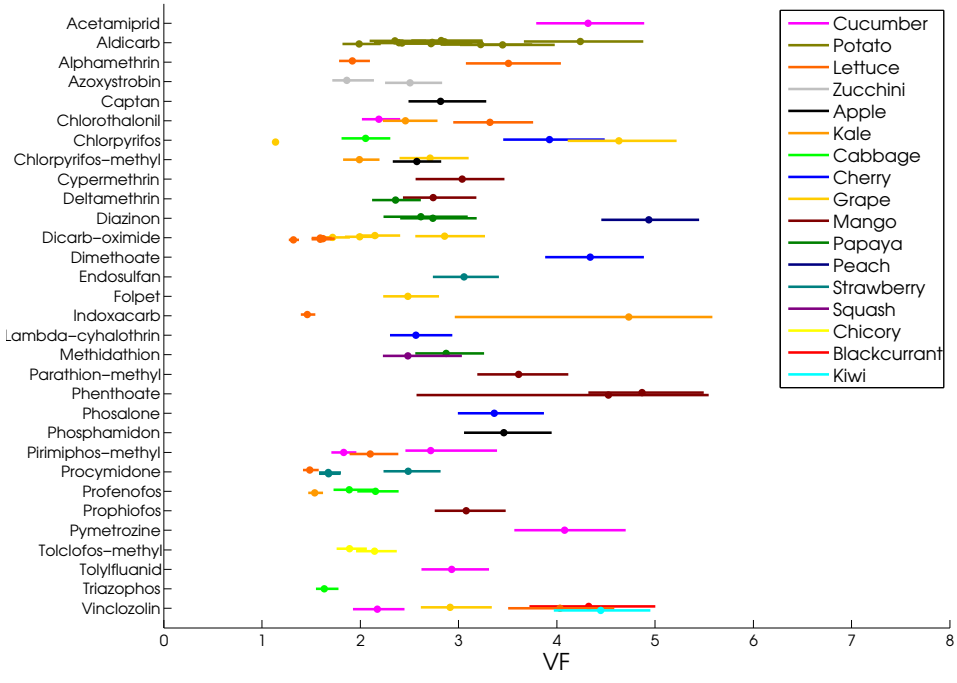
It is interesting to see that the variability factor varies between field trials, for example field trials with aldicarb on potatoes resulted in median estimates of variability factors ranging from 2 to 4. Our analysis produced similar variability factors to those presented in EFSA (2005) and also supports their conclusion that variability factors vary between datasets.

Figure 4.10 – *Comparison of variability factors (VFs), based on field trial data, estimated using the DPMN model with $\kappa = 0.3$ and $\gamma = 10$ for different crops and pesticides. Median estimates (dots) and 95% credible intervals are given for each data set.*

(a) Crop



(b) Active Ingredients



4.5 Inferring the shape distribution for individual data sets

The assumption of a single shape shared between all pesticide/crop combinations may not be realistic. A simple screening analysis to test the shared shape assumption would be to apply the DPMN model to each data set separately. This will lead to an increase in uncertainty for the location, scale and shape parameters for the log-residue data. Although this would take away the advantage of sharing information between data sets, it allows for more flexible shapes and may therefore be useful in cases when unit data are available for a new pesticide/crop combination where (a) the shape seems to be different to the shape obtained from existing data sets and (b) the data cannot be described by a standard parametric family of distributions. Figure 4.11 shows the 97.5^{th} percentile of the DPMN model when applied to the data set described in Section 4.4.1, assuming a shared shape distribution and when applied to each data set individually. The results indicate that the estimates of the 97.5^{th} percentile are very similar for a large proportion of data sets. In those cases where the estimate is different, the data seem to indicate that the distribution shape inferred using the DPMN model may not be applicable to that data set. To investigate this further, we focus on data sets 34 and 48 which appear to have a different shape. Figure 4.12 shows the shared shape distribution compared with the shape obtained by running the DPMN model on data sets 34 and 48 individually. Figure 4.12b indicates that a different shape may be needed to describe data set 34 whereas in Figure 4.12d we observe that there is only a difference in shape due to a cluster of data below the LOD. Therefore, it may still be reasonable to assume a shared shape for this data set. This also illustrates the value of pooling normalised data to obtain a shape estimate because it leads to a smoother representation of the shape distribution than one obtained from one small data set. Ideally, we would like to either select the number of shapes that are necessary to describe the data using objective criteria (e.g. based on crop or pesticide characteristics) or by letting the model determine the number of shapes that are needed. We will discuss further refinements of the model in Chapter 6.

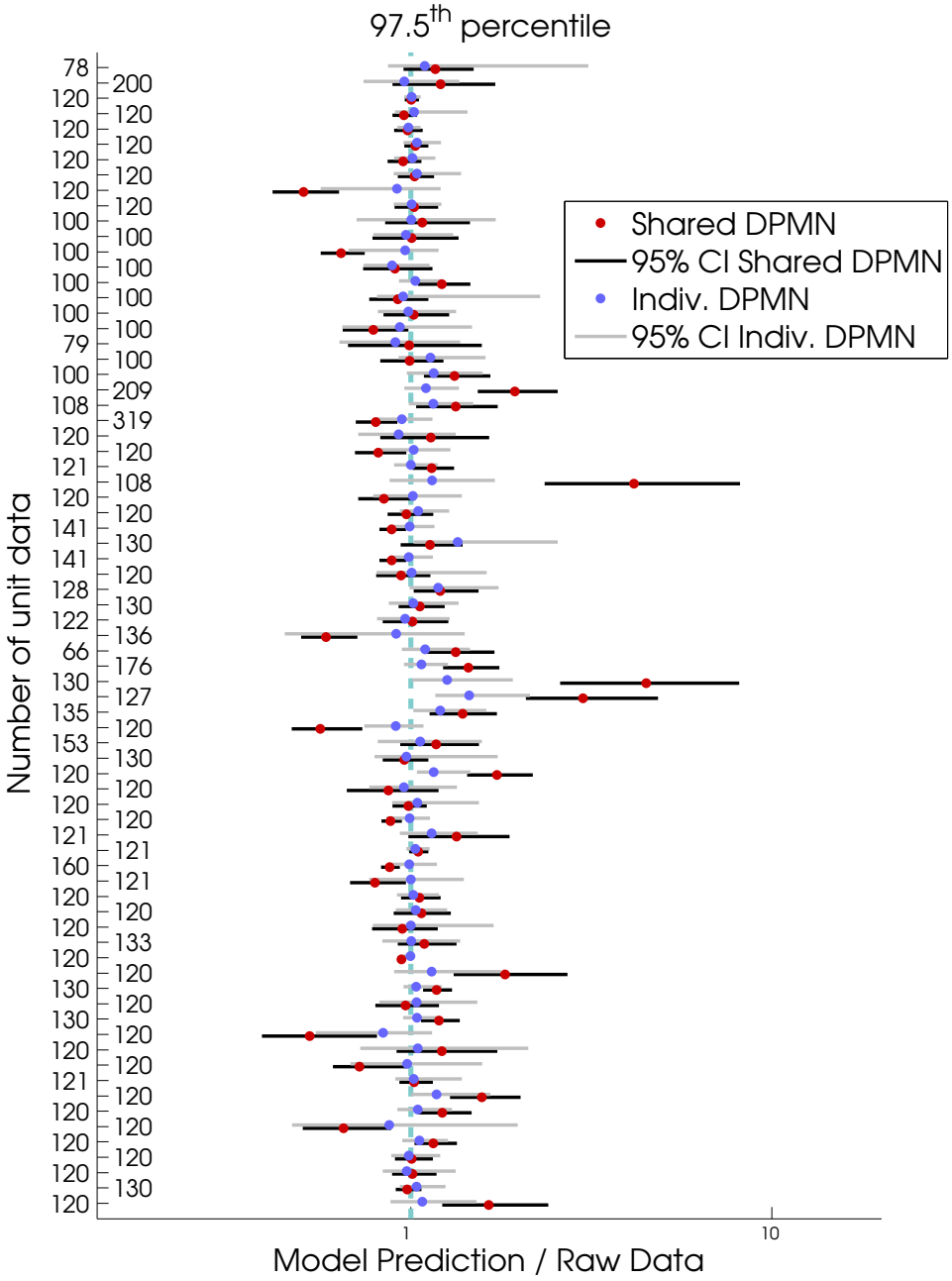
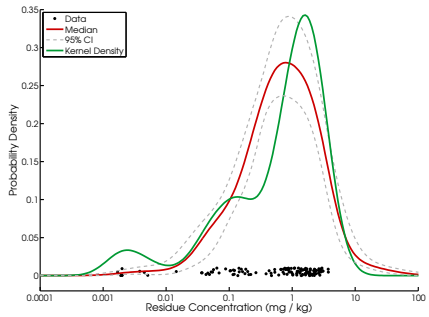


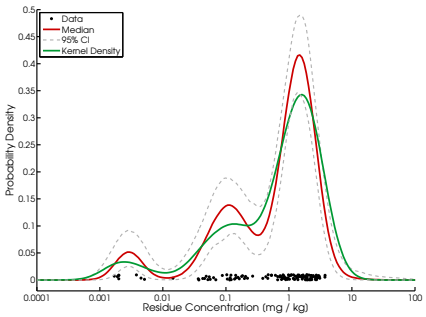
Figure 4.11 – *Comparison of the 97.5th percentile residue level based on the DPMN model ($\gamma = 10$, $\kappa = 0.3$) applied to the selected field trial data sets (median (red dot) and 95% credible intervals (black lines)) and to individual pesticide/crop combinations (median (blue dot) and 95% credible intervals (grey lines)). Data sets 34 and 48 have been numbered as they will be discussed in more detail. The blue dashed line represents the case where the 97.5th percentile of the model predictions is equal to the 97.5th percentile of the data.*

Figure 4.12 – *Difference in shape distributions when the shape distribution is assumed to be shared or not shared between pesticides for data sets 34 and 48. Median (red line) with 95% credible interval (grey dashed line) compared with a kernel density estimate (green line).*

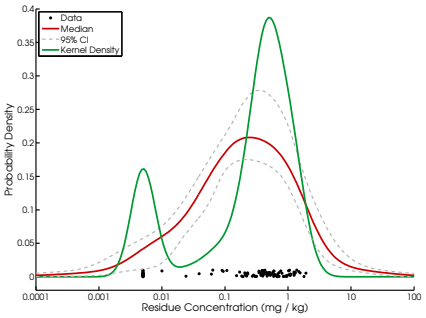
(a) Shared shape for data set 34



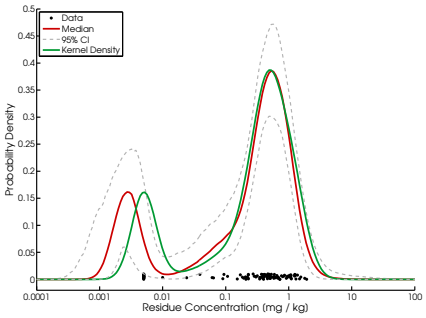
(b) Not shared shape for data set 34



(c) Shared shape for data set 48



(d) Not shared shape for data set 48



4.6 Uncertain γ

The results presented so far required the selection of the Dirichlet Process concentration parameter γ. An alternative approach would be to treat γ as a parameter, which was first suggested by Ishwaran and Zarepour (2000). As demonstrated in Section 2.2.4.3, the stick-breaking representation for a truncated Dirichlet Process, based on C components, leads to weights $\mathbf{w}$ which have a Generalised Dirichlet$(1, \gamma, \ldots, 1, \gamma)$ distribution:

$$p(\mathbf{w}|\gamma) \propto \gamma^{C-1} w_C^{\gamma-1}$$

As the model presented in this chapter consists of a DPMN model for each of the three tertiles, we have:

$$\begin{aligned} p(\mathbf{w}|\gamma) &\propto \gamma^{C^{(1)}-1} w_{C^{(1)}}^{\gamma/3-1} \gamma^{C^{(2)}-1} w_{C^{(2)}}^{\gamma/3-1} \gamma^{C^{(3)}-1} w_{C^{(3)}}^{\gamma/3-1} \\ &\propto \gamma^{C-3} (w_{C^{(1)}} w_{C^{(2)}} w_{C^{(3)}})^{\gamma/3-1} \end{aligned}$$

where $C^{(i)} = C/3$ is the number of components for tertile i. Using a Gamma(v_1, v_2) prior for γ, the posterior distribution becomes:

$$p(\gamma|\mathbf{w}) = \text{Gamma}\left(\gamma; C + v_1 - 3, v_2 - \frac{\log(w_{C^{(1)}}) + \log(w_{C^{(2)}}) + \log(w_{C^{(3)}})}{3}\right) \tag{4.11}$$

After sampling γ from its posterior, we can assign $\gamma/3$ to each tertile.

4.6.1 Choice of Prior Distribution

As γ is now a random variable, we have to assign a prior distribution to γ. In the simulations in the remainder of this section, we have used a Gamma$(2, 0.25)$ distribution as a prior distribution for γ. This is an arbitrary choice that is merely used to illustrate the approach. The prior mean equals 8 which is similar to the fixed value that was used before ($\gamma = 10$). The main reason for selecting this prior was that the probability of $\gamma \leq 1 \approx 2.6\%$ and $\gamma \geq 25 \approx 1.4\%$. We do not want very small values for γ because we want the posterior to be somewhat smooth nor do we want γ to be very large because it leads to issues with high tail probabilities (see Section 2.3.5.3).

4.6.2 Simulation Studies

4.6.2.1 Using validation data sets

To assess the performance of the model with uncertain γ we ran the model on the data sets that we used in the validation studies earlier. The simulations were conducted using $\kappa = 0.3$ and the fitted distributions and posterior distributions of γ are shown in Appendix B. The figures indicate that the distribution fit is very similar to the model output when γ was fixed. The posterior densities of γ for almost all

simulation studies are in the region 1 to 5 with the exception of the simulation study for the Student's t distribution with 3 degrees of freedom, which is in the range 5 to 20 with a mode around 12. Although our fixed value of γ was generally larger than the mode and range for γ inferred from the data, this did not have a noticeable effect on the posterior shape distribution because γ is small in comparison with the sample size ($n = 1000$). These results imply that learning γ from the data has very little influence and that it can therefore be omitted for simplicity. However, it is unclear whether these results can be extrapolated to other applications and therefore we recommend using a model that infers γ from the data to assess whether a chosen fixed value is reasonable.

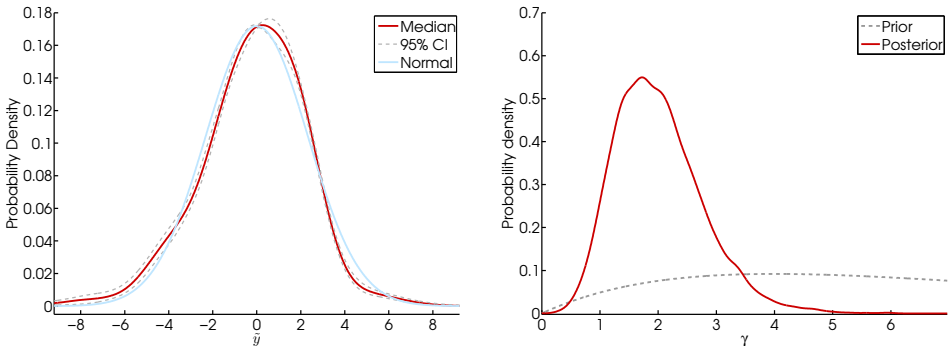
4.6.2.2 Using Unit Log-Residue Data

Figure 4.13a shows the shape distribution resulting from applying the DPMN model to log-residue field trial data when γ was considered to be an uncertain parameter. As in the simulation studies, the shape distribution with uncertain γ is very similar to the results when γ was fixed. If we look at the posterior distribution of γ in Figure 4.13b, we observe that the mode of the posterior distribution is around 2 with γ ranging from 0.5 to 5, slightly lower than the selected fixed value of 10. Although our fixed value of γ was larger than the mode and range for γ inferred from the data, this did not have a noticeable effect on the posterior shape distribution because γ is small in either case in comparison with the size of the normalised log-residue data set which had 9314 values.

Figure 4.13 – *Results from applying the DPMN model (with $\kappa = 0.3$ and uncertain γ) to log-transformed field trial data.*

(a) Median (red line) with 95% credible interval (grey dashed line) compared with a Normal distribution shape (blue line).

(b) Posterior distribution of γ.



4.7 Discussion

In this chapter we have introduced a novel approach which allows information on common population characteristics to be shared between multiple data sets. We use a blocked Gibbs sampler to alternate sampling the individual location and scale parameters of each data set and the common shape distribution using the normalised, pooled log-residue data. The approach for estimating the shape distribution is based on Dirichlet Process Mixture models that have been extensively used in semi-parametric models to describe data sets obtained from a single population. Sharing shape information between samples obtained from multiple populations with a common shape leads to a larger data set from which we can infer the population shape. The Bayesian framework used in this model allows us to account for parameter uncertainty. Model runs generally took around 70 minutes to complete. This is much longer than the currently used models due to the extra complexity. Whilst this complexity is desirable to get better estimates, this may impact on the ability to determine the sensitivity of the model to assumptions (e.g. parameter values, distribution choices).

The model was applied to unit log-residue field trial data sets. As these data are not routinely collected, acute probabilistic dietary risk models need to assume a shape for unit residue levels based on surrogate data. Unlike existing approaches which assume a Normal distribution for log-residues, our method uses data to learn the shape of the distribution. As there is evidence that the Normal distribution may not always provide an adequate fit to the log-residue data (EFSA, 2012), the new approach is an improvement as it provides a better description of the data.

The method assumes that a single distribution shape can be used to model the variation in log-residue levels, whilst acknowledging that the location and scale of log-residue level populations will vary between pesticide/crop combinations. However the field trial data indicates that some of the data sets may in fact have different distribution shapes, although this is difficult to assess for small data sets. If we did not believe that all the data sets shared a common shape, we could identify groups of pesticide/crop combinations which are thought to share a distribution shape and run our model on the subsets of data. Alternatively we could extend the model to allow for multiple shape distributions and infer a clustering from the data. This is discussed further in Chapter 6.

The DPMN model presented here differs from existing DPMN models because it is used to model the distribution shape of log-residues. For our application we also had to refine the DPMN model to deal with censored and rounded data. The performance of our method was extensively tested in validation studies. The validation studies indicated that the model performs well for a range of distributions with short and medium tails. However for heavy-tailed distributions the method would need to be refined if a transformation could not be applied to remove the heavy tail of the data. Suggestions for refining the model to deal with heavy-tailed distributions are discussed in Chapter 6 where we also discuss further options for model validation. We investigated the effect of sample size and concluded that for $n > 50$ the model performed well. The effect of learning the DPMN concentration parameter γ for the validation study data and the unit log-residue data was shown to be negligible

because the inferred range of γ values was relatively similar to the fixed value of γ compared to the pooled sample size.

Applying the DPMN model to unit log-residue data allows us to estimate variability factors for each data set. The VFs obtained using our model on field trial data indicate that a value of 3 as proposed by Hamilton et al. (2004) and Ambrus (2006) may not be sufficiently protective, a result that is in line with the EFSA analysis (EFSA, 2005). Our method improves on the current VFs because it also provides the unit variation distribution which can then be used in a model that accounts for between-field variation (see Chapter 5). A hierarchical model that describes the variation in variability factors between data sets, similar to the EFSA (2005) analysis but based on a DPMN distribution may be more appropriate to describe unit variation. We present such a model in Chapter 5.

The model presented in this chapter can also be used in other applications including modelling consumption data and composite supervised trial data. Outside dietary risk assessment, the approach could also be used for ecotoxicity data that are used to describe the variation in sensitivity between species to chemicals. These applications are discussed in more detail in Chapter 6.

Chapter 5

Modelling within-field and between-field variation in pesticide residues

5.1 Introduction

To estimate the acute dietary risk associated with pesticide residues in food items, we need to know how residue levels vary between food items. Variation in pesticide residues is thought to be affected by four types of factor: application, crop, environmental and dissipation factors (Ambrus, 1979; EFSA, 2005). Prior to pesticide registration for a new use (i.e. the use of a new pesticide on any crop or an existing pesticide on a new crop), there are two sources of data available that could be of use: supervised trial data and unit field trial data (see Section 1.4.1). As unit field trial data are generally not collected for pesticide registration it has been suggested (JMPR, 1999; Ambrus, 2000; JMPR, 2002; Hamilton et al., 2004; EFSA, 2005; Ambrus, 2006) that existing unit data from field trials for other crops and pesticides can be used to provide an estimate of the amount of variation between units. These data were used in Chapter 4 to describe the variation in unit residue levels (within-field variation) for multiple field trials. However, to predict residues levels in consumed food items we also need to account for variation between unit residue levels that are

obtained from different fields (between-field variation).

To describe the within-field and between-field variation in residue levels, we need to obtain residue levels on food items from multiple fields under a wide range of conditions. This is important as for some food items (e.g. a bunch of bananas), a consumer is likely to be exposed to units obtained from a single field, whereas for others, e.g. a bag of apples, the bag may contain units that come from the same or multiple fields. As a consequence, when estimating the dietary exposure of consumers, we need to be able to quantify both the within-field and between-field variation in residue levels.

In current probabilistic dietary risk assessment approaches, a distribution is fitted to composite samples from multiple supervised trials and this is then regarded as a distribution of field means. However, this approach ignores uncertainty in the estimation of the field means and results in a distribution describing a mixture of between-field and unit variation. As a result, the current methods count unit variation twice and do not account for the uncertainty caused by the low number of units used to create composite samples. To overcome this, we propose a new statistical model that will a) provide a more realistic description of residue level variation in units than the currently assumed Lognormal distribution, b) take account of the small number of units used in composite samples and the small number of composite samples used to describe between-field variation and c) account for the fact that composite residue levels from supervised trials already include unit variation. The proposed method uses the same information as existing methods (e.g. composite samples from supervised trial data) and can therefore be implemented in existing software.

5.2 Model Specification

In this section, we propose a novel approach to model variation in residue levels that will not only be based on a data-driven description of unit variation but, un-

like existing approaches, it will not double-count unit variation when accounting for within-field and between-field variation. When inferring the field mean distribution we need to 'remove' the unit variation component in the observed variation in composite samples in order to obtain a distribution describing the variation in field mean residue levels.

Let us assume that residue levels on units k from field l can be described using:

$$y_{kl} = \xi_l^F U_{kl}$$

where ξ_l^F is the mean residue level of field l and U_{kl} is the 'relative unit variation' with $\mathbb{E}[U_{kl}] = 1$. We can use a refined version of the unit residue model developed in Chapter 4 as the basis for our model for U_{kl}. We can then fit this model to the unit log-residue data in Chapter 4 and use the resulting posterior distribution as a prior distribution for U_{kl}.

5.2.1 Refined unit model

Figure 5.1 shows the Directed Acyclic Graph (DAG) of the unit log-residue variation model from Chapter 4 but now refined to model the variation in the scale parameter of the unit log-residue distribution, σ_l^u, using a Gamma distribution with parameters α and β. The data requirements for the registration of a new pesticide use result in a set of composite samples from supervised field trials which will be used to infer the between-field variation. Each composite sample, $\overline{y_l}$, is the average residue level of n_l units obtained from field l, $\overline{y_l} = \frac{\sum_{k=1}^{n_l} y_{kl}}{n_l}$, where $l = 1, \ldots, L$ and L is the number of supervised field trials. As no unit data will be collected as part of the registration process, the information on unit variation in log-residue levels needs to come from other sources. To infer the within-field variation, we have unit field trial data, x_{ij}, with $i = 1, \ldots, n_j$ and $j = 1, \ldots, J$, where n_j is the number of units obtained from field trial j and J is the number of unit field trials. In the model presented in this chapter we have unit data, x_{ij}, from J fields and composite data, y_{kl} from L fields. Throughout this chapter we assume that the variation in unit residue levels in fields from which composite samples were obtained can be described using the unit data even though the measurements were not taken from the same fields. Therefore we

use different indices to denote the different fields that the unit and composite data were collected from. After normalising the unit log-residue data, x_{ij}, using location parameter, μ_j^u, and scale parameter, σ_j^u, we model the normalised log-residue values, z_{ij}, using a Dirichlet mixture shape distribution with weight and location parameters $\mathbf{w}$ and $\boldsymbol{\theta}$. We do not need a hierarchical model for the μ_j^u because changing the value of μ_j^u changes the distribution of x_{ij} but not the distribution of U_{ij}, where U_{ij} is defined as:

$$U_{ij} = \frac{\exp[x_{ij}]}{\mathbb{E}\left[\exp[x_{ij}]|\mu_j^u, \sigma_j^u, \mathbf{w}, \boldsymbol{\theta}\right]}$$

5.2.2 Within- and between-field model

To model composites, $\overline{\mathbf{y}}$, we need to model between-field variation so we introduce a location parameter, μ^F, and a scale parameter, σ^F, to describe the variation in field means, ξ^F. This results in the following joint probability density function:

$$\begin{aligned} &p(\mathbf{w},\boldsymbol{\theta},\alpha,\beta)p(\mathbf{x}|\mathbf{w},\boldsymbol{\theta},\alpha,\beta) \\ &\quad\times p(\mu^F,\sigma^F)p(\overline{\mathbf{y}}|\mu^F,\sigma^F,\mathbf{w},\boldsymbol{\theta},\alpha,\beta) \end{aligned}$$

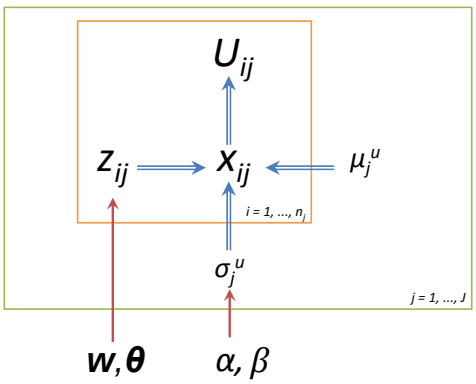


Figure 5.1 *– DAG for refined unit residue generation model which accounts for within-field variation of log-residue levels. The red arrows correspond to the dependencies between the variables and the blue arrows represent deterministic dependencies.*

However, as the distribution $p(\overline{\mathbf{y}}|\mu^F, \sigma^F, \mathbf{w}, \boldsymbol{\theta}, \alpha, \beta)$ is complex, we need to include the field-specific scale parameter, σ_l^u, as an auxiliary parameter in the model:

$$p(\mathbf{w},\boldsymbol{\theta}, \alpha, \beta)p(\mathbf{x}|\mathbf{w}, \boldsymbol{\theta}, \alpha, \beta) \times p(\mu^F, \sigma^F) \prod_l^L p(\sigma_l^u|\alpha, \beta)p(\overline{\mathbf{y}}|\mu^F, \sigma^F, \mathbf{w}, \boldsymbol{\theta}, \sigma_l^u, \alpha, \beta)$$

As the conditional distributions $p(\sigma_l^u|\overline{\mathbf{y}}, \mu^F, \sigma^F, \alpha, \beta)$ and $p(\mu^F, \sigma^F|\overline{\mathbf{y}}, \boldsymbol{\sigma}^{\mathbf{u}}, \alpha, \beta)$ are still complex, we propose to introduce another auxiliary variable, $\overline{U_l} = \frac{1}{n_l}\sum_{k=1}^{n_l} U_{kl}$, describing the variation of composite residue levels around the field mean, ξ_l^F:

$$\log(\overline{y_l}) = \log(\xi_l^F) + \log(\overline{U_l})$$

In other words, residue levels on units and composite samples are a function of the variation in field means and the variation in units, U_{kl}, or composites, $\overline{U_l}$, respectively. The distribution of $\overline{U_l}$ depends on the number of units, n_l, that are used in a composite sample and the relative unit variation, U_{kl}. The refined unit model provides us with the distribution of U_{kl}, but the distribution of $\overline{U_l}$ is unknown. We now have the following joint probability density function:

$$p(\mathbf{w},\boldsymbol{\theta}, \alpha, \beta)p(\mathbf{x}|\mathbf{w}, \boldsymbol{\theta}, \alpha, \beta) \times p(\mu^F, \sigma^F) \prod_l^L p(\sigma_l^u|\alpha, \beta)p(\overline{U_l}|\mathbf{w}, \boldsymbol{\theta}, \sigma_l^u)p(\overline{\mathbf{y}}|\overline{U_l}, \mu^F, \sigma^F)$$

resulting in the blocked Gibbs conditional distributions:

$$p(\mathbf{w}, \boldsymbol{\theta}, \alpha, \beta|\mathbf{x}, \overline{\mathbf{U}}, \boldsymbol{\sigma}^{\mathbf{u}}) \tag{5.1}$$

$$p(\sigma_l^u|\overline{U_l}, \alpha, \beta) \tag{5.2}$$

$$p(\overline{U_l}|\overline{y_l}, \mu^F, \sigma^F, \mathbf{w}, \boldsymbol{\theta}, \sigma_l^u) \tag{5.3}$$

$$p(\mu^F, \sigma^F|\overline{\mathbf{y}}, \overline{\mathbf{U}}) \tag{5.4}$$

5.2.3 MCMC Approach

Equation 5.1 simplifies to $p(\mathbf{w}, \boldsymbol{\theta}, \alpha, \beta|\mathbf{x})$ assuming that the number of data, $\overline{y_l}$, for the crop-pesticide scenario of interest is relatively small and provides little information on $\mathbf{w}$, $\boldsymbol{\theta}$, α and β. Although there is some information about unit variation in

the composite data, $\overline{y_l}$, we ignore this because the number of units in the composite samples is small compared to the number of unit data, x_{ij}. In addition, we expect that a significant proportion of the total variation of $\overline{y_l}$ is due to the variation in field means which makes it difficult to extract information about the unit variation from the composite samples, particularly as the scale parameter of the unit log-residue distribution, σ_l^u is field-specific rather than crop/pesticide specific.
Therefore we only use the unit data, x_{ij}, to infer α, β and the parameters of the Dirichlet Process Mixture distribution obtained from the refined unit model and assume that the distribution for unit variation based on the field trial data, x_{ij}, can be used to describe the unit variation for new pesticide/crop combinations, U_{kl}. This relies on two assumptions:

1. The U_{ij} (or U_{kl}) are exchangeable within a field j (i.e. U_{ij} are independent and identically distributed given σ_j^u, $\mathbf{w}$ and $\boldsymbol{\theta}$).

2. The only parameter that varies between fields is σ_j^u and these are exchangeable and independent and identically distributed given α and β.

The use of existing unit data, x_{ij}, for multiple crop/pesticide combinations to infer the unit variation for a new crop/pesticide combination is supported by the results of the analysis in Chapter 4 and JMPR (2003), which indicated that the variability factor and hence the scale parameter does not seem to be dependent on the crop type or pesticide. As a result, the variation in unit residue levels for a new use can be regarded as a random sample from the refined unit model applied to a suitably representative sample of existing unit data sets. In a similar way to Equation 5.1, we can simplify Equation 5.2 to $p(\sigma_l^u|\alpha,\beta)$ as $\overline{U_l}$ contains little information about σ_l^u.

We cannot calculate nor sample from Equation 5.3 as we do not know the distribution of $\overline{U_l}$, even without conditioning on $\overline{y_l}$. Instead we propose to sample from $p(\mathbf{U_l}|\overline{y_l},\mu^F,\sigma^F,\mathbf{w},\boldsymbol{\theta},\sigma_l^u)$ and use the fact that if we know $\mathbf{U_l}$, we also know $\overline{U_l}$. We now have:

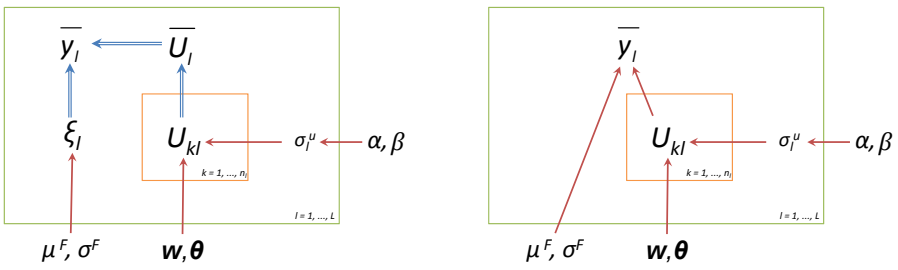
$$p(\mathbf{U_l}|\overline{y_l},\mu^F,\sigma^F,\mathbf{w},\boldsymbol{\theta},\sigma_l^u) \propto p(\overline{y_l}|\overline{U_l},\mu^F,\sigma^F)\prod_{k=1}^{n_l} p(U_{kl}|\mathbf{w},\boldsymbol{\theta},\sigma_l^u) \qquad (5.5)$$

We can calculate Equation 5.5, but as it is difficult to sample from this distribution, we propose a Metropolis-Hastings algorithm in which we use our knowledge of the moments of U_{kl} to obtain a Lognormal approximation of $p(\overline{U_l}|\overline{y_l}, \mu^F, \sigma^F, \mathbf{w}, \boldsymbol{\theta}, \sigma_l^u)$. We can use this approximation to obtain a proposal value for ξ_l^F which, given $\overline{y_l}$, can be used to calculate a proposal value $\overline{U_l}^*$. We subsequently sample proposal values, $\mathbf{U}_l^*$, from $p(\mathbf{U}_l|\overline{U_l}, \mathbf{w}, \boldsymbol{\theta}, \sigma_l^u)$ by approximating $p(U_{kl}|\mathbf{w}, \boldsymbol{\theta}, \sigma_l^u)$ with a Gamma distribution. We then accept/reject the proposed values, $\mathbf{U}_l^*$ and $\overline{U_l}^*$, using the target probability distribution in Equation 5.5. We have chosen this approach instead of using random walk steps because a random walk in n_l dimensions may not be very efficient when n_l is large. We will discuss the logic behind using these proposal distributions in more detail in Section 5.2.6. Updating μ^F and σ^F (Equation 5.4) is straightforward given $\boldsymbol{\xi}^{\mathbf{F}}$ if we assume a conjugate distribution for ξ^F (see Section 5.2.7). Throughout this chapter we will assume that ξ^F follows a Lognormal distribution as is current practice in existing dietary exposure approaches.

Figure 5.2 shows DAGs describing the within-field and between-field variation model. Figure 5.2a highlights the structure of the conceptual model. However, the MCMC algorithm that we implement is a blocked Gibbs sampler in which we replace some Gibbs steps with Metropolis-Hastings steps. As you cannot use a Gibbs sampler on a DAG with logical dependencies (Lunn et al., 2000), Figure 5.2b illustrates which quantities are sampled in the blocked Gibbs algorithm.

Figure 5.2 *– DAGs for our model describing within-field and between-field variation in residue levels.*

(a) Including logical dependencies. **(b)** Without logical dependencies.



5.2.4 Summary

To sample from the posterior distribution $p(\mu^F, \sigma^F, \mathbf{w}, \boldsymbol{\theta}, \boldsymbol{\sigma}^u | \overline{\mathbf{y}}, \mathbf{x})$, we propose the following Metropolis-Hastings within blocked Gibbs algorithm:

1. From the refined unit model in Section 5.2.8 and Figure 5.1, sample values for $\mathbf{w}$, $\boldsymbol{\theta}$, α and β.

2. For each field l:

   (a) Sample σ_l^u for each field from the hierarchical Gamma(α, β) model.

   (b) Impose the constraint $\mathbb{E}[U_{kl}] = 1$ on the unit variation distribution, i.e. determine μ_l^u.

   (c) Update U_{kl} using a Metropolis-Hastings step. Using an approximation to the conditional distribution $\overline{U_l} | \overline{y_l}, \mu^F, \sigma^F, \mathbf{w}, \boldsymbol{\theta}, \sigma_l^u$ (see Section 5.2.6.1) we first propose a value for ξ_l^F. Given $\overline{y_l}$, we can use this to calculate a proposal value for $\overline{U_l}$. Then we propose values for the U_{kl} using an approximation to $U_{kl} | \overline{U_l}, \mathbf{w}, \boldsymbol{\theta}, \sigma_l^u$ (see Section 5.2.6.2). Finally we decide whether or not to accept the U_{kl} proposed.

3. Sample μ^F and σ^F from the conditional distribution given the ξ_l^F (see Section 5.2.7).

Before we can discuss the sampling steps for each distribution in more detail, we first need to consider the distribution of U_{ij}.

5.2.5 Distribution of U_{ij}

For the new use of a pesticide, we are not likely to have unit residue data from a range of fields. Therefore, we need to use information about variation in unit residue levels obtained from the refined unit model (see Figure 5.1) to learn about the distribution of U_{ij}. Let us define $z_{ij} = \frac{\log(x_{ij}) - \mu_j^u}{\sigma_j^u}$. From the unit residue model in Chapter 4 we know that z_{ij} are independent and identically distributed given $\mathbf{w}$ and $\boldsymbol{\theta}$:

$$p\left(z_{ij} | \mathbf{w}, \boldsymbol{\theta}, \sigma_c\right) = \frac{1}{\sigma_c} \sum_{q=1}^{C} w_q \phi\left(\frac{z_{ij} - \theta_q}{\sigma_c}\right)$$

where ϕ is the standard Normal density function and σ_c is the fixed standard deviation of the C Normal components in the refined unit model. Setting $\log(x_{ij}) = \mu_j^u + \sigma_j^u z_{ij}$ and $\hat{U}_{ij} = e^{\sigma_j^u z_{ij}}$, we find $x_{ij} = e^{\mu_j^u} e^{\sigma_j^u z_{ij}} = e^{\mu_j^u} \hat{U}_{ij}$. If we define $\rho_j = \mathbb{E}[\hat{U}_{ij}|\mathbf{w}, \boldsymbol{\theta}, \sigma_j^u]$, we can write $x_{ij} = e^{\mu_j^u} \rho_j U_{ij}$ where $U_{ij} = \hat{U}_{ij}/\rho_j$ does not involve μ_j^u and $\mathbb{E}[U_{ij}|\mathbf{w}, \boldsymbol{\theta}, \sigma_j^u] = 1$. We can now write: $\log(U_{ij}) = \log(x_{ij}) - \mu_j^u - \log(\rho_j) = \sigma_j^u z_{ij} - \log(\rho_j)$. Changing variables from z_{ij} to U_{ij} leads to:

$$p(U_{ij}|\mathbf{w}, \boldsymbol{\theta}, \sigma_c, \sigma_j^u) = \frac{1}{U_{ij}\sigma_j^*} \sum_{q=1}^{C} w_q \phi \left(\frac{\log(U_{ij}) - (\theta_{qj}^* - \log(\rho_j))}{\sigma_j^*} \right) \tag{5.6}$$

where $\theta_{qj}^* = \theta_q \sigma_j^u$ and $\sigma_j^* = \sigma_c \sigma_j^u$. To obtain the distribution of U_{ij} we need to compute $\rho_j = \mathbb{E}[\hat{U}_{ij}|\mathbf{w}, \boldsymbol{\theta}, \sigma_j^u]$.

5.2.5.1 Moments of U_{ij}

In this section we derive the moments of U_{ij} which we will use in the following section to propose values for $\overline{U_l}$ and $\mathbf{U_l}$. The moment generating function for $\log(Y)$, where $\log(Y)$ is Normally distributed, is defined as:

$$\mathbb{E}[Y^t] = \exp\left[\mu t + \frac{\sigma^2 t^2}{2}\right]$$

From this it is easy to find the expected value for Y and Y^2 by setting $t = 1, 2$, respectively:

$$\mathbb{E}[Y] = \exp\left[\mu + \frac{\sigma^2}{2}\right]$$
$$\mathbb{E}[Y^2] = \exp[2(\mu + \sigma^2)]$$

The variance of Y is given by:

$$\text{Var}[Y] = (\mathbb{E}[Y])^2(\exp[\sigma^2] - 1)$$

If the probability distribution of a random variable Y is a mixture distribution with weights, w_q, and component probability density function, $p_q(y)$, then we can use an auxiliary discrete random variable, Q, to select a component of the mixture distribution together with the conditional probability density function of $Y|Q = q$, $p_q(y)$, to obtain the moments of the mixture distribution. With $\mathbb{E}[\hat{U}_{ij}^t|\mathbf{w}, \boldsymbol{\theta}, \sigma_c, \sigma_j^u] =$

$\mathbb{E}[e^{t\hat{z}_{ij}}|\mathbf{w}, \boldsymbol{\theta}, \sigma_c, \sigma_j^u]$ where $\hat{z}_{ij} = \log(\hat{U}_{ij})$, $\hat{z}_{ij}|Q_{ij} = q \sim \mathcal{N}(\theta_{qj}^*, (\sigma_j^*)^2)$ and $p(Q_{ij} = q) = w_q$, we obtain the moments of $\hat{U}_{ij}$ conditional on $\mathbf{w}$, $\boldsymbol{\theta}$, σ_c and σ_j^u:

$$\mathbb{E}[\hat{U}_{ij}^t|\mathbf{w}, \boldsymbol{\theta}, \sigma_c, \sigma_j^u] = \sum_{q=1}^{C} w_q \mathbb{E}[e^{t\hat{z}_{ij}}|Q = q]$$

This leads to:

$$\begin{aligned}\rho_j &= \mathbb{E}[\hat{U}_{ij}|\mathbf{w}, \boldsymbol{\theta}, \sigma_c, \sigma_j^u] = \mathbb{E}[\mathbb{E}[\hat{U}_{ij}|Q = q]] \\ &= \sum_{q=1}^{C} w_q \exp\left[\theta_{qj}^* + \frac{(\sigma_j^*)^2}{2}\right]\end{aligned}$$

where w_q is the weight of component q. Conditional on $\mathbf{w}$, $\boldsymbol{\theta}$, σ_c and σ_j^u, the variance of $\hat{U}_{ij}$ can be obtained using:

$$\mathrm{Var}[\hat{U}_{ij}] = \mathbb{E}[(\hat{U}_{ij} - \mathbb{E}[\hat{U}_{ij}])^2] \quad = \mathbb{E}\left[\mathrm{Var}\left[\hat{U}_{ij}|Q = q\right] + (\mathbb{E}[\hat{U}_{ij}|Q = q] - \mathbb{E}[\hat{U}_{ij}])^2\right]$$

With $\mathrm{Var}[\hat{U}_{ij}|Q = q] = \Big(\mathbb{E}[\hat{U}_{ij}|Q = q]\Big)^2 (\exp[(\sigma_j^*)^2] - 1)$ we obtain:

$$\mathrm{Var}[\hat{U}_{ij}] = \sum_{q=1}^{C} w_q((\mathbb{E}[\hat{U}_{ij}|Q = q])^2(\exp[(\sigma_j^*)^2] - 1) + (\mathbb{E}[\hat{U}_{ij}|Q = q] - \mathbb{E}[\hat{U}_{ij}])^2))$$

This can be simplified by taking out a factor of $f = \exp\left[-\frac{(\sigma^*)^2}{2}\right]$ to find $\mathrm{Var}[\hat{U}_{ij}]$:

$$\begin{aligned}a &= \mathbb{E}[f\hat{U}_{ij}|Q = q] = \exp[\theta_{qj}^*] \\ b &= \mathbb{E}[f\hat{U}_{ij}] = \sum w_q a \\ c &= \exp[(\sigma_j^*)^2] \\ d &= \mathrm{Var}[f\hat{U}_{ij}|Q = q] = a^2(c - 1) \\ \mathrm{Var}[\hat{U}_{ij}] &= c \sum \big(w_q(d + (a - b)^2)\big)\end{aligned} \tag{5.7}$$

With $U_{ij} = \hat{U}_{ij}/\rho_j$, the moments of $p(U_{ij})$ are given by:

$$\mathbb{E}[U_{ij}] = 1 \qquad\qquad \mathrm{Var}[U_{ij}] = \frac{\mathrm{Var}[\hat{U}_{ij}]}{\rho_j^2}$$

5.2.6 Sampling from $p(\mathbf{U}_l, \overline{U_l}|\overline{y_l}, \mu^F, \sigma^F, \mathbf{w}, \boldsymbol{\theta}, \sigma_l^u)$

To sample from $p(\mathbf{U}_l, \overline{U_l}|\overline{y_l}, \mu^F, \sigma^F, \mathbf{w}, \boldsymbol{\theta}, \sigma_l^u)$, we propose a Metropolis-Hastings sampler which uses approximations of $p(\overline{U_l}|\overline{y_l}, \mu^F, \sigma^F, \mathbf{w}, \boldsymbol{\theta}, \sigma_l^u)$ and $p(\mathbf{U}_l|\overline{U_l}, \mathbf{w}, \boldsymbol{\theta}, \sigma_l^u)$ as

the proposal distributions. In the next three sections, we will first derive these proposal distributions and then present the Metropolis-Hastings algorithm to sample from $p(\mathbf{U}, \overline{\mathbf{U}}|\overline{\mathbf{y}}, \mu^F, \sigma^F, \mathbf{w}, \boldsymbol{\theta}, \boldsymbol{\sigma}^u)$.

5.2.6.1 Step 1: Proposal distribution of $p(\overline{U_l}|\overline{y_l}, \mu^F, \sigma^F, \mathbf{w}, \boldsymbol{\theta}, \sigma_l^u)$

Here we consider the distribution of $\overline{U_l}$ and show how approximating $\overline{U_l}$ with a Lognormal distribution allows us to approximate the posterior distribution $p(\log(\xi_l^F)|\overline{y_l})$. Let us assume that:

$$\log(\xi_l^F) \sim \mathcal{N}(\mu^F, (\sigma^F)^2)$$
$$\log(\overline{U_l}) \dot{\sim} \mathcal{N}(\eta, \psi^2)$$

We know that $\log(\overline{y_l}) = \log(\xi_l^F) + \log(\overline{U_l})$. If $\log(\xi_l^F)$ is known, we have:

$$p(\log(\overline{y_l})|\log(\xi_l^F)) \approx \mathcal{N}(\log(\overline{y_l}); \log(\xi_l^F) + \eta, \psi^2)$$

We are actually interested in $p(\log(\xi_l^F)|\log(\overline{y_l}))$ and using Bayes rule we obtain:

$$\begin{aligned} p(\log(\xi_l^F)|\log(\overline{y_l})) &= \frac{p(\log(\xi_l^F), \log(\overline{y_l}))}{p(\log(\overline{y_l}))} \\ &\approx \mathcal{N}\left(\log(\xi_l^F); \frac{\mu^F\psi^2 + (\log(\overline{y_l}) - \eta)(\sigma^F)^2}{(\sigma^F)^2 + \psi^2}, \frac{(\sigma^F)^2\psi^2}{(\sigma^F)^2 + \psi^2}\right) \end{aligned} \quad (5.8)$$

We can now generate a sample from the Normal approximation of $p(\log(\xi_l^F)|\log(\overline{y_l}))$ which will provide a proposal value $\log(\xi_l^{F,*})$. As $\log(\overline{U_l^*}) = \log(\overline{y_l}) - \log(\xi_l^{F,*})$, we can use the proposed value for $\log(\overline{U_l^*})$ in a Metropolis-Hastings step to accept or reject $\log(\xi_l^{F,*})$.

Finding values for η and ψ

To approximate $\overline{U_l}$ with a $\mathcal{LN}(\eta, \psi^2)$ distribution, we need to find values for η and ψ. We propose to do this by matching the moments of this Lognormal distribution with the moments of the distribution of $\overline{U_l}$. As we do not know the distribution of $\overline{U_l}$, we use the moments of U_{kl} to derive the moments of $\overline{U_l}$.

Moments of $\overline{U_l}$

Using the moments of U_{kl}, the expected value of $\overline{U_l}$ is given by:

$$\mathbb{E}[\overline{U_l}] = 1 \quad (5.9)$$

For the variance, we know that if X_i are independent random variables:

$$\text{Var}\left[\,\overline{X}\,\right] = \text{Var}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2}\sum_{i=1}^{n}\text{Var}\left[X_i\right] = \frac{\text{Var}\left[X_i\right]}{n}$$

So:

$$\text{Var}[\overline{U_l}] = \frac{\text{Var}[U_{kl}]}{n_l} \tag{5.10}$$

Calculating parameters of the $\mathcal{LN}(\eta, \psi^2)$ distribution

The moments of the distribution of $\overline{U_l}$ can be matched to the moments of a Lognormal approximation of the distribution of $\overline{U_l}$ with parameters:

$$\psi^2 = \log\left(\frac{\text{Var}[\overline{U_l}]}{(\mathbb{E}[\overline{U_l}])^2} + 1\right)$$
$$\eta = \log(\mathbb{E}[\overline{U_l}]) - \frac{\psi^2}{2}$$

After observing composite residue level, $\overline{y_l}$, from supervised field trials, the approximate conditional posterior distribution that we use to propose values of ξ_l^F is given by Equation 5.8.

5.2.6.2 Step 2: Proposal distribution of $\mathbf{U}_l|\overline{U_l}, \mathbf{w}, \boldsymbol{\theta}, \sigma_l^u$: A constraint problem

The second step of the Metropolis-Hastings algorithm involves proposing values for $\mathbf{U}_l|\overline{U_l}$. Sampling n_l unit values from the mixture distribution of U_{kl} that have mean $\overline{U_l}$ can be achieved using a Gamma approximation.

Gamma approximation

Suppose that $U_{1l}, \ldots, U_{n_l l} \overset{\text{i.i.d.}}{\sim} \text{Gamma}\left(\frac{\zeta}{n_l}, \frac{\gamma}{n_l}\right)$. We need the probability density function of $n_l - 1$ samples when we know the mean of n_l samples, $p(U_{1l}, \ldots, U_{(n_l-1)l}|\overline{U_l})$. This is because if $n_l - 1$ samples and the mean are known, then the n_lth sample is also defined. We change variables from $U_{1l}, \ldots, U_{n_l l}$ to $f_{1l}, \ldots, f_{(n_l-1)l}, n_l\overline{U_l}$, where $f_{hl} = \frac{U_{hl}}{\sum_{k=1}^{n_l} U_{kl}}$ for $h = 1, \ldots, n_l - 1$ and $n_l\overline{U_l} = \sum_{k=1}^{n_l} U_{kl}$. This results in (see Section

2.2.1 for details):

$$p(f_{1l}, \ldots, f_{(n_l-1)l}|n_l\overline{U_l}) = \text{Dirichlet}\left(\frac{\zeta}{n_l}, \ldots, \frac{\zeta}{n_l}\right)$$

Now we can sample $\{f_{1l}, \ldots, f_{(n_l-1)l}\}$ which will provide $f_{n_l l} = 1 - \sum_{k=1}^{n_l-1} f_{kl}$. These are the proportions that each sampled residue value will contribute to $n_l\overline{U_l}$.
Next, assuming that $U_{1l}, \ldots, U_{n_l l} \overset{\text{i.i.d.}}{\sim} \text{Gamma}\left(\frac{\zeta}{n_l}, \frac{\gamma}{n_l}\right)$, we need to determine the distribution of $\overline{U_l}$. If we assume that all U_{kl} are independent, the moment-generating function for $\sum_{k=1}^{n_l} U_{kl}$ is:

$$\begin{aligned} M_{\sum U_{kl}}(t) =& E\left[\exp[t(U_{1l} + \cdots + U_{n_l l})]\right] \\ =& \left(\frac{\gamma^{\frac{\zeta}{n_l}}}{\left(\frac{\gamma}{n_l} - t\right)^{\frac{\zeta}{n_l}}}\right)^{n_l} = \frac{\left(\frac{\gamma}{n_l}\right)^{\zeta}}{\left(\frac{\gamma}{n_l} - t\right)^{\zeta}} \end{aligned}$$

which shows that $\sum_{k=1}^{n_l} U_{kl} \sim \text{Gamma}\left(\zeta, \frac{\gamma}{n_l}\right)$ and thus $\overline{U_l} \sim \text{Gamma}(\zeta, \gamma)$. Given the moments of the distribution of $\overline{U_l}$ from Equations 5.9 and 5.10 we can obtain the parameters ζ and γ by matching the moments of $\overline{U_l}$ with the moments of the Gamma distribution:

$$\begin{aligned} \zeta &= \frac{\mathbb{E}[\overline{U_l}]^2}{\text{Var}[\overline{U_l}]} \\ \gamma &= \frac{\mathbb{E}[\overline{U_l}]}{\text{Var}[\overline{U_l}]} \end{aligned}$$

This allows us to propose values from $p(\mathbf{U}_l|\overline{U_l})$ using:

$$\begin{aligned} f_{1l}, \ldots, f_{n_l l} &\sim \text{Dirichlet}\left(\frac{\zeta}{n_l}, \ldots, \frac{\zeta}{n_l}\right) \\ U_{kl} &= n_l\overline{U_l}f_{kl} \end{aligned}$$

5.2.6.3 Step 3: Metropolis-Hastings Algorithm to sample from $p(\mathbf{U}, \overline{\mathbf{U}}|\overline{\mathbf{y}}, \mathbf{w}, \boldsymbol{\theta}, \boldsymbol{\sigma}^u)$

In Section 5.2.6.1 we learned that if we approximate the distribution of $\overline{U_l}$ with a Lognormal distribution, we can propose values, $\overline{U_l}^*$, from an approximation of the posterior distribution $p(\overline{U_l}|\overline{y_l}, \mu^F, \sigma^F, \mathbf{w}, \boldsymbol{\theta}, \sigma_l^u)$. In Section 5.2.6.2 we observed that if we approximate the distribution of U_{kl} with a Gamma distribution, we can propose

values $\mathbf{U}_l^*$ from an approximation to $p(\mathbf{U}_l|\overline{U_l}, \mathbf{w}, \boldsymbol{\theta}, \sigma_l^u)$. We can use these proposed values in a Metropolis-Hastings step, ultimately leading to samples from our target distribution $p(\mathbf{U}_l, \overline{U_l}|\overline{y_l}, \mu^F, \sigma^F, \mathbf{w}, \boldsymbol{\theta}, \sigma_l^u)$. Let ξ_l^* be the proposed value at iteration t, sampled from proposal density $q(\cdot)$. The acceptance probability, p_{accept}, is defined as:

$$p_{accept} = \min\left(1, \frac{q(\mathbf{U}_l, \overline{U_l}|\overline{y_l}, \mu^F, \sigma^F, \mathbf{w}, \boldsymbol{\theta}, \sigma_l^u)}{q(\mathbf{U}_l^*, \overline{U_l^*}|\overline{y_l}, \mu^F, \sigma^F, \mathbf{w}, \boldsymbol{\theta}, \sigma_l^u)} \frac{p(\mathbf{U}_l^*|\overline{y_l}, \mu^F, \sigma^F, \mathbf{w}, \boldsymbol{\theta}, \sigma_l^u)}{p(\mathbf{U}_l|\overline{y_l}, \mu^F, \sigma^F, \mathbf{w}, \boldsymbol{\theta}, \sigma_l^u)}\right)$$

The first term, $\frac{q(\mathbf{U}_l, \overline{U_l}|\overline{y_l}, \mu^F, \sigma^F, \mathbf{w}, \boldsymbol{\theta}, \sigma_l^u)}{q(\mathbf{U}_l^*, \overline{U_l^*}|\overline{y_l}, \mu^F, \sigma^F, \mathbf{w}, \boldsymbol{\theta}, \sigma_l^u)}$ is referred to as the proposal ratio and the second term, $\frac{p(\mathbf{U}_l^*|\overline{y_l}, \mu^F, \sigma^F)}{p(\mathbf{U}_l|\overline{y_l}, \mu^F, \sigma^F, \mathbf{w}, \boldsymbol{\theta}, \sigma_l^u)}$, as the target ratio. We will discuss both ratios in more detail in the following sections.

Proposal Ratio

As we use a proposal distribution that is based on approximations of the posterior distributions $p(\overline{U_l}|\overline{y_l}, \mathbf{w}, \boldsymbol{\theta}, \sigma_l^u)$ and $p(\mathbf{U}_l|\overline{y_l}, \overline{U_l}, \mathbf{w}, \boldsymbol{\theta}, \sigma_l^u)$, the proposal ratio is given by:

$$\frac{q(\mathbf{U}_l, \overline{U_l}|\overline{y_l}, \mu^F, \sigma^F, \mathbf{w}, \boldsymbol{\theta}, \sigma_l^u)}{q(\mathbf{U}_l^*, \overline{U_l^*}|\overline{y_l}, \mu^F, \sigma^F, \mathbf{w}, \boldsymbol{\theta}, \sigma_l^u)} = \frac{p(\overline{U_l}|\overline{y_l}, \mu^F, \sigma^F, \mathbf{w}, \boldsymbol{\theta}, \sigma_l^u)}{p(\overline{U_l^*}|\overline{y_l}, \mu^F, \sigma^F, \mathbf{w}, \boldsymbol{\theta}, \sigma_l^u)} \frac{p(\mathbf{U}_l|\overline{U_l}, \mathbf{w}, \boldsymbol{\theta}, \sigma_l^u)}{p(\mathbf{U}_l^*|\overline{U_l^*}, \mathbf{w}, \boldsymbol{\theta}, \sigma_l^u)}$$

The first term, $p(\overline{U_l}|\overline{y_l}, \mathbf{w}, \boldsymbol{\theta}, \sigma_l^u)$, can be obtained using the approximation of the posterior distribution of $p(\log(\xi_l^F)|\log(\overline{y_l}))$. Changing variables from $\log(\xi_l^F)$ to $\overline{U_l}$ using $\log(\xi_l^F) = \log(\overline{y_l}) - \log(\overline{U_l})$ and $d\log(\xi_l^F) = \overline{U_l}^{-1} d\overline{U_l}$, results in:

$$p(\overline{U_l}|\overline{y_l}) = \frac{1}{\overline{U_l}} \mathcal{N}\left(\log(\overline{y_l}) - \log(\overline{U_l}); \mu_{post}, \sigma_{post}^2\right) \tag{5.11}$$

where $\mu_{post} = \frac{\mu^F \psi^2 + (\log(\overline{y_l}) - \eta)(\sigma^F)^2}{(\sigma^F)^2 + \psi^2}$ and $\sigma_{post}^2 = \frac{(\sigma^F)^2 \psi^2}{(\sigma^F)^2 + \psi^2}$. For the second term, $p(\mathbf{U}_l|\overline{U_l}, \mathbf{w}, \boldsymbol{\theta}, \sigma_l^u)$, we use a Gamma approximation to propose values $U_{1l}^*, \ldots, U_{n_l l}^*$ and apply a Metropolis-Hastings step to accept/reject them. The proposal distribution for $\mathbf{f}_l = \{f_{1l}, \ldots, f_{(n_l-1)l}\}$ is given by:

$$p\left(\mathbf{f}_l|\overline{U_l}, \frac{\zeta}{n_l}, \mathbf{w}, \boldsymbol{\theta}, \sigma_l^u\right) = \frac{1}{\mathrm{B}\left(\frac{\zeta}{n_l}\right)} \prod_{k=1}^{n_l} f_{kl}^{\frac{\zeta}{n_l}-1}$$

where $\mathrm{B}\left(\frac{\zeta}{n_l}\right) = \frac{\Gamma\left(\frac{\zeta}{n_l}\right)^{n_l}}{\Gamma(\zeta)}$. So let $f_{kl} = \frac{U_{kl}}{n_l \overline{U_l}}$, then $\frac{df_{kl}}{dU_{kl}} = \frac{1}{n_l \overline{U_l}}$, leading to:

$$p\left(\mathbf{U}_l|\overline{U_l}, \frac{\zeta}{n_l}, \mathbf{w}, \boldsymbol{\theta}, \sigma_l^u\right) \propto \frac{1}{\overline{U_l}^{\zeta-1}} \prod_{k=1}^{n_l} U_{kl}^{\frac{\zeta}{n_l}-1} \tag{5.12}$$

The proposal ratio can then be obtained from Equations 5.11 and 5.12:

$$\frac{p(\xi_l^F|\overline{y_l},\mathbf{w},\boldsymbol{\theta},\sigma_l^u)}{p((\xi_l^F)^*|\overline{y_l},\mathbf{w},\boldsymbol{\theta},\sigma_l^u)}\frac{p(U_{1l},\ldots,U_{n_l l})|\overline{y_l},\overline{U_l},\mathbf{w},\boldsymbol{\theta},\sigma_l^u)}{p(U_{1l}^*,\ldots,U_{n_l l)}^*|\overline{y_l},\overline{U_l^*},\mathbf{w},\boldsymbol{\theta},\sigma_l^u)}=\frac{\mathcal{N}\left(\log(\xi_l^F);\mu_{post},\sigma_{post}^2\right)}{\mathcal{N}\left(\log((\xi_l^F)^*);\mu_{post},\sigma_{post}^2\right)}\times \frac{\overline{U_l^*}^{\zeta}\prod_{k=1}^{n_l}U_{kl}^{\frac{\zeta}{n_l}-1}}{\overline{U_l}^{\zeta}\prod_{k=1}^{n_l}(U_{kl}^*)^{\frac{\zeta}{n_l}-1}}$$

Target Ratio

From Equation 5.5 (page 188), we know that we are interested in:

$$p(\mathbf{U}_l|\overline{y_l},\mu^F,\sigma^F,\mathbf{w},\boldsymbol{\theta},\sigma_l^u)\propto p(\overline{y_l}|\overline{U_l},\mu^F,\sigma^F)\prod_{k=1}^{n_l}p(U_{kl}|\mathbf{w},\boldsymbol{\theta},\sigma_l^u)$$

To be able to calculate $p(\overline{y_l}|\overline{U_l},\mu^F,\sigma^F)$, we make use of the logical dependencies, $\log(\overline{y_l})=\log(\xi_l^F)+\log(\overline{U_l})$ and $\overline{U_l}=\frac{U_{1l}+\ldots+U_{n_l l}}{n_l}$. Changing variables and using $\log(\xi_l^F)\sim\mathcal{N}(\mu^F,(\sigma^F)^2)$ leads to:

$$\begin{aligned}p(\overline{y_l}|\overline{U_l},\mu^F,\sigma^F)&=\frac{1}{\overline{U_l}}p_{\xi_l^F}\left(\frac{\overline{y_l}}{\overline{U_l}};\mu^F,\sigma^F\right)\\&=\frac{1}{\overline{y_l}}\mathcal{N}\left(\log\left(\frac{\overline{y_l}}{\overline{U_l}}\right);\mu^F,(\sigma^F)^2\right)\end{aligned}$$

The second term in the target distribution, $p(U_{kl}|\mathbf{w},\boldsymbol{\theta},\sigma_l^u,\sigma_c)$, can be obtained from the revised unit model (Equation 5.6 on page 191). As a result, the target distribution is:

$$\frac{p(\mathbf{U}_l^*|\overline{y_l},\mu^F,\sigma^F,\mathbf{w},\boldsymbol{\theta},\sigma_c,\sigma_l^u)}{p(\mathbf{U}_l|\overline{y_l},\mu^F,\sigma^F,\mathbf{w},\boldsymbol{\theta},\sigma_c,\sigma_l^u)}=\frac{\mathcal{N}\left(\log\left(\frac{\overline{y_l}}{\overline{U_l^*}}\right);\mu^F,(\sigma^F)^2\right)}{\mathcal{N}\left(\log\left(\frac{\overline{y_l}}{\overline{U_l}}\right);\mu^F,(\sigma^F)^2\right)}\times \frac{\prod_{k=1}^{n_l}U_{kl}\sum_{q=1}^{C}w_q\phi\left(\frac{\log(\rho_j U_{kl}^*)-\theta_q\sigma_l^u}{\sigma_c\sigma_l^u}\right)}{\prod_{k=1}^{n_l}U_{kl}^*\sum_{q=1}^{C}w_q\phi\left(\frac{\log(\rho_j U_{kl})-\theta_q\sigma_l^u}{\sigma_c\sigma_l^u}\right)}$$

5.2.7 Distributions of μ^F and σ^F for various prior distributions

Now we have samples of ξ_l^F, we want to obtain samples of the posterior distribution $p(\mu^F,\sigma^F|\boldsymbol{\xi}^{\mathbf{F}})$. If we assume $\log(\xi_l^F)\sim\mathcal{N}(\mu^F,(\sigma^F)^2)$, we have various options for the prior distribution, $\pi(\mu^F,\sigma^F)$. In this section, we explore three different prior distributions.

5.2.7.1 $\pi(\boldsymbol{\mu}^{\mathbf{F}}, \boldsymbol{\sigma}^{\mathbf{F}}) = 1/\boldsymbol{\sigma}^{\mathbf{F}}$

Using a $1/\sigma^F$ prior (Box and Tiao, 1973), updating the parameters μ^F and σ^F of the Lognormal distribution for ξ_l^F given the data is relatively simple. However, Gelman (2006) states that for hierarchical variance parameters in a one-way ANOVA setting, with group-level effect $\alpha_j \sim \mathcal{N}(0, \sigma_\alpha^2)$, a $1/\sigma_\alpha$ prior results in an improper posterior distribution. As the data cannot rule out a group-level variance of zero, a $1/\sigma_\alpha$ prior distribution that puts an infinite mass near zero will result in inferences that favour the absence of a group effect. For our model, when using a $1/\sigma^F$ prior distribution, the data, $\overline{y}_l$, cannot rule out that σ^F can be zero as the model will use the distribution of $\overline{U_l}$ to explain the observed variation in $\overline{y}_l$. Therefore, alternative prior distributions for σ^F should be considered.

5.2.7.2 Normal-Gamma prior

An alternative to the $1/\sigma^F$ prior distribution is a Normal-Gamma prior for (μ^F, τ^F) where $\tau^F = (1/sigma^F)^2$. This prior distribution allows available information on μ^F and σ^F to be incorporated into the model, e.g. by restricting σ^F to be more likely in a certain range. The Normal-Gamma distribution is a 4 parameter distribution, defined as:

$$\begin{aligned}\pi(\mu^F, \tau^F|\mu_0, \kappa_0, \alpha_0, \beta_0) &= \mathcal{N}(\log(\xi^F); \mu_0, (\kappa_0\tau^F)^{-1})\text{Gamma}(\tau^F; \alpha_0, \beta_0) \\ &\propto (\tau^F)^{\frac{1}{2}} \exp\left[-\frac{\kappa_0\tau^F}{2}(\log(\xi^F) - \mu_0)^2\right](\tau^F)^{\alpha_0-1}\exp\left[-\tau^F\beta_0\right]\end{aligned}$$

A sequential sampler can be used to sample τ^F from a Gamma distribution and $\mu^F|\tau^F$ from a Normal distribution:

$$\begin{aligned}\tau^F|\log(\boldsymbol{\xi}^{\mathbf{F}}) \sim \text{Gamma}\Bigg(&\frac{L}{2} + \alpha_0, \beta_0 + \frac{L\kappa_0(\overline{\log(\xi^F)} - \mu_0)^2}{2(L+\kappa_0)} \\ &+\frac{1}{2}\sum_{l=1}^{L}\left(\log(\xi_l^F) - \overline{\log(\xi^F)}\right)^2\Bigg)\end{aligned}$$

$$\mu^F|\tau^F, \log(\boldsymbol{\xi}^{\mathbf{F}}) \sim \mathcal{N}\left(\frac{\kappa_0\mu_0 + L\overline{\log(\xi^F)}}{\kappa_0 + L}, \frac{1}{(L+\kappa_0)\lambda^F}\right)$$

where $\overline{\log(\xi^F)} = \frac{\sum_{l=1}^{L}\log(\xi_l^F)}{L}$.

5.2.7.3 Uniform Prior

A non-informative alternative to the $1/\sigma^F$ prior distribution is to use a Uniform prior distribution for μ^F and σ^F. The conditional distribution of μ^F can be obtained by treating σ^F as a constant and factorising:

$$p(\mu^F|\sigma^F, \log(\boldsymbol{\xi}^{\mathbf{F}})) \propto \exp\left[\frac{L}{2(\sigma^F)^2}\left(\mu^F - \overline{\log(\xi^F)}\right)^2\right]$$

in which we recognise a $\mathcal{N}\left(\overline{\log(\xi^F)}, \frac{(\sigma^F)^2}{L}\right)$ distribution. The marginal distribution of $\tau^F = (\sigma^F)^{-2}$ can now be obtained using:

$$p\left(\tau^F|\log(\boldsymbol{\xi}^{\mathbf{F}})\right) \propto (\tau^F)^{\frac{L-2}{2}-1}\exp\left[-\frac{\tau^F(L-1)s^2}{2}\right]$$

where s^2 is the sample variance of $\log(\boldsymbol{\xi}^{\mathbf{F}})$. In this we recognise that $\tau^F(L-1)s^2|\log(\boldsymbol{\xi}^{\mathbf{F}}) \sim \chi^2_{L-2}$, i.e. a χ^2-distribution with $L-2$ degrees of freedom.

5.2.7.4 Choice of prior distribution for within-field and between-field model

If no information is available on σ^F, the recommended prior distribution for σ^F is the Uniform distribution (Gelman, 2006) as it does not favour small values of σ^F and does not require the specification of many parameters. However, if information was available that would allow us to specify an informative prior distribution on σ^F, it might be useful to select a Normal-Gamma prior distribution for τ^F. Although the Normal-Gamma distribution may provide more flexibility than the other two prior distributions, it requires 4 parameters to be specified. A weakly informative non-conjugate alternative would be the half-Cauchy distribution (Gelman, 2006). As we do not have prior information for σ^F we use the Uniform prior distribution for the model runs illustrated in this chapter.

5.2.8 Hierarchical model for the scale parameter of the unit model

The refined unit model (see Figure 5.1) allows us to infer the shape and scale parameters for the unit log-residue distribution for a new pesticide for which unit data are

not available. We considered a hierarchical Gamma model for the scale parameter σ_j^u and for the precision $\tau_j^u = (\sigma_j^u)^{-2}$ of log-residue data obtained from field trial j. However, in tests we found that the model worked better when parameterised with the scale parameter so we only discuss that model here. Figure 5.3 shows the results of an exploratory data analysis of the unit field trial data. We calculated the scale parameter, σ_j^u, defined as half the intertertile range, for each of the 75 available field trial data sets and fitted a Gamma distribution. The results indicate that the model provides a reasonable fit, but may predict more extreme values for σ_j^u than we observed in the data.

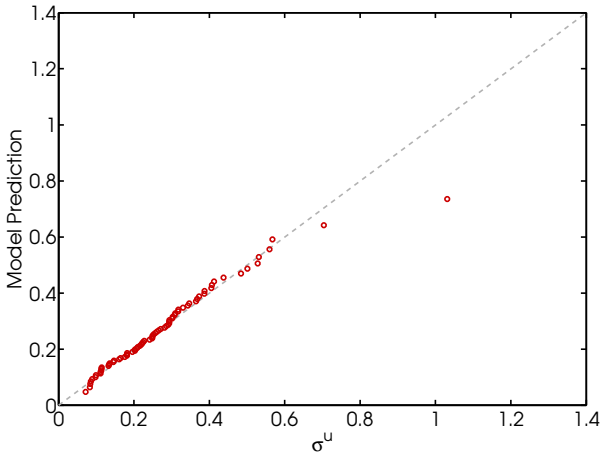


Figure 5.3 – *QQ Plot of scale parameter $\hat{\sigma}_j^u$. Empirical estimates are plotted against quantiles based on a maximum likelihood estimate of the parameters of a Gamma distribution.*

5.2.8.1 Model Description

Metropolis-Hastings algorithm for μ_j^u and σ_j^u

Using a hierarchical model for $\boldsymbol{\sigma}^u$ affects the posterior distribution of the location and scale parameters, μ_j and σ_j respectively, of the unit model. Therefore in this section we discuss the new Metropolis-Hastings algorithm that is necessary to sample these parameters.

Proposal ratio

For the proposal ratio we use the previously used proposal distributions (see Section 4.2.2 for details) leading to:

$$\frac{q(\mu_j^u, \sigma_j^u; (\mu_j^u)^*, (\sigma_j^u)^*)}{q((\mu_j^u)^*, (\sigma_j^u)^*; \mu_j^u, \sigma_j^u)} = \left(\frac{(\sigma_j^u)^*}{\sigma_j^u}\right)^2$$

Target ratio

Let us replace the $1/\sigma_j^u$ prior from the unit model in Chapter 4 with a Gamma(α, β) prior distribution for the scale parameter $\boldsymbol{\sigma}^u = \{\sigma_1^u, \ldots, \sigma_J^u\}$ where J is the number of unit data sets as before:

$$\pi(\mu_j^u, \sigma_j^u | \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} (\sigma_j^u)^{\alpha-1} \exp[-\beta\sigma_j^u]$$

This leads to:

$$p(\boldsymbol{\mu}, \boldsymbol{\sigma} | \alpha, \beta) = \frac{\beta^{J\alpha}}{\Gamma(\alpha)^J} \exp\left[-\beta \sum_{j=1}^{J} \sigma_j^u\right] \prod_{j=1}^{J} (\sigma_j^u)^{\alpha-1}$$

We know that the shape distribution for normalised log-residue data, $\log(\tilde{\mathbf{x}}_{ij})$, is given by:

$$p_{\log(\tilde{x})}(\log(\tilde{x}_{ij}) | \boldsymbol{\theta}, \mathbf{w}, \sigma_c) = \sum_{q=1}^{C} \frac{w_q}{\sigma_c} \phi\left(\frac{\log(\tilde{x}_{ij}) - \theta_q}{\sigma_c}\right)$$

where C is the number of components in the mixture and ϕ is the standard Normal density. Given α, β and the shape distribution, the $\{\mu_j^u, \sigma_j^u\}$ sets are independent, so:

$$\begin{aligned} p(\mu_j^u, \sigma_j^u | \mathbf{x}, \boldsymbol{\theta}, \mathbf{w}, \alpha, \beta) &\propto p(\mu_j^u, \sigma_j^u | \alpha, \beta) p(x | \boldsymbol{\theta}, \mathbf{w}, \mu_j^u, \sigma_j^u) \\ &\propto \exp[-\beta\sigma_j^u] (\sigma_j^u)^{\alpha-1-n_j} \prod_{i=1}^{n_j} p_{\log(\tilde{x})}(\tilde{x}_{ij}) \end{aligned}$$

where $\mathbf{x}$ are the unit log-residue data from field trials. The target ratio for each $\{\mu_j^u, \sigma_j^u\}$ pair is:

$$\begin{aligned} \frac{p((\mu_j^u)^*, (\sigma_j^u)^* | \mathbf{x}, \boldsymbol{\theta}, \mathbf{w}, \alpha, \beta)}{p(\mu_j^u, \sigma_j^u | \mathbf{x}, \boldsymbol{\theta}, \mathbf{w}, \alpha, \beta)} &= \frac{\exp[-\beta(\sigma_j^u)^*]((\sigma_j^u)^*)^{\alpha-1-n_j} \prod_{i=1}^{n_j} p_{\log(\tilde{x})}(\tilde{x}_{ij}^*)}{\exp[-\beta(\sigma_j^u)](\sigma_j^u)^{\alpha-1-n_j} \prod_{i=1}^{n_j} p_{\log(\tilde{x})}(\tilde{x}_{ij})} \\ &= \exp[-\beta((\sigma_j^u)^* - \sigma_j^u)] \left(\frac{(\sigma_j^u)^*}{\sigma_j^u}\right)^{\alpha-1-n_j} \prod_{i=1}^{n_j} \frac{p_{\log(\tilde{x})}(\tilde{x}_{ij}^*)}{p_{\log(\tilde{x})}(\tilde{x}_{ij})} \end{aligned}$$

Distributions of α and β

The next step is to update the parameters α and β of the Gamma distribution. Let us assume $v \overset{\text{i.i.d.}}{\sim} \text{Gamma}(\alpha, \beta)$ and use the following Jeffreys prior for α and β (Miller, 1980):

$$\pi(\alpha, \beta) = \frac{1}{\alpha\beta}$$

Therefore the posterior distribution for α and β is:

$$p(\alpha, \beta|\mathbf{v}) \propto \frac{1}{\alpha\beta} \frac{\beta^{J\alpha}}{\Gamma(\alpha)^J} \exp\left[-\beta \sum_{j=1}^{J} v_j\right] \prod_{j=1}^{J} v_j^{\alpha-1}$$

where $\mathbf{v} = \{v_1, \ldots, v_J\}$. From this, we can obtain:

$$p(\beta|\alpha, \mathbf{v}) \propto \beta^{J\alpha-1} \exp\left[-\beta \sum_{j=1}^{J} v_j\right]$$

So:

$$p(\beta|\alpha, \mathbf{v}) = \text{Gamma}\left(\beta; J\alpha, \sum_{j=1}^{J} v_j\right)$$

We know α has a marginal distribution given by:

$$\begin{aligned} p(\alpha|\mathbf{v}) &= \int_0^\infty p(\alpha, \beta|\mathbf{v}) d\beta \\ &\propto \int_0^\infty \frac{\beta^{J\alpha-1}}{\alpha\Gamma(\alpha)^J} \exp\left[-\beta \sum_{j=1}^{J} v_j\right] \prod_{j=1}^{J} v_j^{\alpha-1} d\beta \end{aligned}$$

This leads to:

$$p(\alpha|\mathbf{v}) \propto \frac{\Gamma(J\alpha)}{\left(\sum_{j=1}^{J} v_j\right)^{J\alpha}} \frac{\prod_{j=1}^{J} v_j^{\alpha-1}}{\alpha\Gamma(\alpha)^J} \tag{5.13}$$

Since we cannot sample easily from $p(\alpha|\mathbf{v})$, we use a Metropolis-Hastings step.

Proposal ratio

We use the following proposal density:

$$q(\alpha^*|\alpha) \sim \mathcal{N}(\alpha, W(\alpha)^2)$$

where $W(\alpha)$ is the Wald Standard Error for α. The log-likelihood for J observations $\mathbf{v}$ from a Gamma(α, β) distribution is given by:

$$\log(L(\alpha, \beta)) = J\alpha \log(\beta) - J \log\left(\Gamma(\alpha)\right) + (\alpha - 1) \sum_{j=1}^{J} \log(v_j) - \beta \sum_{j=1}^{J} v_j$$

Maximising with respect to β results in $\hat{\beta} = \frac{\alpha}{\bar{v}}$. Substitution in the log-likelihood function leads to the profile log-likelihood of α:

$$\log(L(\alpha)) = J\alpha\log(\alpha) - J\alpha\log(\bar{v}) - J\log\left(\Gamma(\alpha)\right) - J\alpha + (\alpha - 1)\sum_{j=1}^{J}\log(v_j)$$

Now we calculate the Fisher information $\mathcal{I}$ for α:

$$\begin{aligned}\mathcal{I}(\alpha) &= -E\left[\frac{\partial^2}{\partial\alpha^2}\log(L(\alpha))\middle|\,\alpha\right] \\ &= J\Psi'(\alpha) - \frac{J}{\alpha} = \frac{J(\alpha\Psi'(\alpha) - 1)}{\alpha}\end{aligned}$$

where $\Psi'(z)$ is the trigamma function $\Psi'(z) = \frac{d^2}{dz^2}\log(\Gamma(z))$. The Wald Standard Error is given by:

$$W(\alpha) = \frac{1}{\sqrt{\mathcal{I}(\alpha)}} = \sqrt{\frac{\alpha}{J(\alpha\Psi'(\alpha) - 1)}}$$

Although the maximum likelihood estimator, $\hat{\alpha}$, is normally used in this expression, we instead use $W(\alpha)$ as a quick approximation in a Metropolis-Hastings step. The proposal ratio is:

$$\frac{q(\alpha|\alpha^*)}{q(\alpha^*|\alpha)} = \frac{W(\alpha)}{W(\alpha^*)}\exp\left[\frac{(\alpha^* - \alpha)^2}{2}\frac{W(\alpha^*)^2 - W(\alpha)^2}{W(\alpha^*)^2 W(\alpha)^2}\right]$$

Target ratio

The target distribution is given by Equation 5.13 leading to the target ratio:

$$\frac{p_{\alpha|\mathbf{v}}(\alpha^*)}{p_{\alpha|\mathbf{v}}(\alpha)} = \left(\sum_{j=1}^{J} v_j\right)^{J(\alpha-\alpha^*)}\frac{\Gamma(J\alpha^*)}{\Gamma(J\alpha)}\frac{\alpha\Gamma(\alpha)^J}{\alpha^*\Gamma(\alpha^*)^J}\left(\prod_{j=1}^{J} v_j\right)^{\alpha^*-\alpha}$$

Now that we have defined a hierarchical model for the unit variation of log-residue data sets, we can use a Metropolis-Hastings within Gibbs algorithm to sample μ_j^u, σ_j^u, α and β.

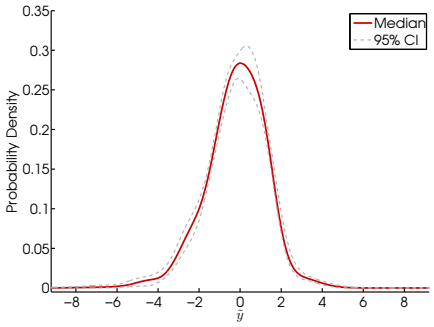
5.2.8.2 Results of applying the hierarchical unit model to field trial data

In this section we show the results from applying the hierarchical unit model to the unit log-residue data from field trials. We considered two different setups: the first model run used a fixed value of $\gamma = 10$ (see Chapter 4 for details about γ), whereas

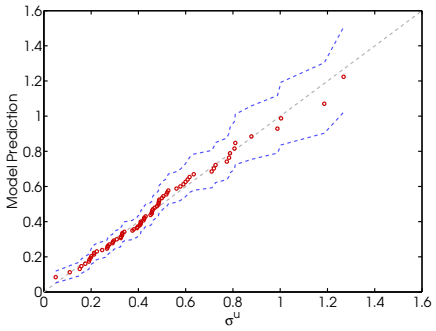
the second run aimed to learn γ from the data. As there was little difference between these two cases we only show the results using uncertain γ here and we use these in the model runs later in this chapter. Figure 5.4 shows the results from running the hierarchical unit model on unit log-residue field trial data. Figure 5.4a shows

Figure 5.4 – *Results from running the hierarchical model applied to unit log-residue data from field trials with uncertain γ.*

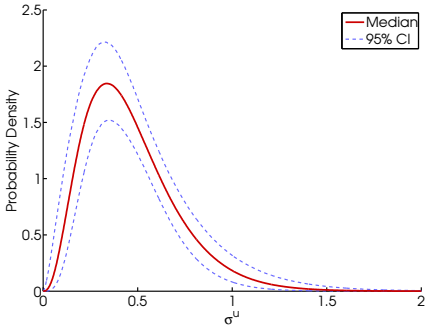
(a) Shape distribution



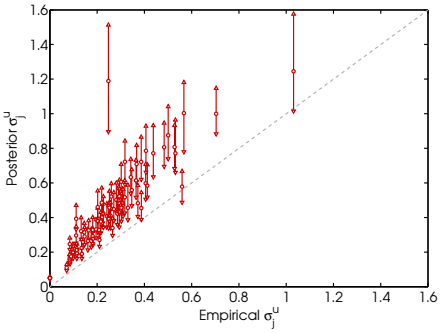
(b) QQ plot



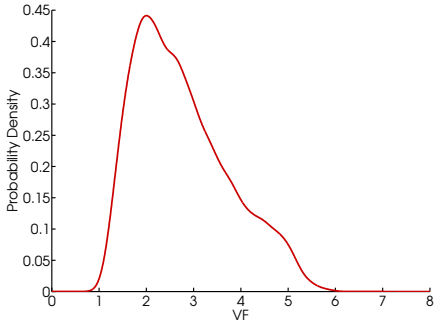
(c) Posterior Gamma distribution for $\boldsymbol{\sigma}^u$



(d) Empirical versus posterior distribution of $\boldsymbol{\sigma}^u$



(e) Variation in Variability Factor



the unit shape distribution which is similar to the shape obtained in Chapter 4. Figure 5.4b shows the QQ plot for the posterior mean for $\boldsymbol{\sigma}^u$ for each of the 75 data sets versus the median and 95% credible interval of the predictions based on the Gamma distribution. This suggests that the Gamma distribution provides an adequate fit to the scale parameters, $\boldsymbol{\sigma}^u$. Figure 5.4c shows the posterior Gamma distribution of the scale parameters $\boldsymbol{\sigma}^u$. It is clear that for a new use for which no unit data are available, predicting the scale of the unit log-residue distribution is going to be very uncertain as the variation in the scale parameter is relatively large with expected values of the 2.5^{th} and 97.5^{th} percentiles being 0.12 and 1.06 respectively. Figure 5.4d shows the empirical scale parameter, $\boldsymbol{\sigma}^u$, defined as half the intertertile range of the unit data, plotted against the median and 95% credible interval of the posterior distributions of $\boldsymbol{\sigma}^u$. The results indicate that the posterior distributions tend to be higher than the empirical estimates. The reason for this is that the shape model does not restrict the tertiles of the shape distribution to be at -1 and 1 due to probability leaching (see Chapter 4). In our application half the intertertile range is approximately 0.7 (rather than 1) for the normalised log-residue unit data leading to a narrower shape distribution. Therefore σ_j^u needs to be larger to match the scale of the shape distribution to the unit data. Figure 5.4e shows the predictive distribution of the variability factor based on the revised unit model. Samples were generated from the shape and scale distributions to obtain realisations of the unit distribution. Subsequently, the ratio of the 97.5^{th} percentile and mean was calculated for each realisation. The results indicate that based on the hierarchical, common-shape distribution, the median estimate of the variability factor is 2.6 with a 95% credible interval ranging from 1.3 to 5.0, which is in line with the VF estimated in EFSA (2005).

5.3 Summary of MCMC Algorithm

Let us define:

$\overline{y_l}$	Composite residue level measurement from supervised field trial l.
n_l	Number of units that are used in a composite sample.
U_{kl}	Random quantity describing variation in unit data $1, \ldots, n_l$ from field l, defined by the scale and shape parameters of the unit data distribution with $\mathbb{E}[U_{kl}] = 1$.
$\overline{\mathrm{U_l}}$	Random quantity describing variation in composite samples around the field mean.
ξ_l^F	Mean residue in field l.
L	Number of fields from which composite samples, $\overline{y_l}$, were collected.
$\mathcal{N}(\mu^F, \sigma^F)$	Normal distribution with parameters μ^F and σ^F describing the variation in log field means, $\log(\xi_l^F)$.
x_{ij}	Residue on unit i from field j.

The inference steps can be summarised using the following algorithm:

1. Sample a unit distribution $(\mathbf{w}, \boldsymbol{\theta}, \alpha, \beta)$ from our posterior distributions for U_{kl} obtained from the hierarchical unit model.

2. Sample a scale parameter σ_l^u for each field for which we have a supervised trial composite value $\overline{y_l}$: $\sigma_l^u \overset{\text{i.i.d.}}{\sim} \text{Gamma}(\alpha, \beta)$.

3. Rescale the unit distribution using parameter $\rho_j = \sum_{q=1}^{C} w_q \exp\left[\theta_{qj}^* + \frac{(\sigma_j^*)^2}{2}\right]$ where $\theta_{qj}^* = \theta_q \sigma_l^u$ and $\sigma_j^* = \sigma_l^u \sigma_c$.

4. Calculate moments of the rescaled unit distribution and hence those of $\overline{U_l}$: $\mathbb{E}[\overline{U_l}] = 1$ and $\text{Var}[\overline{U_l}] = \frac{\text{Var}[\hat{U}_{kl}]}{n_l \rho_j^2}$, where $\text{Var}[\hat{U}_{kl}]$ is given by Equation 5.7 (page 192).

5. Use a Metropolis-Hastings step to sample values for ξ_l^F:

   (a) Propose $\overline{U_l}|\overline{y_l}$:

i. Calculate parameters η, ψ^2 of the $\mathcal{LN}(\eta, \psi^2)$ distribution used to approximate $p(\overline{U_l})$.

$$\eta = \log(\mathbb{E}[\overline{U_l}]) - \frac{1}{2}\log\left(1 + \frac{\text{Var}[\overline{U_l}]}{\mathbb{E}[\overline{U_l}]^2}\right)$$
$$\psi^2 = \log\left(1 + \frac{\text{Var}[\overline{U_l}]}{\mathbb{E}[\overline{U_l}]^2}\right).$$

ii. Propose a new value for $(\xi_l^F)^*$ (see Section 5.2.6.1):

$$\log((\xi_l^F)^*)|\log(\overline{y_l}) \sim \mathcal{N}\left(\frac{\mu^F\psi^2 + (\log(\overline{y_l}) - \eta)(\sigma^F)^2}{(\sigma^F)^2 + \psi^2}, \frac{(\sigma^F)^2\psi^2}{(\sigma^F)^2 + \psi^2}\right)$$

iii. Calculate $\overline{U_l}^* = \frac{\overline{y_l}}{(\xi_l^F)^*}$.

(b) Propose $\mathbf{U}_l|\overline{U_l}^*$:

i. Calculate the parameters of the Gamma(ζ, γ) distribution used to approximate $p(\overline{U_l})$ (see Section 5.2.6.2):

$$\zeta = \frac{\mathbb{E}[\overline{U_l}]^2}{\text{Var}[\overline{U_l}]}$$
$$\gamma = \frac{\mathbb{E}[\overline{U_l}]}{\text{Var}[\overline{U_l}]}$$

ii. Propose $U_{1l}^*, ..., U_{n_l l}^*$:

$$f_{1l}, \ldots, f_{n_l l} \sim \text{Dirichlet}\left(\frac{\zeta}{n_l}, \ldots, \frac{\zeta}{n_l}\right)$$
$$U_{kl}^* = n_l\overline{U_l}f_{kl}$$

(c) Sample $u \sim \text{Uniform}(0, 1)$.

(d) Accept or reject the proposed $\left(\xi_l^F\right)^*$, $U_{1l}^*, \ldots, U_{n_l l}^*$ if $u \leq p_{accept}$ where p_{accept} is as described in Section 5.2.6.3.

6. Sample μ^F and σ^F assuming a Uniform prior distribution:

$$p(\sigma^F|\log(\boldsymbol{\xi}^{\mathbf{F}})) = \sqrt{\frac{(L-1)s^2}{\chi^2_{L-2}}}$$
$$p(\mu^F|\sigma^F, \log(\boldsymbol{\xi}^{\mathbf{F}})) \sim \mathcal{N}\left(\overline{\log(\xi^F)}, \frac{(\sigma^F)^2}{L}\right)$$

where s is the standard deviation of the $\log(\xi_l^F)$, $\overline{\log(\xi^F)} = \frac{\sum_{l=1}^{L}\log(\xi_l^F)}{L}$ and χ^2_{L-2} is a χ^2 distribution with $L-2$ degrees of freedom.

7. Store μ^F, σ^F, $\boldsymbol{\sigma}^u$, $\mathbf{w}$, $\boldsymbol{\theta}$, α and β.

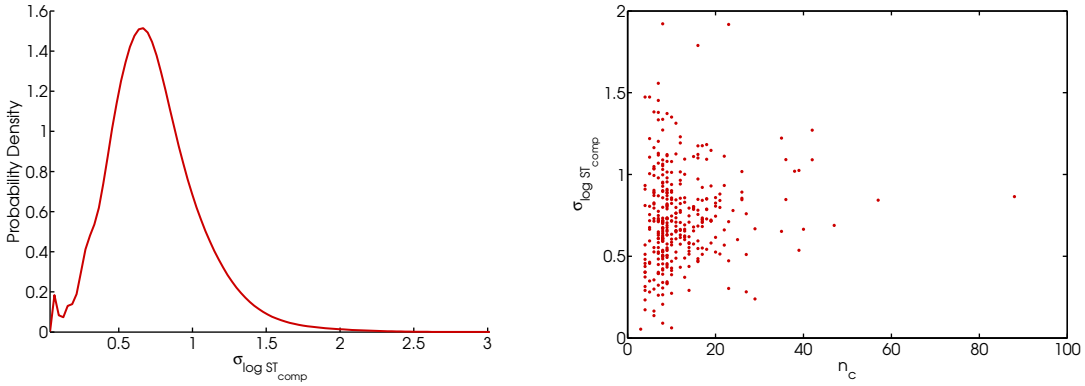
5.4 Validation Studies

To assess the performance of our new model we conduct a series of validation exercises. As we do not have information about the distribution of field means, the validation exercises had to be based on synthetic examples. To inform us about typical values to use for these examples, we analysed a large number of composite supervised trial data with no values below the limit of determination. For each supervised trial with five or more values we calculated the standard deviation of the log-transformed values, resulting in 345 standard deviations. The distribution of standard deviations is given in Figure 5.5.

Figure 5.5 – *Distribution of the standard deviations of log composite residue levels observed in 345 supervised field trials with five or more values.*

(a) Kernel Density Plot

(b) Standard deviation for each of the data sets plotted against the number of composite samples, n_c.



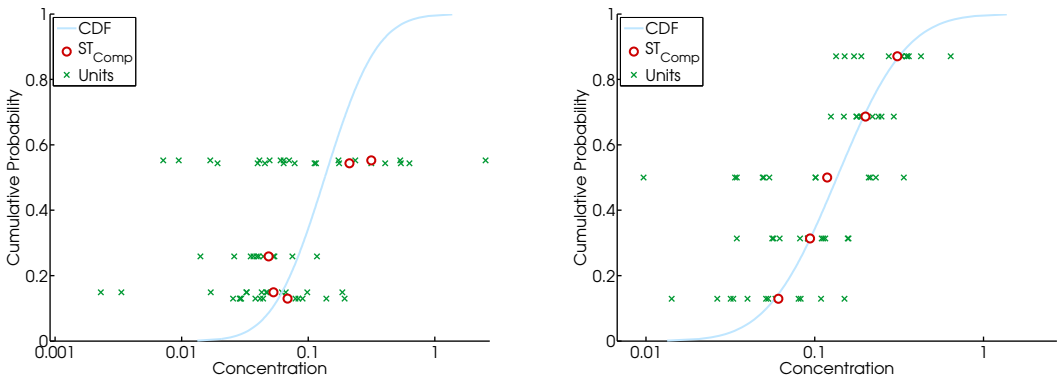
Next we calculated the 2.5^{th}, 50^{th} and 97.5^{th} percentiles of this distribution (0.24, 0.71 and 1.4 respectively). Although the variation in composite supervised trial data includes unit variation, we decided to use similar values as examples of small (0.25), typical (0.75) and large (1.5) values for σ^F in our validation studies.

Next we created samples from a known target distribution of field means to determine whether the model was capable of retrieving this distribution. To do this

we first sampled L log field means from the log field mean distribution $\xi_l^F \sim \mathcal{N}(-2, (\sigma^F)^2)$. From each of the L fields, we then generated n_l units from which we calculated a composite supervised trial value for each field. For a wide range of crops (including pome fruit, citrus fruit, tropical fruit with inedible skin, root and tuber vegetables, bulb vegetables, stem vegetables and brassicas), $n_l = 12$ is used in regulatory practice, so we used $n_l = 12$ in the first validation study (Section 5.4.1). We also explored a range of values for n_l to assess whether 12 units is sufficient to obtain a reliable estimate of the field mean (Section 5.4.2).

Figure 5.6 – *Two sampling approaches used to create validation data sets for $L = 5$ and $n_l = 12$. The true field mean distribution is shown in blue. The 12 unit residue values are indicated by green crosses and the 5 composite samples, based on the 12 units, are red circles.*

(a) Random Sample **(b)** Stratified Sample



We implemented two sampling approaches to generate the field means and units. The first approach generated both field means and units at random. However, for small L this may result in a poor representation of the field mean distribution (blue line) as shown in Figure 5.6a where no field means were sampled from the upper third of the distribution. The composite samples (red circles) do not fall on the field mean distribution so they do not provide good estimates of the field mean. Therefore, we also applied a stratified sampling approach to obtain the field means and unit residues (Figure 5.6b). Another reason for using this alternative approach

is that as there are generally very few composite samples available for a new use of a pesticide we would have to repeat all validation studies several times to assess the performance of the model whilst accounting for sampling variation. As the alternative approaches that we compare our model with would also struggle with 'poor' samples, we instead show how well the model performs using both random and stratified data sets. Each study was run in Matlab 2012a on a computer with an Intel i7-860 2.80 Ghz processor and 8GB RAM and took between 10 minutes and 9 hours depending on the scenario (number of composite samples and number of units per composite sample).

5.4.1 Validation Study 1: Multiple runs with typical between-field variation

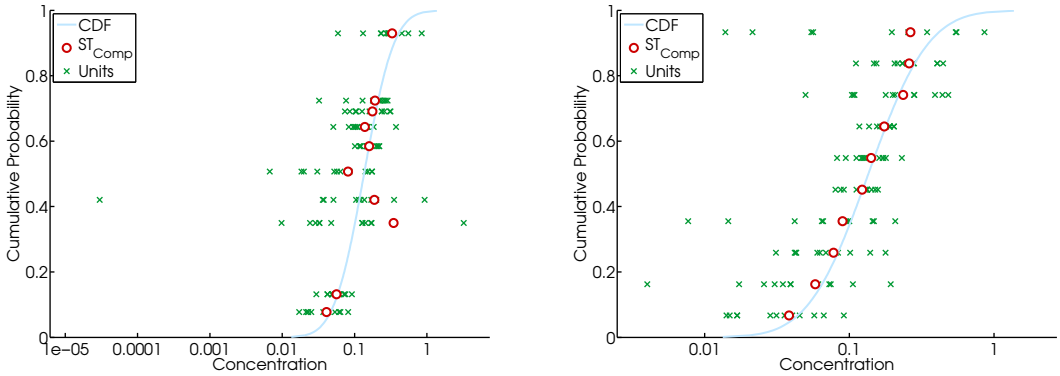
The first validation exercise consists of applying the model to a data set generated using a 'typical' value for σ^F, i.e. $\sigma^F = 0.75$. Data sets were generated using both the random and stratified sampling approaches to simulate 10 composite samples from supervised field trials based on 12 units per composite sample. The model output in Figure 5.7 shows that the field mean distribution inferred by the model is close to the true distribution, indicating that the model performs well for a typical data set.

Next, we compare the outcome of the models with existing approaches recommended by EFSA (2012). Figure 5.8 shows the posterior predictive distribution of unit log-residue levels, assuming that all units are obtained from separate fields, compared with existing approaches. These results indicate that our model is much better at describing the true distribution of within-field and between-field variation than the alternative approaches. The 'EFSA - Optimistic' model is based on bootstrapping the composite samples and does not explicitly model unit variation. This model provides a poor estimate of the distribution tails. The 'EFSA - Pessimistic' model assumes a Lognormal distribution for the composite samples and uses a Lognormal variability factor approach to describe unit variation. Using a variability factor of 6.82, the pessimistic EFSA model overestimates extreme residue levels (longer upper tail) and underestimates the main body of the unit residue distribution. EFSA

Figure 5.7 – *The top two panes show the sample obtained from the target field mean distribution (blue) with the 12 unit residue values indicated by green crosses and the 10 composite samples, based on the 12 units, as red circles. The bottom two panes show the output of running the within-field and between-field model on the random (left pane) and stratified data (right pane) with the target distribution in blue. The red lines represent the median estimate of the posterior field mean distribution and the grey dashed lines represent the 95% credible intervals.*

(a) Random data set

(b) Stratified data set



(c) Posterior field mean distribution - Random data set

(d) Posterior field mean distribution - Stratified data set

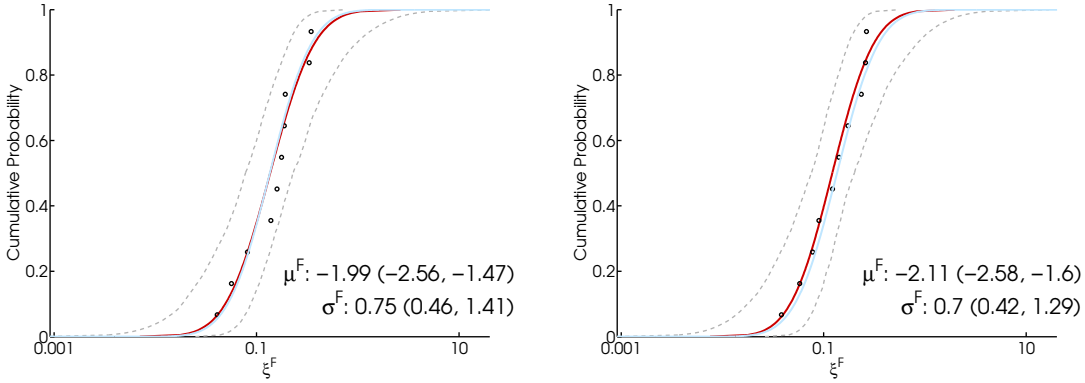
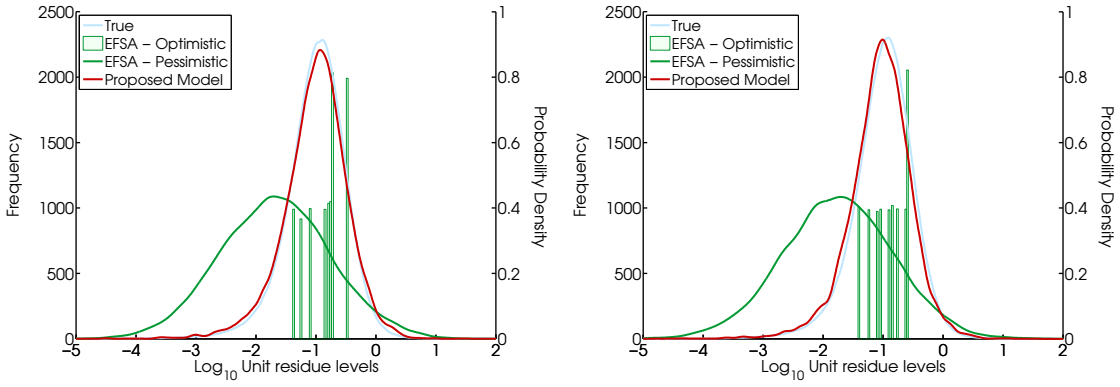


Figure 5.8 – *Comparison of the posterior predictive distributions obtained from applying the new within-field and between-field model (red) with the currently recommended approaches (EFSA, 2012) in green on the random and stratified samples. The blue line indicates the true unit distribution.*

(a) Random Sample **(b)** Stratified Sample



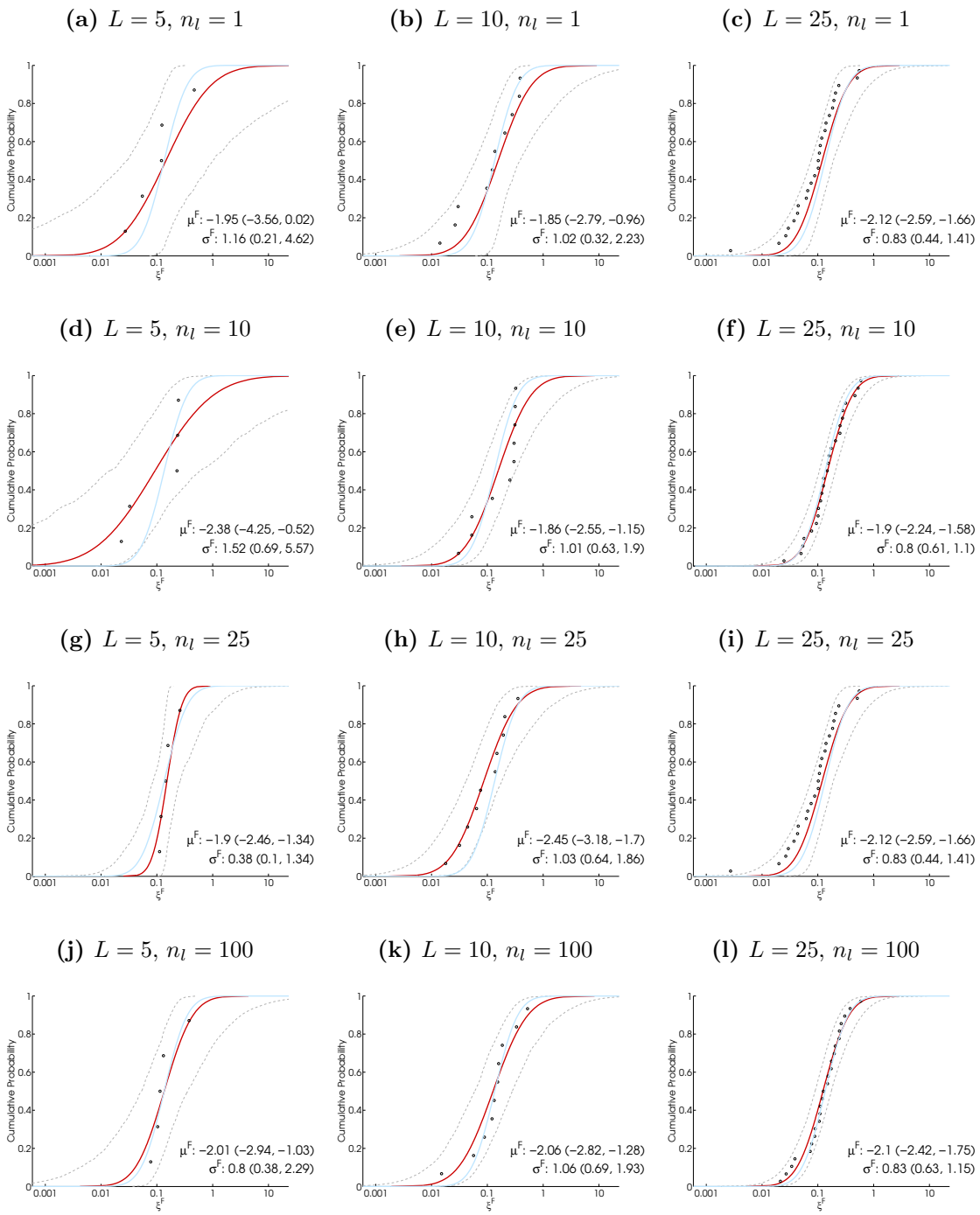
(2012) state that a variability factor of 6.82 'generates an excessive proportion of very high residues', so one could argue that the comparison presented here may be inappropriate. However, EFSA (2012) does not provide an alternative suggestion for the VF. Moreover, using any other value for the VF will result in two Lognormal distributions (because $\log(VF) = k\sigma - \sigma^2/2$ where $k = \Phi^{-1}(0.975) \approx 1.96$ and $\sigma > 0$ has two solutions for $1 < VF < \exp(k^2/2)$, a single solution for $0 < VF < 1$ and $VF = \exp(k^2/2)$ and no solutions for $VF > \exp(k^2/2)$, where $\exp(k^2/2) \approx 6.82$). Even though it is common to select the solution with smaller σ (Kennedy and Hart, 2009), no justification is given. Therefore we decided to use a variability factor of 6.82 in these simulations as it is the only option mentioned in the EFSA (2012) guidance document.

5.4.2 Validation Study 2: Effect of sample size with typical between-field variation

We ran our new model on data sets with varying numbers of simulated field composite samples ($L \in \{5, 10, 25\}$) and varying numbers of units per composite sample ($n_l \in \{1, 10, 25, 100\}$) with $\sigma^F = 0.75$. Figure 5.9 shows the field mean distribution

as inferred by the model using a random sample from the target distribution.

Figure 5.9 – *Effect of sample size using a random sample (black circles) from the target distribution (blue line). The red line indicates the median field mean distribution estimate with a 95% credible interval indicated by grey dashed lines.*



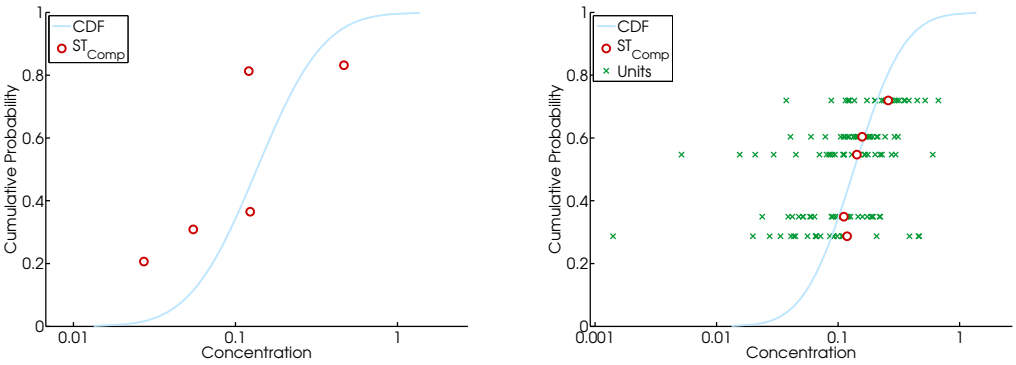
It is clear that the model does well for most sample sizes but for small sample sizes it has a tendency to assign the observed variation in log composite-residue values to between-field variation, thereby overestimating the between-field variation. Gelman (2006) stated that for small sample sizes ($L = 4$ or 5), the heavy right tail of the Uniform prior is likely to result in an overestimate of σ^F. Therefore one could consider using a more appropriate prior distribution for small L. For example, if information was available that suggests our new approach overestimates the between-field variation, we could replace the Uniform prior distribution for σ^F with a prior distribution (e.g. a Normal-Gamma or a half-Cauchy distribution) that takes that information into account, such as by specifying limits for the variation in field means.

If the number of composite samples, L, increases, the model is able to obtain the target distribution even when the number of units per composite is very small. Although our new approach may overestimate the between-field variation for data sets with small L it is an improvement on current dietary risk assessment approaches for the following reasons. Current approaches ignore the uncertainty about σ^F and are likely to overestimate σ^F because they assign all the observed variation in composite samples to between-field variation. This ignores the fact that some of the variation should be attributed to within-field variation. As current composite data sets only provide poor estimates of field means, we have no evidence to establish whether the observed variation in composite samples stems from between-field variation or unit variation.

The issue when we have a small number of fields in combination with a small number of sampled units per composite is that the data may provide a poor estimate of the between-field variation as illustrated in Figure 5.10.

Figure 5.10 – *Sampling small numbers of field trials may result in a poor representation of the variation in field means. In this example, five field studies were simulated at random by generating 1 or 25 units (green crosses) for each of the five fields. The red circles indicate the composite sample and the blue line indicates the true field mean distribution.*

(a) $L = 5$, $n_l = 1$ **(b)** $L = 5$, $n_l = 25$

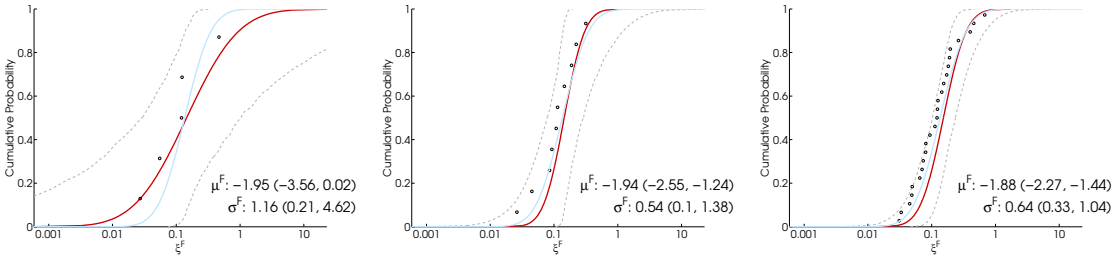


With only one unit sampled per field, the units provide poor estimates of the field means despite the fact that the field means sampled (red circles) represented a wide range of values from the target distribution. In the example with 25 units per composite, the five simulated fields were all sampled from the central part of the distribution, so despite the fact that the 25 units from each field provide a reasonable estimate of each field mean, the model has underestimated the variance of the field mean distribution (see Figure 5.9g).

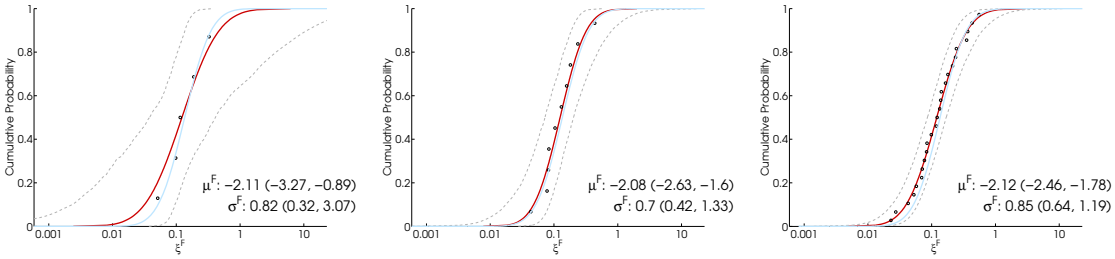
If we take stratified samples (Figure 5.11), we observe an improvement in the model performance as the median field mean distribution now better reflects the target distribution.

Figure 5.11 – *Effect of sample size using stratified samples from the target distribution (blue line). The red line indicates the median field mean distribution estimate with a 95% credible interval indicated by grey dashed lines.*

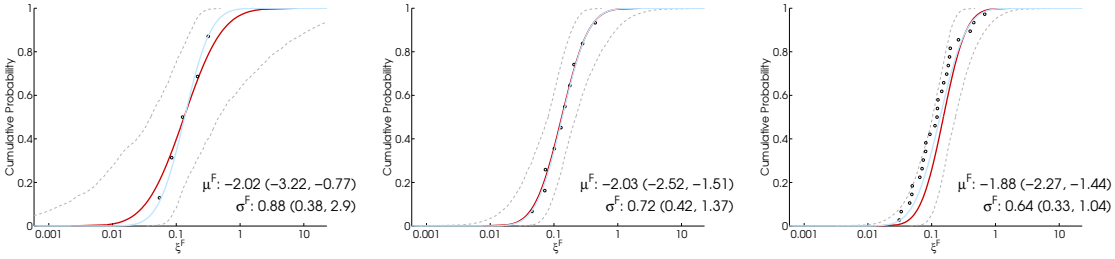
(a) $L = 5$, $n_l = 1$ **(b)** $L = 10$, $n_l = 1$ **(c)** $L = 25$, $n_l = 1$



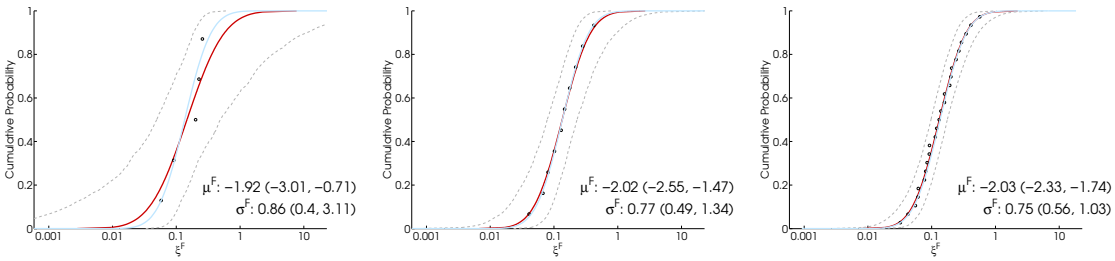
(d) $L = 5$, $n_l = 10$ **(e)** $L = 10$, $n_l = 10$ **(f)** $L = 25$, $n_l = 10$



(g) $L = 5$, $n_l = 25$ **(h)** $L = 10$, $n_l = 25$ **(i)** $L = 25$, $n_l = 25$



(j) $L = 5$, $n_l = 100$ **(k)** $L = 10$, $n_l = 100$ **(l)** $L = 25$, $n_l = 100$



The model still overestimates the variation in field means for $L = 5$ for small values of n_l but in general the model is able to infer the target field mean distribution. This implies that if field trials are designed in such a way that we observe the full range of application conditions (in terms of equipment and weather conditions), as few as 5 trials could be sufficient to determine the variation in mean residue levels between fields. However, given that the simulations are based on a typical value for the between-field variation and that in practice we are unlikely to know whether the fields from which composite samples are obtained provide a representative sample of the variation between fields, further work is needed to recommend sample sizes for the number of field trials and the number of units needed per field trial.

The results indicate that the effect of reducing the uncertainty about the field mean by obtaining more units per composite sample is limited in comparison to a reduction in uncertainty that could be achieved by using a larger number of composite samples. For $L = 25$, the model provides a better and less uncertain estimate of the field mean distribution, even for $n_l = 1$. In contrast, for $L = 5$, the field mean distribution is very uncertain even when n_l is large. For typical data sets, $L = 10$ and $n_l = 10$ and 25, the model performs well.

5.4.3 Validation Study 3: Effect of σ^F

To assess the impact of σ^F on the model performance, we ran further simulation studies in which we set σ^F to be the small (0.25), typical (0.75) and large (1.5) observed variation in composite samples. We used stratified samples to ensure that the simulated data provided a reasonable spread of field means and we generated 12 units per composite sample.

Figure 5.12 – *Effect of σ^F for various numbers of field trials. The target distribution is indicated by the blue line. The red line indicates the median field mean distribution estimate with a 95% credible interval indicated by grey dashed lines.*

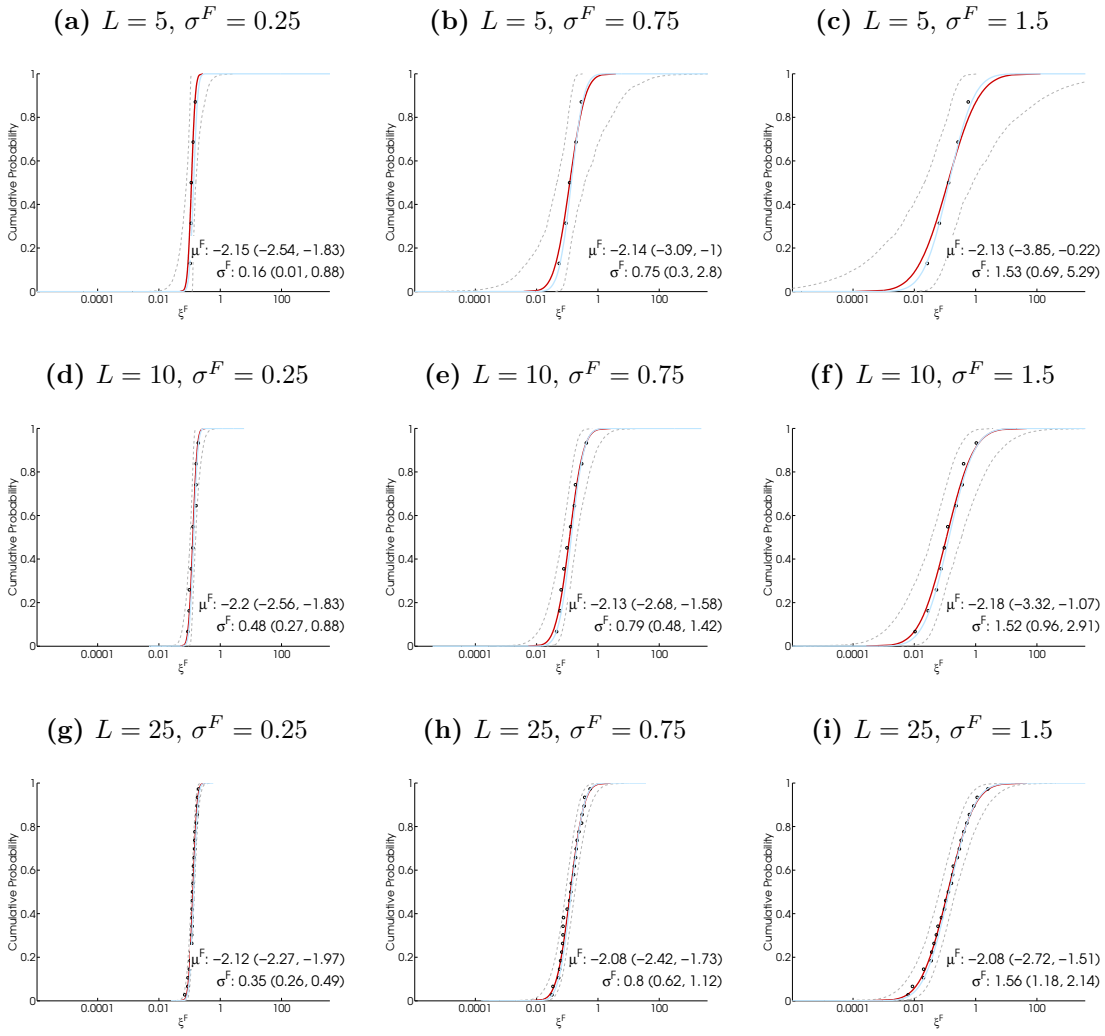


Figure 5.12 shows that the model is capable of inferring the field mean distribution for all values of σ^F, although the performance is better for large values of L. Figure 5.5b seems to indicate that small variation in composite residue levels only occur in supervised trials in which a small number of fields were sampled. It is possible that the variation in equipment, weather conditions, etc. in those supervised trials is smaller than the variation that would be observed if the pesticide was applied in practice.

5.4.4 Validation Study 4: Removing uncertainty about the unit distribution

The fourth set of validation studies aim to remove the uncertainty about the unit distribution to assess whether the model will be better able to infer the between-field distribution shape if we have a better estimate of the unit distribution shape. The first approach taken was to fix the shape of the unit distribution for every field whilst allowing different fields to have different scale parameters. Using a typical $\sigma^F = 0.75$, $L \in \{5, 10, 25\}$ and $n_l \in \{1, 10, 25, 100\}$, the model results, not shown here, indicate that the performance of the model is better than before with the model no longer overestimating σ^F. This suggests that the inference of the between-field distribution will profit from obtaining a more precise estimate of the unit log-residue distribution.

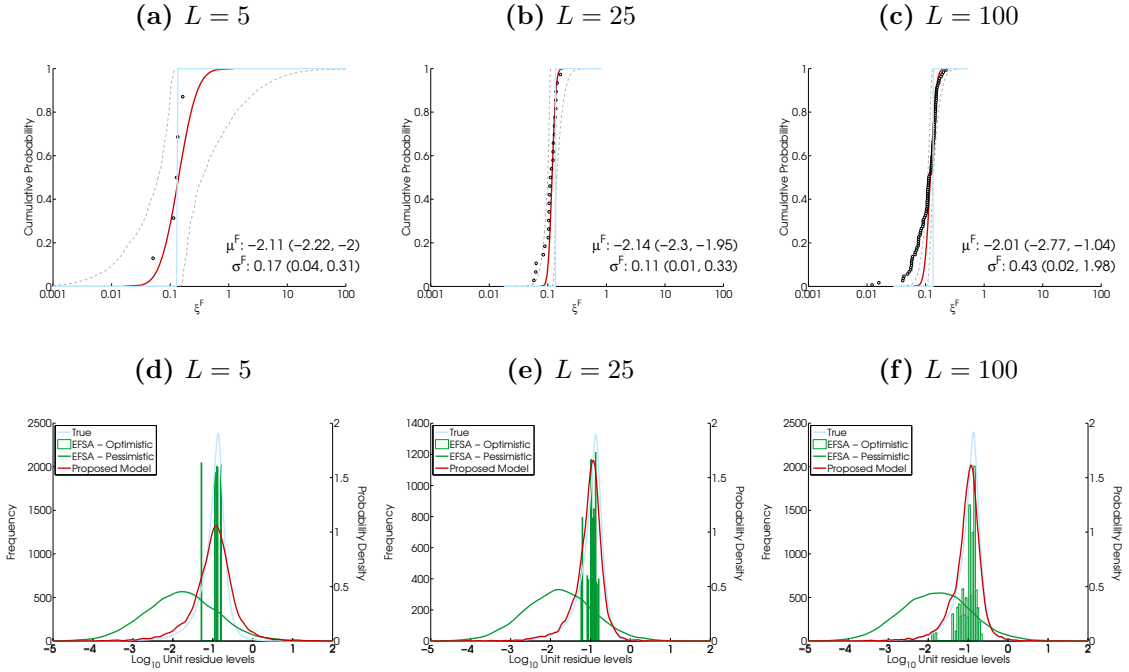
As we may not be able to obtain a more precise estimate of the unit log-residue distribution, another approach is to remove the effect that the uncertainty about the unit distribution has on the inference of the field mean by taking a large sample (e.g. $n_l = 400$) of units per simulated field. In theory we can collect more units to obtain composite samples in supervised field trials, so it is worth exploring what the benefits of doing so would be for the inference of the field mean distribution. Using more units to estimate $\overline{U_l}$ will result in a more certain estimate of ξ_l^F and the field mean distribution parameters. As in the $n_l = 100$ cases, shown in Figures 5.9 and 5.11, the model output (not shown) provides a good estimate of the field mean distribution. However, as mentioned before, the inference of the field mean distribution benefits more from increasing the number of composite samples than from increasing the number of units per composite sample.

5.4.5 Validation Study 5: Simulating no between-field variation

Here we assess the performance of the model if we effectively remove the between-field variation by setting $\sigma^F = 10^{-8}$ and simulate only 1 unit per field. This should in theory result in a simulated sample that reflects the unit distribution and a field

mean distribution that indicates that there is no between-field variation. Figure 5.13 shows the model results for different numbers of simulated field trials. The model struggles to infer the correct field mean distribution as it assigns some of the observed variation to between-field variation, potentially as a result of using a Uniform prior distribution for σ^F. However, if we look at the posterior predictive distributions, the model's performance is acceptable with the possible exception of the $L = 5$ case which overestimates the variation in unit residue levels. The model appears to perform better than the current approaches (Figures 5.13d-f).

Figure 5.13 – *Model results for $\sigma^F = 10^{-8}$ for various field trial sizes and $n_l = 1$. The upper panes (a-c) show the target distribution (blue line), the median field mean distribution estimate (red line) and a 95% credible interval (grey dashed lines). The bottom panes (d-f) show the predictive distributions of our within-field and between-field model in red and the currently recommended approaches (EFSA, 2012) in green together with the target distribution (blue line).*

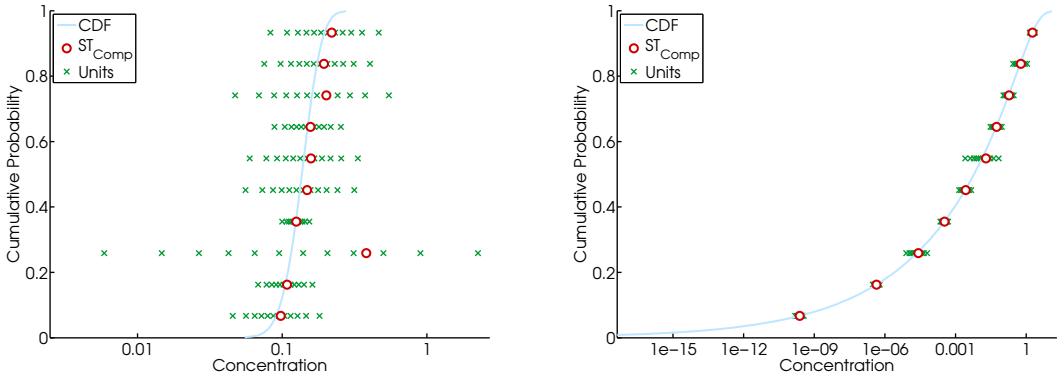


5.4.6 Validation Study 6: Using different distributions for field means and units

The validation studies presented thus far were all based on a Lognormal distribution for field means and the distribution obtained from the DPMN model for units, i.e. the same distributions as used by the within-field and between-field model. As current literature and methods use a Lognormal distribution for field means, we feel that it is not unreasonable to make this assumption here. Nevertheless, to assess how sensitive the results are to the selection of these distributions, an additional set of validation studies was conducted in which we assume a Gamma distribution for the field mean distribution and a Lognormal distribution for the unit variation. Figure 5.14 shows two examples of samples taken from the new target distribution with different values of σ^F. It is clear that assuming a Gamma field mean distribution results in a much longer lower tail for larger values of σ^F.

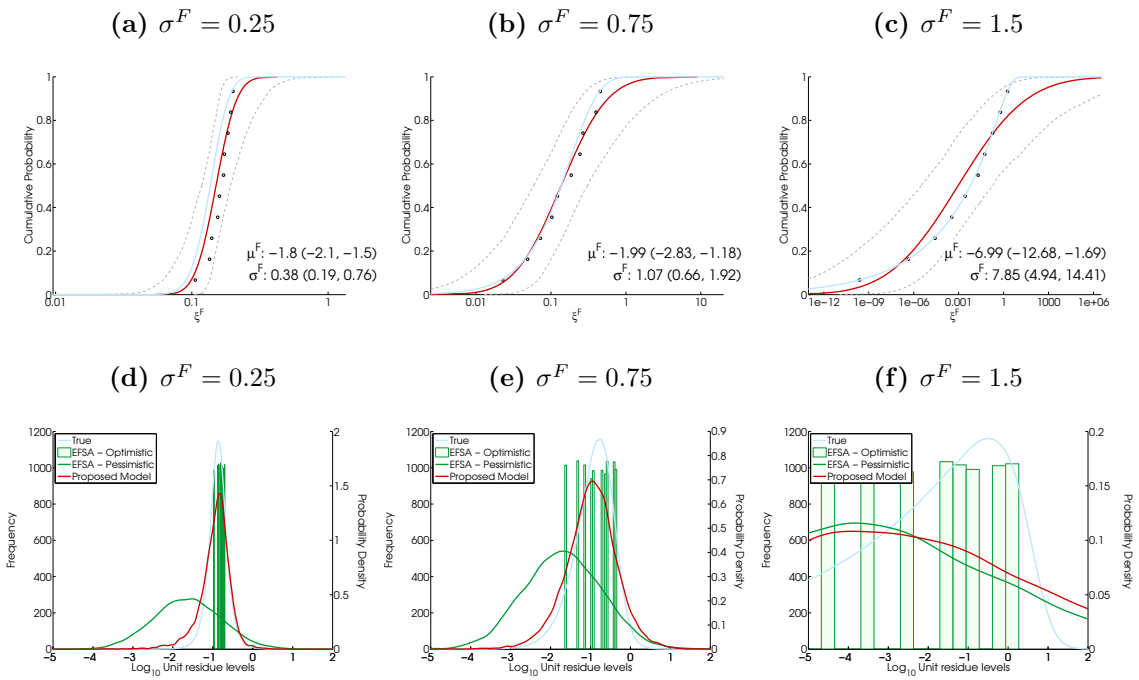
Figure 5.14 – *Twelve unit residue values (green crosses) generated from a Lognormal distribution with mean $\xi_l \sim Gamma(a, b)$ where a and b are set so that the mean equals 0.01 and the standard deviation equals σ^F. The red circles indicate the composite samples and the blue line indicates the cumulative distribution function of the Gamma field mean distribution.*

(a) $\sigma^F = 0.25$ **(b)** $\sigma^F = 1.5$



As expected, the model struggles to infer the target distribution from the simulated data set for large values of σ^F due to the long lower tail. This is not as obvious in the posterior distributions in Figure 5.15, but if we look at the predictive distributions (Figure 5.15d-f) it becomes clear that the model is struggling to some extent with the Gamma shape of the field mean distribution and the Lognormal distribution for units. For $\sigma^F = 0.25$ and $\sigma^F = 0.75$, the model's performance is acceptable as the true distribution falls within the 95% credible intervals. For those two values of σ^F the posterior predictive distributions do well given the fact that both the field mean distribution and the unit distribution in our model have a different shape to the target distributions. These results imply that the method may, within reason, not be very sensitive to slight deviations from the Lognormal shape assumption for the field mean distribution for small and typical σ^F, however it struggles for $\sigma^F = 1.5$. As the target σ^F has been calculated based on composite samples containing a mixture of within-field and between-field variation, we would expect that σ^F is more likely to be smaller than 1.5. For the more likely value $\sigma^F = 0.75$, our model clearly performs much better than the methods currently recommended by EFSA (Figure 5.15d-f).

Figure 5.15 – *Field mean and posterior predictive distributions based on simulation studies in which the 10 field means were generated from a Gamma distribution and the units were generated from a Lognormal distribution. For the field mean distributions (a-c), the target distribution is indicated by the blue line. The red line indicates the median field mean distribution estimate with a 95% credible interval indicated by grey dashed lines. For the posterior predictive distributions (d-f), the target distribution is indicated by the blue line, two alternative EFSA approaches are shown in green and the results from applying our model are shown in red.*



5.4.7 Summary of Validation Studies

The validation simulations indicate that the model is capable of retrieving the target distribution, even for sample sizes as small as $L = 5$, as long as the samples provide a good representation of the true distribution of field means. As we may not know this in practice, the model results seem to suggest that if we want to obtain an estimate of the between-field variation, field trials should be conducted on more than five fields or the number of units used to obtain a composite sample should be at least 25.

The validation studies demonstrated that taking 12 units from a single field may result in very poor representations of the field means. Figure 5.6a shows an example of this where the red circles, representing composite residue levels, are poor estimates of the field means. This issue would affect any model so we recommend increasing the minimum data requirements if regulators want to obtain reliable estimates of residue levels on food items.

A second issue is that the model may overestimate the field mean variation if the variation between composite samples is small due to the Uniform prior distribution for σ^F. In these cases, the model attributes the observed variation in composite samples to the between-field variation, even when it is caused by unit variation. This effect is less pronounced if composite samples from more than 5 fields are collected or if the number of units used to create a composite sample is increased. The obvious solution is to replace the Uniform prior distribution, used in all validation studies, with a different prior distribution (e.g. Normal-Gamma) to express one's beliefs about reasonable values for σ^F. However this would require eliciting values from experts so it was considered to be outside the scope of this thesis.

As no data exist that would provide more information about the field mean distribution shape, we have to rely on simulation studies to assess the sensitivity of the method when other distribution shapes are used to simulate data. The results of these simulation studies imply that the proposed model performs reasonably well and is better at describing the variation in residue levels than existing approaches for small and typical σ^F when the true distributions are Gamma (field mean) and Log-normal (unit residue levels). If evidence became available which suggests different distributions for field means, the model could be adjusted to reflect this.

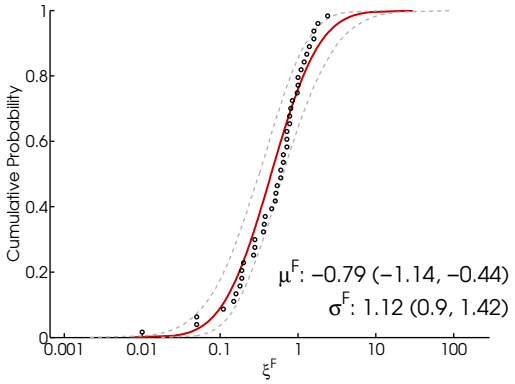
5.5 Case Studies

To illustrate how the model can be used in practice, we ran the model on various supervised trial data sets. Figure 5.16 shows the field mean distribution for four data sets. As we do not know what the true field mean distribution is, we can only

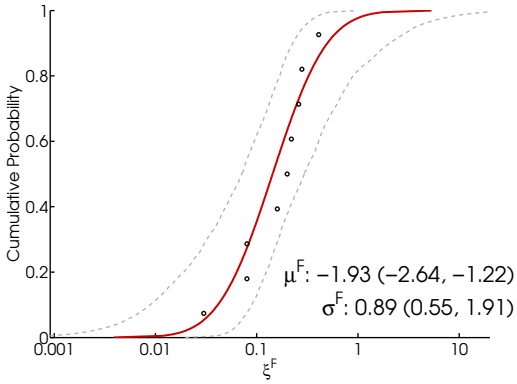
demonstrate that the model can be applied to existing residue data. It is interesting to see the effect of the number of composite samples, L, on the uncertainty of the field mean distribution. In all cases the model attributed some of the observed variation to between-field variation with $\mathbb{E}[\sigma^F]$ of the log field mean distribution being 1.1 for apple, 0.97 for peach, 0.30 for orange and 0.89 for kiwi. Comparing these values to the posterior distribution of σ^u (Figure 5.4c), it seems that the expected between-field variation for all but one of the data sets (orange) is larger than the expected within-field variation ($\mathbb{E}[\sigma^u] = 0.46$). However, as both σ^F and σ^u are uncertain and of the same order of magnitude, both within-field and between-field variation should be modelled in dietary exposure assessments.

Figure 5.16 – *Field mean distribution obtained from applying the model to four supervised trial data sets. As before L indicates the number of supervised field trials and n_l is the number of units per composite sample.*

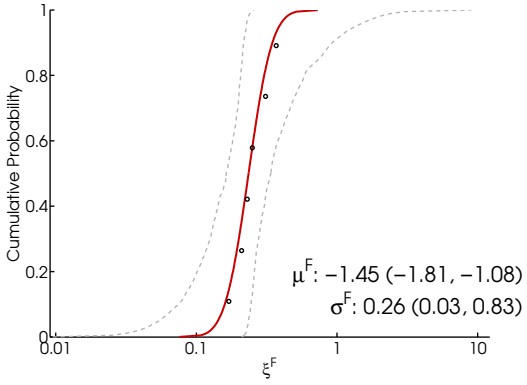
(a) Apple - Benomyl, $L = 42$, $n_l = 12$



(b) Peach - Etofenprox, $L = 9$, $n_l = 24$



(c) Orange - Buprofezin, $L = 6$, $n_l = 12$



(d) Kiwi - Boscalid, $L = 4$, $n_l = 12$

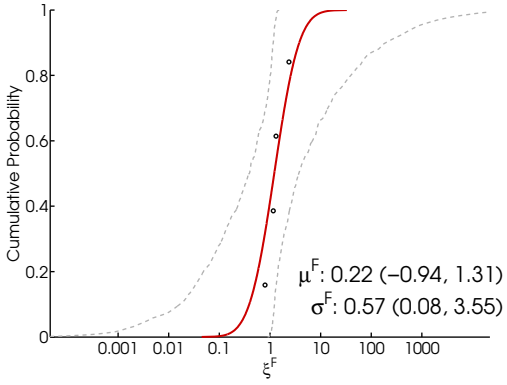


Figure 5.17 – *Posterior predictive distributions obtained by applying the model (green line) and two alternative approaches (blue), recommended by EFSA, to two supervised trial data sets.*

(a) Apple - Benomyl **(b)** Orange - Buprofezin

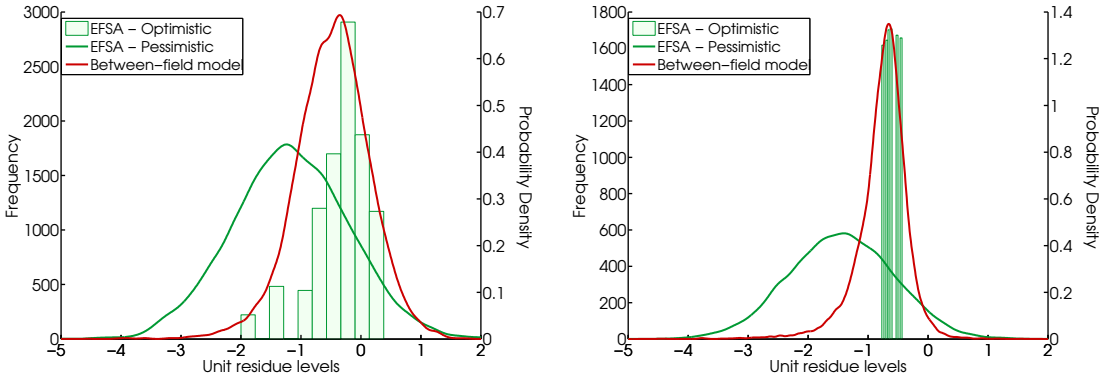


Figure 5.17 shows a comparison of the posterior predictive distributions of the proposed approach with two approaches recommended by EFSA (2012). As the results are similar for all case studies, we only show two examples here. In the validation studies we observed that the 'EFSA Optimistic' approach, based on bootstrapping the composite samples, provided a poor estimate of the residue distribution. This is because the L composite samples are resampled and as L is typically less than ten, this is unlikely to provide a good description of the distribution tails. The 'EFSA Pessimistic' approach includes a large variability factor so it provides the most conservative estimates due to its longer upper tail. However it cannot be considered conservative because the use of a large variability factor leads to a lower mean of the assumed Lognormal distribution and therefore it also provides lower estimates of unit residues than the other approaches.

5.6 Residue Generator

The within-field and between-field model can be used to generate samples from residue level distributions that can be used in probabilistic dietary risk assessment using a series of straightforward algorithms, depending on the aim of the risk as-

sessment.

1. For acute risk assessments for food items that are consumed as single units (e.g. apples, pears, etc.):

   (a) Select the number of units that are obtained from a single field, n_F.

   (b) Select the number of residue levels that you want to simulate, n, where n is a multiple of n_F and set the field index $m = 0$.

   (c) If $m \times n_F \leq n$, repeat the following steps:

      i. Select a set of parameters, $(\mu^F, \sigma^F, \alpha, \beta)$, from the n_{iter} samples of their posterior distributions:

         A. Sample $u \sim$ Discrete Uniform$(1, n_{iter})$.

         B. Select u^{th} value from the posterior distribution sample of ξ^F, σ^F, α and β.

      ii. Sample a field mean: $\log(\xi_l^F) \sim \mathcal{N}\left(\mu^F, (\sigma^F)^2\right)$.

      iii. Sample a scale parameter for the unit distribution: $\sigma_l^u \sim \text{Gamma}\,(\alpha, \beta)$ and calculate $\rho_j = \sum_{q=1}^{C} w_q \exp\left[\theta_{qj}^* + \frac{(\sigma_j^*)^2}{2}\right]$ with $\theta_{qj}^* = \theta_q \sigma_l^u$ and $\sigma_j^* = \sigma_l^u \sigma_c$.

      iv. Repeat the following steps n_F times:

         A. Sample $u \sim$ Uniform$(0, 1)$.

         B. Select component j from the unit mixture distribution if $\sum_{j=0}^{k-1} w_j < u \leq \sum_{j=1}^{k} w_j$, where $w_0 = 0$.

         C. Sample log unit residue: $\log(z_{kl}) \sim \mathcal{N}\,(\theta_j, \sigma_c^2)$ where θ_j is the location parameter of the selected component and σ_c is the component scale parameter.

         D. Rescale z_{kl} so that $\mathbb{E}[z_{kl}] = 1$: $R_{unit} = \frac{\exp\left[\xi_l^F + \log(\tilde{U}_{kl})\sigma_l^u\right]}{\rho_j}$.

      v. Set $m = m + 1$, i.e. select the next field from which we will generate residue levels.

2. For acute risk assessments for food items that are either consumed in bulk per eating event (e.g. grapes, peanuts) or which have been blended during processing (e.g. fruit juice):

(a) If all units are expected to originate from a single field:

i. Select the number of unit residue levels that need to be sampled from a single field (e.g. portion size, number of units in one portion of juice): $n_{portion}$.

ii. Sample unit residue levels R_{unit} using the previous algorithm for single units with $n = n_F = n_{portion}$.

(b) If units are expected to originate from multiple fields:

i. Select the number of units in a portion: $n_{portion}$.

ii. Select the number of unit residue levels that need to be sampled from a single field: n_F. Note that $n_{portion}$ should be divisible by n_F.

iii. Sample unit residue levels R_{unit} using the previous algorithm for single units with $n = n_{portion}$.

(c) Calculate the average concentration over n_F units assuming that the unit weights are constant: $R = \frac{\sum_{i=1}^{n_{portion}} R_{unit}}{n_{portion}}$.

5.7 Discussion

In this section we will provide a brief discussion of issues related to the data, the model performance and the model's sensitivity to prior distributions.

5.7.1 Data

We assume that variation in unit field trial data is representative of variation in residue levels on units in supervised trials, despite the fact that field trials are often conducted at higher residue levels and consist of mixtures of pesticides. This is supported by both Ambrus (2006) and MacLachlan and Hamilton (2011) who state that studies indicate that the variation in log-residues levels is not significantly influenced by the application rate.

We also assume that the variation in the scale parameter of the unit log-residue model based on the 75 unit data sets provides a representative estimate of the variation in unit residue levels in supervised trials for a new pesticide. It is possible

that the variation in residue levels from field trial unit data is larger than one would expect for a new pesticide use as field trial data include a range of pesticides and crops and may consist of a wider range of spraying techniques, environmental conditions etc.

5.7.2 MCMC Performance

The MCMC performance of the model was monitored post-analysis by plotting the chains and assessing the auto-correlation. Generally, a thinning factor of 300 was used for each model run to overcome mixing problems. These problems were caused by the fact that we sample unit scale parameters σ_l^u from a Gamma distribution independently of U_{kl}. Large increases in σ_l^u (e.g. moving from small within-field to large within-field variation) result in a high rejection probability which causes the chain to get stuck temporarily. Although thinning overcomes this problem, a neater solution that could be explored in future research would be to either sample values of σ_l^u dependent on U_{kl} or to propose a new σ_l^u that is not too far from the current value. However, as both options may result in smaller step changes between iterations, it is unclear whether they will lead to an improved exploration of the parameter space.

5.7.3 Choice of Prior distributions for σ^F

All model calculations presented in this chapter are based on a Uniform prior distribution for σ^F. The reason for selecting this prior distribution is that it is non-informative and unlike the informative Normal-Gamma prior, it does not require the specification of 4 input parameters. As the Normal-Gamma prior distribution is not recommended if one wants to use a non-informative prior (Gelman, 2006), we suggest using a Uniform prior distribution for σ^F. However, when L is small this distribution may result in an overestimation of σ^F, so care must be taken when few composite samples are available.

5.8 Conclusions

The novel approach presented in this chapter accounts for within-field and between-field variation and provides a better estimate than current methods of the variation in residue levels on crop units. The within-field model, used to describe variation between unit residue levels, was adapted from the model described in Chapter 4 which also included a full discussion of the benefits and disadvantages of that model. The hierarchical refinements made to the unit model in this chapter allow us to estimate the unit variation over the selected data sets, conditional on the common shape assumption. The analysis of the available unit data sets (Figure 5.4e) implies that the median variability factor is 2.6 with a 95% credible interval ranging from 1.3 to 5.0.

It is clear from the application of the recommended EFSA methods in the validation exercises that using a parametric or empirical distribution for composite values may lead to a poor estimate of residue levels. Figure 5.7a shows that the link between composite values and field means is relatively weak and that we should account for the fact that composite samples are an uncertain estimate of the field mean when inferring the field mean distribution. Methods that are based on fitting a distribution directly to composite samples double-count the unit variation and incorrectly assume that the resulting distribution is a distribution of field means to which a variability factor can be applied. Figure 5.8 shows how two approaches, currently recommended by EFSA (2012), provide poor estimates of the target distribution whilst the novel within-field and between-field model performs well.

The results from four case studies indicate that a significant proportion (up to 60%, calculated as the median of $\mathbb{E}[\text{Var}[\overline{U_l}]]/(\mathbb{E}[\text{Var}[\overline{U_l}]] + (\sigma^F)^2)$) of the observed variation in composite samples from supervised trials is a result of the variation in units. This illustrates the importance of accounting for the unit variation when inferring the field mean distribution. Applications of the model to various validation data sets indicate that the model may sometimes overestimate the variance of the

field mean distribution, particularly when only a few fields were included in the field trials. To overcome this, either residue data needs to be collected from more fields or the Uniform prior distribution used in the validation studies could be replaced by a suitable alternative.

One of the main challenges for probabilistic dietary risk assessment is the absence of good quality residue level data. The large variation in the scale parameter of the unit log-residue distribution means that any predictions based on this scale parameter distribution will be very uncertain. If notifiers were asked to provide unit data from supervised field trials we would have a much better picture of the within-field variation. This may lead to less uncertainty about the residue levels on units if the scale parameters of the provided unit data have a smaller range than the range suggested by the hierarchical log-residue unit model. However, as it is unlikely that regulators will reconsider the data requirements for dietary risk assessments, the model presented here provides a better estimate of residue levels given the available data than alternative approaches as proposed by EFSA (2012).

Chapter 6

Conclusions and Future Research

This chapter provides a summary of this thesis, highlights the novel approaches developed to describe the variation of pesticide residue levels on food items and discusses future areas of research.

6.1 Summary

Chapter 1 presents an overview of the current regulatory framework for dietary risk assessment for plant protection products in the EU. It also introduces current deterministic and probabilistic methods and provides an overview of issues with the data that are routinely collected, the use of these data by current methods and the methods themselves. Chapter 2 introduces several mathematical concepts that are used throughout the thesis. This includes an introduction to Bayesian methods, Monte Carlo algorithms and Dirichlet Distributions and Processes which are essential for the novel approaches developed in this thesis. We also provide an overview of various methods for sampling from a Dirichlet Process.

Residue levels of multiple pesticides may occur on individual food items when more than one pesticide is applied to a crop. To model the variation in these residue levels multivariate techniques are needed to account for any correlations in residue levels. Chapter 3 presented two novel approaches to model pesticide log-residues

in composite samples using available monitoring data and Pesticide Usage Survey (PUS) data. The models use the PUS data to inform them about the proportion of composite samples that have been treated with pesticides. The independent mixture model assumes independence between pesticide log-residue levels on composite samples whereas the bivariate mixture model infers the correlation between them. We used validation studies to show that our models performed well for synthetic data sets and compared the novel approaches with the currently used approaches. These comparisons indicated that the use of PUS data improved the inference of the log-residue distributions, particularly when commonly observed high levels of censoring were induced. For the validation simulations presented in this thesis the novel approaches performed better than the existing approaches. The independent mixture model is an improvement on the existing Paulo et al. (2005) approach as the use of PUS data reduces the uncertainty about the proportion of untreated samples. Therefore this method is useful when log-residue levels in a composite sample are assumed to be independent. The bivariate mixture model is an improvement on the existing bootstrap methods because it provides a better description of the log-residue distribution as it is not restricted to the observed values and provides an estimate of the correlation between log-residue levels. Therefore these novel approaches offer a promising alternative to current approaches for dietary risk assessment. However, we only consider the bivariate case in this thesis as the number of model parameters increases considerably in higher dimensions. If more than two pesticides have been applied to a crop, there may not be enough data to infer the parameters reliably unless more data are available or stronger assumptions are made. In addition, high levels of censoring in residue data means that the results rely on the choice of prior distributions and the availability and relevance of PUS data.

In Chapter 4 we introduced a novel non-parametric Bayesian approach to describe the distribution of unit log-residue levels. As unit log-residue data sets are relatively small, the novel approach shares information on the shape distribution between samples obtained from multiple populations, leading to a larger data set from which the population shape distribution can be inferred. We use a blocked Gibbs sampler to

alternately sample the individual location and scale parameters of each log-residue data set and the common shape distribution using the normalised, pooled log-residue data. The shape distribution for the log-residue levels is modelled using a Dirichlet Process mixture of Normal distributions (DPMN) model and accounts for uncertainties introduced by censored and rounded data. The Bayesian framework used in this model also allows us to account for parameter uncertainty. Despite EFSA (2012) suggesting that a Normal distribution may not always be appropriate, current approaches tend to assume a Normal distribution for log-residues. Therefore our new approach is an improvement as it learns the shape distribution from the data. Validation studies showed that the model performed well for a range of distributions with short and medium tails. For heavy-tailed distributions a refinement may be needed if the distribution cannot be transformed to have a shorter tail. The validation studies also indicated that the model performed well for $n > 50$ and that learning the DPMN concentration parameter γ did not have a large effect. The method depends on the assumption that the individual log-residue data sets share the same shape distribution. However, it may not always be easy to assess whether this assumption is justified due to the small sample sizes available. To assess the impact of the common shape assumption, the model can be run on individual data sets or subsets of the data.

In Chapter 5 we proposed a novel approach to model variation in residue levels that not only uses a data-driven description of unit variation but, unlike existing approaches, does not double-count unit variation when accounting for within-field and between-field variation. When inferring the field mean distribution we need to 'remove' the unit variation component in the observed variation in composite samples to obtain a distribution describing the variation in field mean residue levels. The within-field model, used to describe variation between unit residue levels, was adapted from the model described in Chapter 4. The new approach accounts for the small number of units used in composite samples, the small number of composite samples used to describe between-field variation and for the fact that composite residue levels from supervised trials already contain unit variation.

Validation studies indicated that the novel approach performs better than existing approaches. This may be because current approaches are based on fitting a distribution directly to composite samples and therefore double-count the unit variation. If only a few composite samples (i.e. <5) are available, the new approach may overestimate the variance of the field-mean distribution. Possible solutions include either collecting data from more fields or replacing the uniform prior distribution by a suitable alternative (Gelman, 2006). The novel method can be applied to data sets that are routinely collected as part of the pesticide registration process and therefore provides a feasible alternative to current approaches.

A major challenge for probabilistic dietary risk assessment is that there is a lack of good quality residue level data. If it was a requirement that notifiers had to provide unit field trial data, this would result in a much clearer picture of within-field and between-field variation. However, as regulators are unlikely to impose this requirement, the models presented in this thesis provide better estimates than existing approaches and account for the relevant uncertainties.

6.2 Future Research

In this section we discuss areas of research which would either be beneficial to refine the new approaches presented in this thesis or to explore areas where the approaches could be applied in the future.

6.2.1 Ideas for future research

6.2.1.1 Extending the bivariate mixture model

The bivariate mixture model presented in Chapter 3 is based on a mixture of univariate and bivariate Normal distributions. This is in line with current dietary exposure models which make the assumption that log-residue data are Normally distributed. However EFSA (2012) suggested that this assumption is not always valid and therefore it would be useful to extend the model to mixtures of other distributions.

Another area of research would be to extend the bivariate mixture model to more dimensions as there may be occasions where more than two pesticides are used on units in a single composite sample. As the number of model parameters that need to be inferred increases considerably for more than two pesticides, it is unlikely that they can be estimated reliably from the limited number of monitoring data that are generally available. As a result, any attempt to model the cumulative exposure for more than two pesticides will have to rely heavily on assumptions, e.g. by eliciting prior distributions from experts or by reducing the number of model parameters as discussed in Section 3.5.3.

6.2.1.2 Modelling of correlations of unit log-residue levels

Correlations in composite monitoring data are unlikely to provide any indication about the correlation in unit log-residue levels. As there is generally little information available about these correlations we proposed a scenario-based approach in Section 3.8 which can be used to estimate log-residue levels on individual food items. However, more research is needed to validate this approach and/or to reduce the number of scenarios considered.

6.2.1.3 Refinement of common shape assumption

The DPMN model presented in Chapter 4 is based on the assumption that log-residue data for multiple pesticide/crop combinations can be described by a single shape. This reflects the current state of the art which assumes that a Normal shape can be fitted to unit log-residue data. However, QQ-plots of the field trial data indicate that instead of a single shape, the log-residue data may be better described by multiple shapes. One solution would be to model each data set separately but this approach effectively reverts back to existing implementations of DPMN models. This also has the disadvantage that the shape distribution will have to be inferred from relatively small sample sizes, leading to more uncertainty in the parameter estimates. Therefore it may be preferable to replace the common shape assumption with a more flexible approach.

One option is to ask pesticide/crop specialists to select subsets of data for which

the common shape assumption seems reasonable (i.e. supervised learning). Alternatively, we can use unsupervised learning approaches (e.g. a hierarchical DP or reversible jump MCMC) to let the data determine how many shapes are necessary to describe the data. The advantage of this is that it would only use the data to infer the subsets of data which share a common shape.

One drawback of any approach that involves fitting multiple shape distributions to the data is that fewer data sets will be available to infer each shape distribution. In other words, the more shapes we fit, the fewer residue data that will be available to estimate the parameters of each shape distribution. As a result, we may be more uncertain about each shape distribution. Therefore the number of shapes that we use to describe the data needs to be balanced with the amount of data available to learn each of the shapes. However, as forcing all data to share a single shape introduces uncertainty about the distribution shape as well, we do not know the net effect. Another issue is that both supervised and unsupervised learning may result in undesirable clustering of data sets. For example, if multiple unit residue data sets for pesticide X on crop Y are assigned to different shape distributions (by experts or a model), predicting unit residue levels for pesticide X on crop Y may be quite complex to explain.

6.2.1.4 Mixture of other distributions

Another area of future research to improve the DPMN model from Chapter 4 is to explore mixtures of other distributions. The validation studies for the DPMN model indicated that the model struggled to describe distributions with long tails because of the tail characteristics of the Normal components. To overcome these problems, we could either consider using a different shape for the components or develop a new approach that combines a DPMN model with another distribution, for example a Generalised Pareto distribution (GPD). The DPMN model can then be used for the main body of the distribution and the GPD can be used to model the lower and upper tails. Alternatively, we can apply data transformation techniques, use the existing DPMN model on the transformed data and transform back to the original

scale afterwards.

6.2.1.5 Accounting for non-treated data

The DPMN model, introduced in Chapter 4, could also be extended to allow for non-treated data, offering a more flexible approach to the univariate mixture model introduced in Chapter 3, e.g. $p(y) = w_0\delta_0 + (1 - w_0)DPMN(y|G_0, \gamma)$, where w_0 is the proportion of untreated data and δ is the Kronecker delta function. Such a model would be particularly helpful to describe composite data from monitoring programmes as they may contain untreated and treated samples.

6.2.1.6 Further Validation Studies

The existing validation studies that were carried out for the DPMN model from Chapter 4 consisted of generating a sample from either multiple populations with a common shape (mixture of Normal distributions) or a single population (all other simulations). Even though the method assumes a common shape for all the data, it might be useful to assess how sensitive the method is to deviations to the common shape assumption. This would provide us with some indication of how robust the approach is and how much effort should be put into establishing that a data set obtained from multiple populations can be considered to be a data set from populations with a common shape.
Another aspect that could be explored further is to assess to what extent the conclusions regarding the effect of sample size, the DP concentration parameter γ and the smoothness parameter κ from the existing validation studies can be extrapolated to distribution shapes other than the mixture of two Normal distributions.
A final set of additional validation studies could explore the effect of the DP base distribution G_0. All the model runs in Chapters 4 and 5 were based on $G_0 \sim \mathcal{N}(\mu_0, \sigma_0^2)$. We proposed the Student's t distribution as an alternative for G_0 but the effect of changing G_0 has not been assessed.

The validation studies for the within-field and between-field model from Chapter 5 can be extended to assess the performance of the method when other distributions

are used to simulate the validation data sets. As we do not have information on the field mean distribution, this would provide us with a better understanding of the model's limitations.

6.2.1.7 Other applications of the DPMN model

The DPMN model, introduced in Chapter 4, was developed to model the variation in unit log-residue levels and is based on the assumption that multiple data sets can be described using a common shape distribution. However, it may be applicable to other risk-related problems where distributional assumptions are made. One example could be to model consumption data, possibly by allowing different shapes for different types of food. For example we could distinguish staple foods, food items that are eaten regularly and in such quantities that it constitutes a dominant portion of a diet and food items that are consumed rarely.

The approach could also be applied outside the field of dietary risk assessments. For example, the DPMN model can be applied to ecotoxicity data that are used to describe the variation in sensitivity between species. Even though the number of species tested for each chemical is relatively small, large data sets exist for a wide range of chemicals. Both examples indicate that there is a much wider scope for common shape DPMN models in probabilistic modelling. Further applications of the model will help us obtain a better understanding of the model's behaviour and may provide new application-specific challenges to overcome.

6.2.1.8 Refinements to the within-field and between-field model

The within-field and between-field model, introduced in Chapter 5, offers various options for refinements. Firstly, we could explore whether it is feasible to replace the non-informative Uniform prior distribution for the scale parameter of the field mean distribution with alternative distributions, in particular by using a Normal-Gamma distribution. Secondly, we could replace the Lognormal distribution assumption for field means and/or the unit residue distribution with (an) alternative distribution(s), e.g. the unit distribution could be obtained using a different grouping of data sets than the group used in Chapter 4. A further refinement that could be considered is

to improve the MCMC performance for the within-field and between-field model by sampling σ_l^u in the Metropolis-Hastings step t to be dependent on $\mathbf{U}_l$ in step $t-1$. However, although this would overcome the high rejection probability, smaller step changes in σ_l^u may not result in a more efficient exploration of the sample space.

6.2.1.9 Elicitation of prior distributions

Using a weakly informative prior for the bivariate mixture model in Chapter 3, improved the estimates for the parameters. When there are high levels of censoring, e.g. validation data set C, the use of informative prior distributions seems to provide a transparent approach for predicting residue levels in food items which allows for an assessment of the impact of choosing different prior distributions. For the model presented in Chapter 5, using information from other pesticides may improve the estimate of the field mean distribution, particularly for small data sets. Expert elicitation is a systematic approach that aims to translate subjective judgements into a probability distribution (Slottje et al., 2008). This approach could be used to incorporate any available information into prior distributions for the models presented in this thesis.

6.2.2 Prioritisation of refinement options

In this chapter we have discussed several possibilities for future work to improve or further test the approaches developed in this thesis. We will now briefly discuss the three options which we feel should be the main priorities.

1. Refining the common shape assumption is an important area for future research because the unit data used in Chapter 4 suggested that the data sets may not share a common shape. However this is difficult to assess for small sample sizes which contain rounded and censored values. If the model in Chapter 5 is to be used in a regulatory context it is important that the shape distribution provides a good representation of the variation in residue levels.

2. Although we have tested the models introduced in this thesis extensively in validation studies, additional validation studies as described in Section 6.2.1.6

would be useful to assess how robust the common shape DPMN model is, particularly to deviations from the common shape assumption.

3. There are various areas where the models presented in this thesis may improve current practice as described in Section 6.2.1.7 so it would be useful to explore these fully to assess in which other application areas the models can be useful. Using the model in different areas would also provide more opportunities for validation and further development of the model.

Bibliography

Ambrus, A. (1979). 'The Influence of Sampling Methods and Other Field Techniques on the Results of Residue Analysis.' In: *Pesticide residues: A contribution to their interpretation, relevance, and legislation.* Ed. by H. Frehse and H. Geissbühler. Oxford: Pergamon Press, 6–18.

Ambrus, A. (1995). *Residues of Chlorpyrifosmethyl in Apples at Intervals Following a Single Application of Reldan 50 (EF 917).* R95-143. Dow Elanco, Indianapolis, USA.

Ambrus, A. (2000). Within and Between Field Variability of Residue Data and Sampling Implications. *Food Additives and Contaminants.* **17** (7), 519–537.

Ambrus, A. (2006). Variability of Pesticide Residues in Crop Units. *Pest Management Science.* **62** (8), 693–714.

Antoniak, C. E. (1974). Mixtures of Dirichlet Processes with Applications to Bayesian Nonparametric Problems. *The Annals of Statistics.* **2** (6), 1152–1174.

Blackwell, D. and J. MacQueen (1973). Ferguson Distributions via Pólya Urn Schemes. *The Annals of Statistics.* **1** (2), 353–355.

Blei, D. M. and M. J. Jordan (2006). Variational Inference for Dirichlet Process Mixtures. *Bayesian Analysis.* **1** (1), 121–144.

Boon, P. E., H. Van der Voet, and J. Van Klaveren (2003). *Dietary Exposure to Pesticides: Relevant Variables and Probabilistic Modelling.* Report 2003.008. RIKILT - Institute of Food Safety.

Box, G. E. and G. C. Tiao (1973). *Bayesian Inference in Statistical Analysis.* Wiley - InterScience.

Brandstetter, B., A. Korfmann, A. Kroke, N. Becker, M. Schulze, and H. Boeing (1999). Dietary Habits in the German EPIC Cohorts: Food Group Intake

Estimated with the Food Frequency Questionnaire. *Annals of Nutrition and Metabolism.* **43**, 246–257.

Connor, R. J. and J. E. Mosimann (1969). Concepts of Independence for Proportions with a Generalization of the Dirichlet Distribution. *Journal of the American Statistical Association.* **64** (325), 194–206.

De Boer, W. and H. Van der Voet (2011). *MCRA 7, a Web-based Program for Monte Carlo Risk Assessment.* Reference Manual 2011-12-19. Biometris.

De Finetti, B. (1931). Funzione Caratteristica di un Fenomeno Aleatorio. *Atti della R. Academia Nazionale dei Lincei, Serie 6. Memorie, Classe di Scienze Fisiche, Mathematice e Naturale.* **4**, 251–299.

DeGroot, M. H. (1970). *Optimal Statistical Decisions.* Wiley - InterScience.

Dorne, J. L. C. M. and A. G. Renwick (2005). The Refinement of Uncertainty/Safety Factors in Risk Assessment by the Incorporation of Data on Toxicokinetic Variability in Humans. *Toxicological Sciences.* **86** (1), 20–26.

EC (1991). *Council Directive 91/414/EC of 15 July 1991 Concerning the Placing of Plant Protection Products on the Market.*

EC (1997). *General Recommendations for the Design, Preparation and Realization of Residue Trials. 7029/VI/95 rev.5, VI B II-1 Appendix B, 22 July 1997.* Available from: http://ec.europa.eu/food/plant/protection/resources/app-b.pdf.

EC (2002). *Commision Directive 2002/63/EC of 11 July 2002 Establishing Community Methods of Sampling for the Official Control of Pesticide Residues in and on Products of Plant and Animal Origin and Repealing Directive 79/700/EEC.*

EC (2005). *Regulation (EC) No. 396/2005 of the European Parliament and of the Council of 23 February 2005 on Maximum Residue Levels of Pesticides in or on Food and Feed of Plant and Animal Origin and Amending Council Directive 91/414/EEC.*

EC (2009). *Regulation (EC) No. 1107/2009 of the European Parliament and of the Council of 21 October 2009 Concerning the Placing of Plant Protection Products on the Market and Repealing Council Directives 79/117/EEC and 91/414/EEC.*

EC (2011a). *Commision Implementing Regulation (EU) No. 540/2011 of 25 May 2011 Implementing Regulation (EC) No. 1107/2009 of the European Parliament and of the Council as Regards the List of Approved Active Substances.*

EC (2011b). *Commision Regulation (EU) No. 545/2011 of 10 June 2011 Implementing Regulation (EC) No. 1107/2009 of the European Parliament and of the Council as Regards the Data Requirements for Plant Protection Products.*

Efron, B (1979). Bootstrap Methods: Another Look at the Jacknife. *The Annals of Statistics*. **7** (1), 1–26.

EFSA (2005). Opinion of the Scientific Panel on Plant Health, Plant Protection Products and their Residues on a Request from the Commission Related to the Appropriate Variability Factor(s) to be Used for Acute Dietary Intake Assessment of Pesticide Residues in Fruit and Vegetables. *EFSA Journal*. **177**, 1–61.

EFSA (2007a). Opinion of the Scientific Panel on Plant Protection Products and their Residues on a Request from the Commission on Acute Dietary Intake Assessment of Pesticide Residues in Fruit and Vegetables. *EFSA Journal*. **538**, 1–88.

EFSA (2007b). Reasoned Opinion on the Potential Chronic and Acute Risk to Consumers' Health Arising from Proposed Temporary EU MRLs. *EFSA Journal*. **5** (3).

EFSA (2009). Scientific Opinion on Risk Assessment for a Selected Group of Pesticides from the Triazole Group to Test Possible Methodologies to Assess Cumulative Effects from Exposure through Food from these Pesticides on Human Health. *EFSA Journal*. **7** (9), 1167.

EFSA (2010a). 2008 Annual Report on Pesticide Residues. *EFSA Journal*. **8** (6), 1646.

EFSA (2010b). Management of Left-censored Data in Dietary Exposure Assessment of Chemical Substances. *EFSA Journal*. **8** (3), 1557.

EFSA (2011). The 2009 European Union Report on Pesticide Residues in Food. *EFSA Journal*. **9** (11), 2430.

EFSA (2012). Scientific Opinion: Guidance on the Use of Probabilistic Methodology for Modelling Dietary Exposure to Pesticide Residues. *EFSA Journal*. **10** (10), 2839.

Escobar, M. D. (1994). Estimating Normal Means with a Dirichlet Process Prior. *Journal of the American Statistical Association*. **89** (425), 268–277.

Escobar, M. D. and M. West (1995). Bayesian Density Estimation and Inference Using Mixtures. *Journal of the American Statistical Association*. **90** (430), 577–588.

Fearnhead, P. (2004). Particle Filters for Mixture Models with an Unknown Number of Components. *Statistics and Computing*. **14** (1), 11–21.

FEPA (1985). *Food and Environment Protection Act. Part III*. http://www.legislation.gov.uk/ukpga/1985/48/pdfs/ukpga_19850048_en.pdf. Accessed 8 January 2013.

Fera (2011). *Surveys of the Agri-environment.* Available from: http://www.fera.defra.gov.uk/scienceResearch/science/lus/pesticideUsage.cfm.

Ferguson, T. S. (1973). A Bayesian Analysis of Some Nonparametric Problems. *The Annals of Statistics*. **1** (2), 209–230.

Folland, G. B. (1999). *Real Analysis: Modern Techniques and Their Applications*. Wiley - InterScience.

Gamerman, D. and H. F. Lopes (2006). *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*. 2nd ed. Chapman and Hall/CRC.

Gelfand, A. E. and A. Kottas (2002). A Computational Approach for Full Nonparametric Bayesian Inference under Dirichlet Process Mixture Models. *Journal of Computational and Graphical Statistics*. **11** (2), 289–305.

Gelfand, A. E. and A. F. Smith (1990). Sampling-Based Approaches to Calculating Marginal Densities. *Journal of the American Statistical Association*. **85** (410), 398–409.

Gelman, A. (2006). Prior Distributions for Variance Parameters in Hierarchical Models. *Bayesian Analysis*. **1** (3), 515–533.

Geman, S. and D. Geman (1984). Stochastic Relaxation, Gibbs Distributions and the Bayesian Restoration of Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. **6**, 721–741.

Gibney, M. and H. Van der Voet (2003). Introduction to the Monte Carlo Project and the Approach to the Validation of Probabilistic Models of Dietary Exposure to Selected Food Chemicals. *Food Additives and Contaminants*. **20**. Supplement 1, S1–S7.

Gilks, W. (1992). 'Derivative-free Adaptive Rejection Sampling for Gibbs Sampling.' In: *Bayesian Statistics*. Ed. by J. Bernardo, J. Berger, A. Dawid, and A. Smith. Oxford: Oxford University Press, 169–194.

Gilks, W. and P. Wild (1992). Adaptive Rejection Sampling for Gibbs Sampling. *Applied Statistics*. **4** (2), 337–348.

Görür, D. and C. Rasmussen (2010). Dirichlet Process Gaussian Mixture Models: Choice of the Base Distribution. *Computer Science and Technology*. **25** (4), 615–626.

Green, P. J. and S. Richardson (2001). Modelling Heterogeneity With and Without the Dirichlet Process. *Scandinavian Journal of Statistics*. **28**, 355–375.

Gregory, J., S. Lowe, C. Bates, A. Prentice, L. Jackson, G. Smithers, R. Wenlock, and M. Farron (2000). *National Diet and Nutrition Survey: Young People Aged 4 to 18 Years, Vol. 1. Report of the Diet and Nutrition Survey.* London: TSO.

Hamilton, D., A. Ambrus, R. Dieterle, A. Felsot, C. Harris, B. Petersen, K. Racke, S.-S. Wong, R. Gonzalez, K. Tanaka, M. Earl, G. Roberts, and R. Bhula (2004). Pesticide Residues in Food - Acute Dietary Exposure. *Pest Management Science*. **60**, 311–339.

Hastings, W. (1970). Monte Carlo Sampling Methods Using Markov Chains and their Applications. *Biometrika*. **57** (8), 311–339.

Hill, A. and S. Reynolds (2002). Unit-to-unit Variability of Pesticide Residues in Fruit and Vegetables. *Food Additives and Contaminants*. **19** (8), 733–747.

Hoare, J., L. Henderson, C. Bates, A. Prentice, M. Birch, G. Swan, and M. Farron (2004). *The National Diet and Nutrition Survey: Adults Aged 19 to 64 Years. Volume 5: Summary Report.* London: TSO.

Holland, P. and C. Malcolm (2002). *Distribution of Pesticide Residues on Fruit Within the Canopy.* Unpublished. New Zealand Kiwifruit Marketing Board.

Ishwaran, H. and L. F. James (2001). Gibbs Sampling Methods for Stick-Breaking Priors. *Journal of the American Statistical Association.* **96** (453), 161–173.

Ishwaran, H. and L. F. James (2003). Some Further Developments for Stick-breaking Priors: Finite and Infinite Clustering and Classification. *Sankhyā: The Indian Journal of Statistics.* **65**, 577–592.

Ishwaran, H. and M. Zarepour (2000). Markov Chain Monte Carlo in Approximate Dirichlet and Beta Two-parameter Process Hierarchical Models. *Biometrika.* **87** (2), 371–390.

Jain, S. and R. M. Neal (2004). A Split-Merge Markov Chain Monte Carlo Procedure for the Dirichlet Process Mixture Model. *Journal of Computational and Graphical Statistics.* **13** (1), 158–182.

JMPR (1999). *Joint FAO/WHO Meeting on Pesticide Residues.* Report of the Joint Meeting of the FAO Panel of Experts on Pesticide Residues in Food, the Environment, and the WHO Core Assessment Group on Pesticide Residues. Rome, Italy.

JMPR (2002). *Pesticide residues in food - 2002.* Report of the Joint Meeting of the FAO/WHO Meeting of Experts. Rome, Italy.

JMPR (2003). *Pesticide Residues in Food - 2003.* 2003 Joint FAO/WHO Meeting on Pesticide Residues. Geneva.

Kaethner, M. (2001a). *Determination of Residue Variability in Head Lettuce Following a Tank-mix Application of Anilinopyrimidine, Triazole, Pyrethroid, Organophosphate and Dicarboximide Crop Protection Products, France/Germany 2000 to 2001.* BASF DocID 2002/1007078. Unpublished. European Crop Protection Association (ECPA), Residue Expert Group. Belgium.

Kaethner, M. (2001b). *Determination of Residue Variability in Table and Wine Grapes After a Tank-mix Application of Anilinopyrimidine, Triazole, Pyrethroid, Organophosphate and Dicarboximide Crop Protection Products, France/Germany 2000 to 2001.* BASF DocID 2002/1007077. Unpublished. European Crop Protection Association (ECPA), Residue Expert Group. Belgium.

Kennedy, M. and A. Hart (2009). Bayesian Modeling of Measurement Errors and Pesticide Concentration in Dietary Risk Assessments. *Risk Analysis*. **29** (10), 1427–1442.

Kennedy, M., V. Roelofs, C. Anderson, and J. Salazar (2011). A hierarchical Bayesian model for extreme pesticide residues. *Food and Chemical Toxicology*. **49** (1), 222–232.

Kitagawa, G. (1996). Monte Carlo Filter and Smoother for Non-Gaussian Nonlinear State Space Models. *Journal of Computational and Graphical Statistics*. **5** (1), 1–25.

Lallukka, T., M. Lahti-Koski, and M. Ovaskainen (2001). Vegetable and Fruit Consumption and its Determinants in Young Finnish Adults. *Scandinavian Journal of Nutrition*. **45**, 120–126.

Lehman, A. J. and O. G. Fitzhugh (1954). 100-Fold Margin of Safety. *Association for Food and Drug Official US Quarterly Bulletin*. **18**, 33–35.

Liu, J. S. (1996). Nonparametric Hierarchical Bayes via Sequential Imputations. *The Annals of Statistics*. **24** (3), 911–930.

Lo, A. Y. (1984). On a Class of Bayesian Nonparametric Estimates: I. Density Estimates. *The Annals of Statistics*. **12**, 351–357.

Lunn, D., A. Thomas, N. Best, and D. Spiegelhalter (2000). WinBUGS - a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing*. **10** (4), 325–337.

MacEachern, S. N. and P. Müller (1998). Estimating Mixture of Dirichlet Process Models. *Journal of Computational and Graphical Statistics*. **7** (2), 223–238.

MacEachern, S. N., M. Clyde, and J. S. Liu (1999). Sequential Importance Sampling for Nonparametric Bayes Models: The Next Generation. *The Canadian Journal of Statistics*. **27** (2), 251–267.

MacLachlan, D. J. and D. Hamilton (2011). A Review of the Effect of Different Application Rates on Pesticide Residue Levels in Supervised Residue Trials. *Pest Management Science*. **67**, 609–615.

McNamara, C., B. Naddy, D. Rohan, and J. Sexton (2003). Design, Development and Validation of Software for Modelling Dietary Exposure to Food Chemicals and Nutrients. *Food Additives and Contaminants*. **20** (Supplement 1), S8–26.

Metropolis, N. and S. Ulam (1949). The Monte Carlo Method. *Journal of the American Statistical Association*. **44** (247), 335–341.

Metropolis, N., A. Rosenbluth, M. Rosenbluth, and A. Teller (1953). Equation of State Calculations by Fast Computing Machines. *The Journal of Chemical Physics*. **21** (6), 1087–1092.

Miller, R. B. (1980). Bayesian Analysis of the Two-Parameter Gamma Distribution. *Technometrics*. **22** (1), 65–69.

Neal, R. M. (2000). Markov Chain Sampling Methods for Dirichlet Process Mixture Models. *Journal of Computational and Graphical Statistics*. **9** (2), 229–265.

Neal, R. M. (2003). Slice Sampling. *The Annals of Statistics*. **31** (3), 705–767.

Ocké, M., H. van Rossum, and H. Fransen (2007). *Dutch National Food Consumption Survey - Young Children 2005/2006*. Bilthoven, The Netherlands: RIVM.

OECD (2009). *OECD Guidelines for the Testing of Chemicals, Section 5. Test 509: Crop Field Trials*. Available from: http://www.oecd-ilibrary.org/test-no-509-crop-field-trial_5ksb6nhrnk5k.pdf?contentType=/ns/Book&itemId=/content/book/9789264076457-en&containerItemId=/content/serial/20745796&accessItemIds=/content/serial/20745796&mimeType=application/pdf. OECD.

OECD (2011a). *Guidance Document on Crop Field Trials. Version 6*. OECD.

OECD (2011b). *OECD MRL Calculator: Statistical White Paper*. Series of pesticides No. 56, ENV/JM/MONO(2011)2, 1 March 2011. In: Pesticide Publications/Publications on Pesticide Residues. Available from: http://www.oecd.org/env/pesticides. OECD.

Papaspiliopoulos, O. (2008). *A Note on Posterior Sampling from Dirichlet Mixture Models. Technical Report*. Available from: http://www2.warwick.ac.uk/fac/sci/statistics/crism/research/2008/08-20wv2.pdf. Centre for Research in Statistical Methodology, University of Warwick, Coventry.

Papaspiliopoulos, O. and G. O. Roberts (2008). Retrospective Markov Chain Monte Carlo Methods for Dirichlet process Hierarchical Models. *Biometrika.*, 1–18.

Paulo, M., H. van der Voet, M. Jansen, C. ter Braak, and J. van Klaveren (2005). Risk Assessment of Dietary Exposure to Pesticides Using a Bayesian Method. *Pest Management Science*. **61**, 759–766.

Porteous, I., A. Ihler, P. Smyth, and M. Welling, eds. (2006). *Gibbs Sampling for (Coupled) Infinite Mixture Models in the Stick Breaking Representation.* In Proceedings of the 22nd Annual Conference on Uncertainty in Artificial Intelligence (UAI-06). Arlington, VA: AUAI Press.

PRC (2008). *Pesticide Residues Committee Reports 2008. Second Quarter (April to June 2008, Published 9th December 2008).* Available from: http://www.pesticides.gov.uk/prc.asp?id=2536.

PRC (2009). *Pesticide Residues Committee Reports 2008. Fourth Quarter Results (October to December 2008, Published 25 June 2009).* Available from: http://www.pesticides.gov.uk/prc.asp?id=2536.

PRC (2010). *Pesticide Residues Committee Reports 2010. Quarter 1 Results (January to March 2010, Published 8 October 2011).* Available from: http://www.pesticides.gov.uk/Resources/CRD/Migrated-Resources/Documents/Other/2010_Q1_PRC_report.pdf.

PRC (2011a). *Pesticide Residues Committee Reports 2010. Quarter 2 Results (April to June 2010, Published 31 January 2011).* Available from: http://www.pesticides.gov.uk/Resources/CRD/Migrated-Resources/Documents/Other/2010_Q2_Report.pdf.

PRC (2011b). *Pesticide Residues Committee Reports 2010. Quarter 3 Results (July to September 2010, Published 10 March 2011).* Available from: http://www.pesticides.gov.uk/Resources/CRD/Migrated-Resources/Documents/Other/2010_Q3_Report.pdf.

PRC (2011c). *Pesticide Residues Committee Reports 2010. Quarter 4 Results (October to December 2010, Published 16 June 2011).* Available from: http://www.pesticides.gov.uk/Resources/CRD/Migrated-Resources/Documents/Other/2010_Q4_Final.pdf.

Renwick, A. (2002). Pesticide Residue Analysis and its Relationship to Hazard Characterisation (ADI/ARfD) and Intake Estimations (NEDI/NESTI). *Pest Management Science*. **58**, 1073–1082.

Renwick, A. and N. Lazarus (1998). Human Variability and Noncancer Risk Assessment - An Analysis of the Default Uncertainty Factor. *Regulatory Toxicology and Pharmacology*. **27**, 3–20.

Roberts, C. and G. Casella (2005). *Monte Carlo Statistical Methods*. 2nd ed. Springer.

SANCO, D. (2012). *EU Pesticide Database*. http://ec.europa.eu/sanco_pesticides/public/?event=activesubstance.selection. Accessed 10 October 2012.

Sethuraman, J. (1994). A Constructive Definition of Dirichlet Priors. *Statistica Sinica*. **4**, 639–650.

Silverman, B. (1981). Using Kernel Density Estimates to Investigate Multimodality. *Journal of the Royal Statistical Society. Series B (Methodological)*. **43** (1), 97–88.

Slottje, P., J. van der Sluijs, and A. Knol (2008). *Expert Elicitation. Methodological suggestions for its use in environmental health impact assessments.* Report 630004001/2008. Bilthoven, The Netherlands: RIVM.

Tanner, B. and W. Wong (1987). The Calculation of Posterior Distributions by Data Augmentation. *Journal of the American Statistical Association*. **82** (398), 528–550.

Tew, E. (1993). *Determination of Variation in Magnitude and Character of Aldicarb Residues in Potatoes from Temik-Treated Fields: Final Report.* EC-90092: EC-90-117: 41207. Unpublished Report. See also: http://www.fao.org/docrep/W5897E/W5897E00.htm. Rhône-Poulenc Agricultural Company.

Valdez-Flores, C., L. R. Holden, and R. L. Sielken Jr (2002). Generating Single-unit Residue Concentration Distributions Based on Maximum Likelihood Estimation from Composite Data. *Environmetrics*. **13**, 711–724.

Van der Velde-Koerts, T., G. Van Donkersgoed, N. Koopman, and B. Ossendorp (2011). *Revision of Dutch Dietary Risk Assessment Models for Pesticide Authorisation Purposes.* Report 320005006/2010. Bilthoven, The Netherlands: RIVM.

Van Klaveren, J., G. Van Donkersgoed, H. Van der Voet, C. Stephenson, and P. Boon (2009). *Cumulative Exposure Assessment of Triazole Pesticides.* Report 2009.008. Rikilt.

VCP (1998). *Zo eet Nederland. Resultaten van de Voedselconsumptiepeiling 1997-1998.* The Hague, The Netherlands: Voedingscentrum.

Vermeire, T., H. Stevenson, M. Pieters, M. Renne, and B. Slob W. Hakkert (1999). Assessment Factors for Human Health Risk Assessment: A Discussion Paper. *Critical Reviews in Toxicology.* **29** (5), 439–490.

Von Neumann, J., ed. (1951). *Various Yechniques Used in Connection with Random Digits.* 12. In Proceedings of symposium on 'Monte Carlo Method' held June-July 1949 in Los Angeles. Los Angeles: National Bureau of Standards, Applied Math. Chap. 13, 36–38.

Walker, S. G. (2007). Sampling the Dirichlet Mixture Model with Slices. *Communications in Statistics - Simulation and Computation.* **36**, 45–54.

WHO (1997). *Guidelines for Predicting Dietary Intake of Pesticide Residues.* World Health Organisation. Global Environment Monitoring System / Food Contamination Monitoring and Assessment Programme (GEMS /Food).

WHO (2012). *GEMS/Food Cluster Diets.* http://www.who.int/foodsafety/chem/gems/en/index1.html. Accessed 10 October 2012.

Xu, X., R. A. Murray, J. D. Salazar, and K. Hyder (2008). The Temporal Pattern of Captan Residues on Apple Leaves and Fruit under Field Conditions in Relation to Weather and Canopy Structure. *Pest Management Science.* **64**, 565–578.

Appendix A

Unit Residue Data

Table A.1 *– Unit Field Trial data used for DPMN model.*

Data Set	Commodity	Pesticide	n	n >LOD	Rounding method*	Reference
1	Blackcurrant	Vinclozolin	120	120	-	Ambrus (2006)
2	Cabbage	Triazophos	130	130	4 dp	Ambrus (2006)
3	Cabbage	Chlorpyrifos	120	120	4 dp	Ambrus (2006)
4	Cabbage	Profenofos	120	120	3 dp	Ambrus (2006)
5	Cabbage	Profenofos	120	120	3 dp	Ambrus (2006)
6	Cherry	Chlorpyrifos	120	120	3 dp	Ambrus (2006)
7	Cherry	Lambda-cyhalothrin	120	120	3 dp	Ambrus (2006)
8	Cherry	Phosalone	120	120	5 dp	Ambrus (2006)
9	Chicory	Tolclofos-methyl	121	121	3 dp	Ambrus (2006)
10	Cucumber	Pirimyphos-methyl	120	120	-	Ambrus (2006)
11	Cucumber	Acetamiprid	120	115	3 dp	Ambrus (2006)
12	Cucumber	Pymetrozine	120	106	3 dp	Ambrus (2006)
13	Cucumber	Chlorothalonil	130	130	4 dp	Ambrus (2006)
14	Cucumber	Tolylfluanid	120	120	-	Ambrus (2006)
15	Cucumber	Pirimiphos-methyl	130	130	-	Ambrus (2006)
16	Grape	Chlorpyrifos	120	114	3 dp	Ambrus (2006)
17	Grape	Chlorpyrifos	120	120	2 dp	Ambrus (2006)
18	Grape	Chlorpyrifos-methyl	133	133	-	Ambrus (2006)
19	Grape	Vinclozolin	120	120	3 dp	Ambrus (2006)
20	Grape	Folpet	120	120	-	Ambrus (2006)
21	Kale	Chlorpyrifos-methyl	120	120	3 dp	Ambrus (2006)
22	Kale	Chlorothalonil	121	121	-	Ambrus (2006)
23	Kale	Profenofos	160	160	-	Ambrus (2006)
24	Lettuce	Indoxacarb	121	121	3 dp	Ambrus (2006)
25	Lettuce	Vinclozolin	121	120	2 dp	Ambrus (2006)

Continued on next page

Table A.1 – continued from previous page

Data Set	Commodity	Pesticide	n	n >LOD	Rounding method*	Reference
26	Lettuce	Procymidone	120	120	3 dp	Ambrus (2006)
27	Lettuce	Alphamethrin	120	120	3 dp	Ambrus (2006)
28	Lettuce	Alphamethrin	120	120	3 dp	Ambrus (2006)
29	Lettuce	Chlorothalonil	120	120	2 dp	Ambrus (2006)
30	Lettuce	Pirimiphos-methyl	130	130	3 dp	Ambrus (2006)
31	Mango	Parathion-methyl	153	153	3 dp	Ambrus (2006)
32	Mango	Deltamethrin	120	120	3 dp	Ambrus (2006)
33	Mango	Cypermethrin	135	135	-	Ambrus (2006)
34	Mango	Phenthoate	127	127	4 dp	Ambrus (2006)
35	Mango	Phenthoate	130	130	5 dp	Ambrus (2006)
36	Mango	Prophiofos	176	176	-	Ambrus (2006)
37	Papaya	Diazinon	66	66	2 dp	Ambrus (2006)
38	Papaya	Methidathion	136	136	2 dp	Ambrus (2006)
39	Papaya	Diazinon	122	122	3 dp	Ambrus (2006)
40	Papaya	Deltamethrin	130	130	4 dp	Ambrus (2006)
41	Squash	Methidathion	128	128	-	Ambrus (2006)
42	Strawberry	Procymidone	120	120	3 dp	Ambrus (2006)
43	Strawberry	Procymidone	141	141	-	Ambrus (2006)
44	Strawberry	Endosulfan	130	130	4 dp	Ambrus (2006)
45	Strawberry	Procymidone	141	141	5 dp	Ambrus (2006)
46	Zucchini	Azoxystrobin	120	120	3 dp	Ambrus (2006)
47	Zucchini	Azoxystrobin	120	120	3 dp	Ambrus (2006)
48	Kale	Indoxacarb	108	90	3 dp	Ambrus (2006)
49	Chicory	Tolclofos-methyl	121	121	3 dp	Ambrus (2006)
50	Cucumber	Vinclozolin	120	120	3 dp	Ambrus (2006)
51	Cherry	Dimethoate	120	120	3 dp	Ambrus (2006)
52	Apple	Chlorpyrifos-methyl	319	319	-	Ambrus (1995)
53	Apple	Phosphamidon	108	108	2 dp	Ambrus (1979)
54	Kiwi	Vinclozolin	209	209	2 dp	Holland and Malcolm (2002)
55	Potato	Aldicarb	100	100	2 dp	Tew (1993)
56	Potato	Aldicarb	100	100	2 dp	Tew (1993)
57	Potato	Aldicarb	79	79	2 dp	Tew (1993)
58	Potato	Aldicarb	100	100	2 dp	Tew (1993)
59	Potato	Aldicarb	100	100	2 dp	Tew (1993)
60	Potato	Aldicarb	100	100	2 dp	Tew (1993)
61	Potato	Aldicarb	100	100	2 dp	Tew (1993)
62	Potato	Aldicarb	100	100	2 dp	Tew (1993)
63	Potato	Aldicarb	100	100	2 dp	Tew (1993)
64	Potato	Aldicarb	100	100	2 dp	Tew (1993)
65	Potato	Aldicarb	100	100	2 dp	Tew (1993)
66	Grape	Dicarb-oximide	120	120	2 sf	Kaethner (2001b)
67	Grape	Dicarb-oximide	120	120	2 sf	Kaethner (2001b)
68	Grape	Dicarb-oximide	120	120	2 sf	Kaethner (2001b)

Continued on next page

Table A.1 – continued from previous page

Data Set	Commodity	Pesticide	n	n >LOD	Rounding method*	Reference
69	Grape	Dicarb-oximide	120	120	2 sf	Kaethner (2001b)
70	Lettuce	Dicarb-oximide	120	120	2 sf	Kaethner (2001a)
71	Lettuce	Dicarb-oximide	120	120	2 sf	Kaethner (2001a)
72	Lettuce	Dicarb-oximide	120	120	2 sf	Kaethner (2001a)
73	Lettuce	Dicarb-oximide	120	120	2 sf	Kaethner (2001a)
74	Peach	Diazinon	200	200	3 dp	Valdez-Flores et al. (2002)
75	Apple	Captan	348	78	0.0001	Xu et al. (2008)

* *When the data were rounded to n_d decimal places, this is represented by n_d dp where $n_d \leq 5$ and when the data were rounded to n_s significant figures, this is represented by n_s sf. When data was rounded above 5 decimal places, rounding error was ignored. These data sets are indicated by '-'.*

Table A.2 *– Unit Market Survey data used to obtain weakly informative prior distributions for the bivariate mixture model in Chapter 3. The data set is a subset from the data set reported by Hill and Reynolds (2002). Data sets 31, 32, 36, 53 and 63 were excluded because >50% of the observations had residue levels below the LOD.*

Data Set	Commodity	Pesticide	n	n >LOD	LOD
1	Apple	Carbaryl	108	108	0.001
2	Apple	Carbaryl	95	78	0.01
3	Apple	Carbaryl	100	90	0.01
4	Apple	Phosalone	100	100	0.001
5	Apple	Phosalone	100	100	0.001
6	Apple	Chlorpyrifos	110	108	0.001
7	Apple	Chlorpyrifos	110	103	0.001
8	Apple	Carbaryl	100	100	0.001
9	Apple	Carbaryl	100	100	0.01
10	Apple	Chlorpyrifos	100	100	0.001
11	Apple	Carbaryl	100	99	0.001
12	Banana	Chlorpyrifos	100	100	0.0001
13	Banana	Chlorpyrifos	100	93	0.0001
14	Kiwi	Phosmet	100	98	0.001
15	Kiwi	Parathion-methyl	100	99	0.001
16	Kiwi	Parathion-methyl	100	100	0.001
17	Kiwi	Quinalphos	100	91	0.001
18	Kiwi	Diazinon	100	97	0.001
19	Orange	Imazalil	100	100	0.001
20	Orange	Imazalil	110	109	0.001
21	Orange	Chlorpyrifos	100	88	0.001
22	Orange	Imazalil	100	99	0.001
23	Orange	Imazalil	100	92	0.001

Continued on next page

Table A.2 – continued from previous page

Data Set	Commodity	Pesticide	n	n >LOD	LOD
24	Peach	Dimethoate	100	89	0.001
25	Peach	Carbaryl	100	71	0.01
26	Peach	Carbaryl	100	68	0.01
27	Peach	Methamidophos	100	65	0.001
28	Peach	Phosalone	100	90	0.001
29	Pear	Phosalone	110	75	0.001
30	Pear	Phosalone	100	100	0.001
33	Pear	Carbaryl	110	95	0.001
34	Pear	Carbaryl	100	86	0.001
35	Plum	Chlorpyrifos	100	74	0.001
37	Plum	Phosalone	100	81	0.001
38	Plum	Pirimicarb	100	59	0.001
39	Plum	Phosalone	100	100	0.001
40	Plum	Acephate	100	98	0.001
41	Plum	Dimethoate	100	65	0.001
42	Plum	Pirimiphos-methyl	100	99	0.001
43	Plum	Fenitrothion	100	99	0.001
44	Plum	Acephate	101	101	0.001
45	Potato	Aldicarb	100	81	0.001
46	Potato	Aldicarb	100	84	0.001
47	Potato	Aldicarb	100	67	0.001
48	Potato	Aldicarb	100	72	0.001
49	Potato	Aldicarb	100	94	0.001
50	Potato	Aldicarb	100	99	0.001
51	Potato	Aldicarb	100	85	0.001
52	Potato	Aldicarb	100	94	0.001
54	Potato	Aldicarb	100	100	0.001
55	Tomato	Methamidophos	100	95	0.001
56	Tomato	Formetanate	100	94	0.001
57	Tomato	Methamidophos	100	62	0.001
58	Celery	Tolclofos-methyl	40	40	0.001
59	Celery	Heptenophos	40	40	0.001
60	Celery	Disulfoton	40	39	0.001
61	Celery	Disulfoton	40	40	0.001
62	Celery	Phorate	40	40	0.001
64	Celery	Chlorpyrifos	40	40	0.01
65	Celery	Chlorpyrifos	40	40	0.001
66	Celery	Chlorpyrifos	40	40	0.001

Appendix B

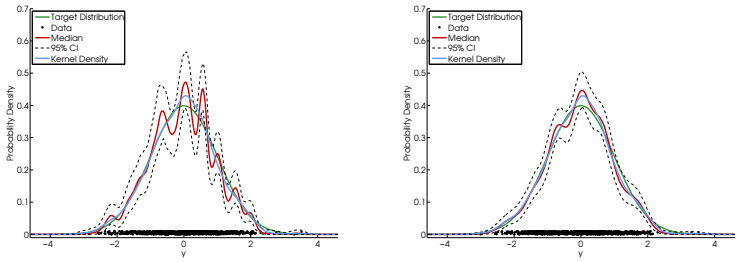
Validation Studies for Chapter 4

B.1 DPMN model output when $\gamma = 10$

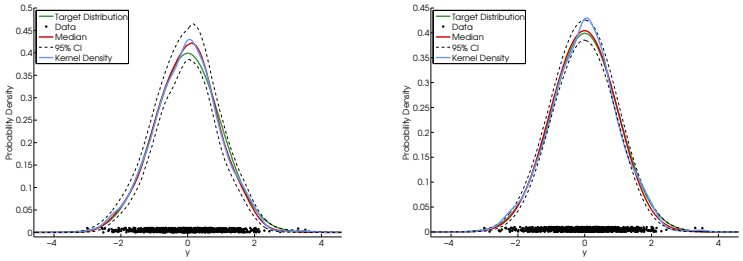
B.1.1 Normal Distribution

Figure B.1 – *Output of DPMN model using a Normal target distribution with* $\gamma = 10$.

(a) $\kappa = 0.1$ **(b)** $\kappa = 0.2$



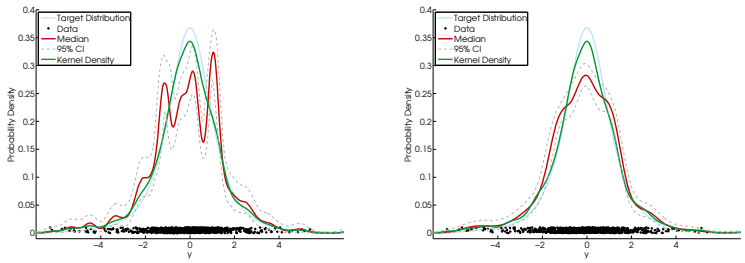
(c) $\kappa = 0.4$ **(d)** $\kappa = 0.8$



B.1.2 Student's t Distribution

Figure B.2 – *Output of DPMN model using a Student-t target distribution with* $\gamma = 10$.

(a) $\kappa = 0.1$ **(b)** $\kappa = 0.2$



(c) $\kappa = 0.4$ **(d)** $\kappa = 0.8$

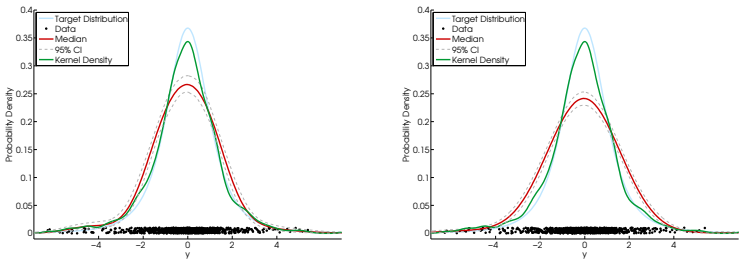
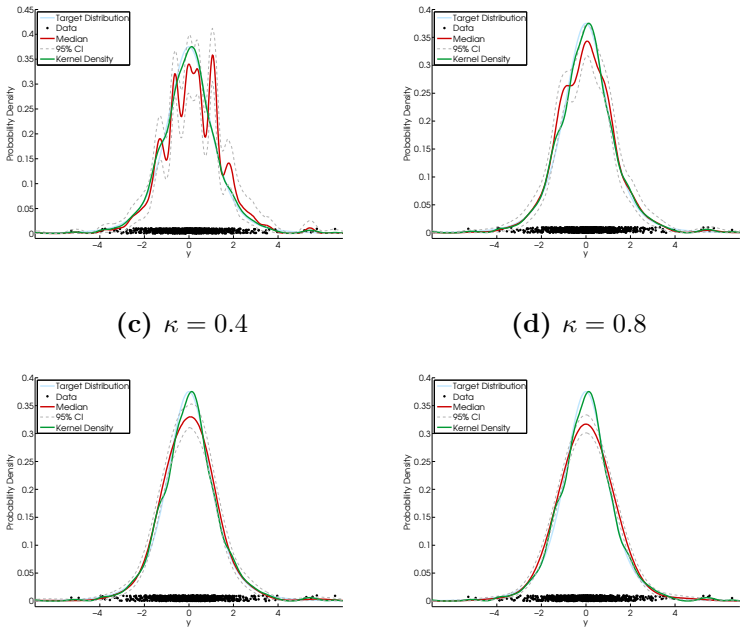


Figure B.3 – *Output of DPMN model using a Student-$t_{\nu=4}$ target distribution with $\gamma = 10$.*

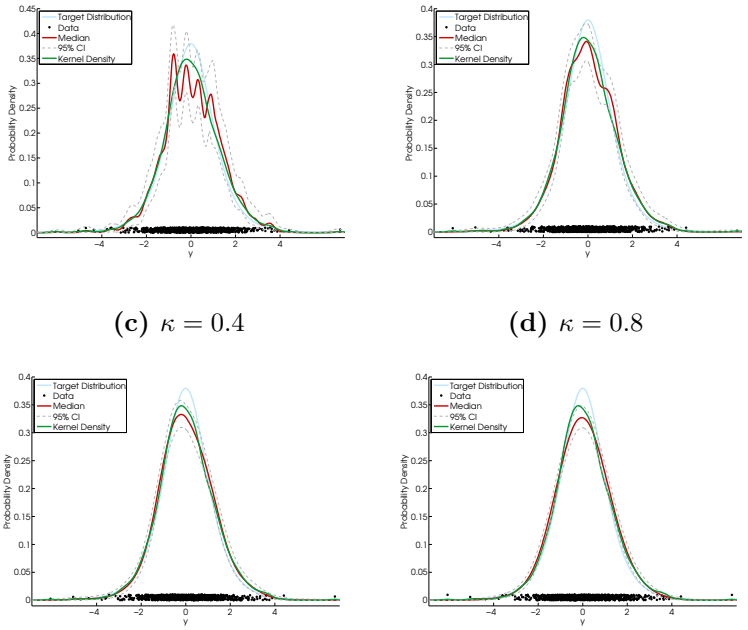
(a) $\kappa = 0.1$ **(b)** $\kappa = 0.2$



(c) $\kappa = 0.4$ **(d)** $\kappa = 0.8$

Figure B.4 – *Output of DPMN model using a Student-$t_{\nu=5}$ target distribution with $\gamma = 10$.*

(a) $\kappa = 0.1$ **(b)** $\kappa = 0.2$

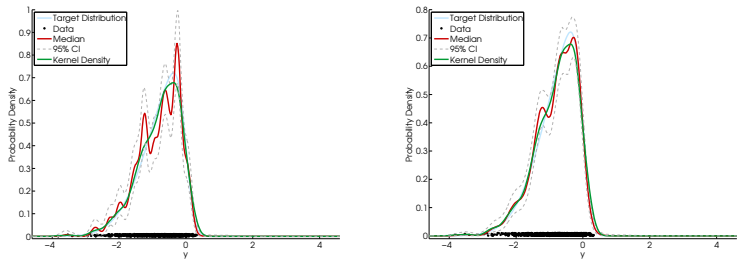


(c) $\kappa = 0.4$ **(d)** $\kappa = 0.8$

B.1.3 Skew-Normal Distribution

Figure B.5 – *Output of DPMN model using a Skew Normal target distribution with* $\lambda = -5$ *and* $\gamma = 10$.

(a) $\kappa = 0.1$ **(b)** $\kappa = 0.2$



(c) $\kappa = 0.4$ **(d)** $\kappa = 0.8$

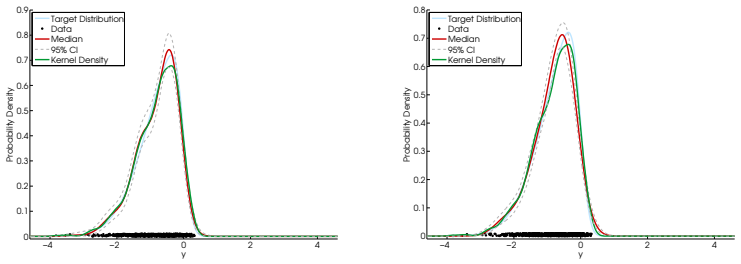
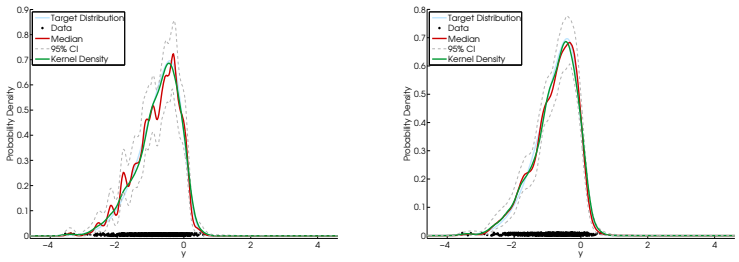


Figure B.6 – *Output of DPMN model using a Skew Normal target distribution with $\lambda = -4$ and $\gamma = 10$.*

(a) $\kappa = 0.1$ **(b)** $\kappa = 0.2$



(c) $\kappa = 0.4$ **(d)** $\kappa = 0.8$

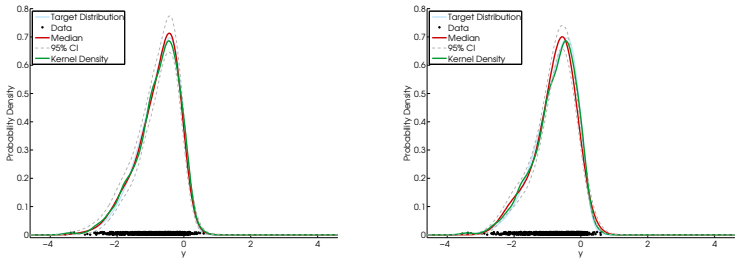
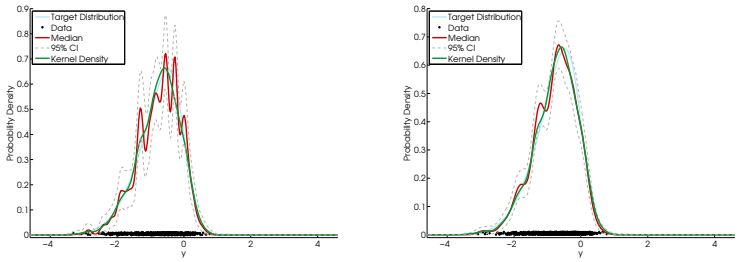


Figure B.7 – *Output of DPMN model using a Skew Normal target distribution with $\lambda = -3$ and $\gamma = 10$.*

(a) $\kappa = 0.1$ **(b)** $\kappa = 0.2$



(c) $\kappa = 0.4$ **(d)** $\kappa = 0.8$

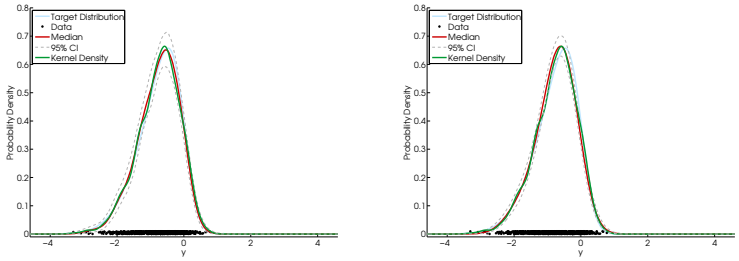
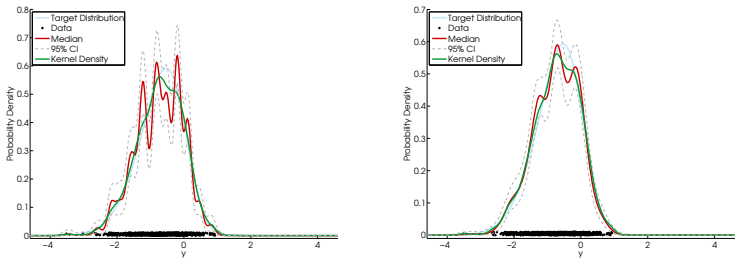


Figure B.8 – *Output of DPMN model using a Skew Normal target distribution with $\lambda = -2$ and $\gamma = 10$.*

(a) $\kappa = 0.1$ **(b)** $\kappa = 0.2$



(c) $\kappa = 0.4$ **(d)** $\kappa = 0.8$

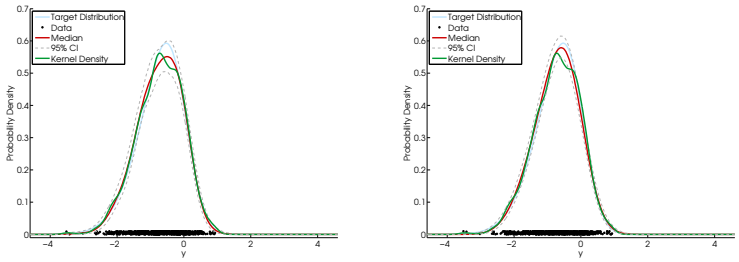
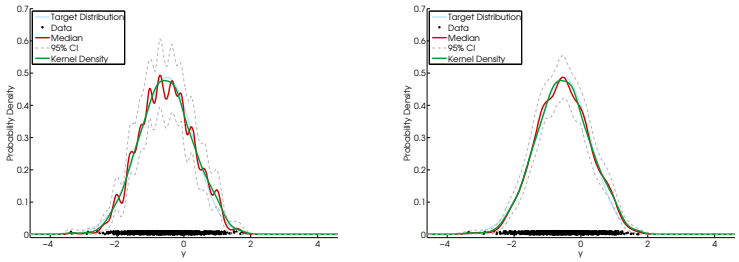


Figure B.9 – *Output of DPMN model using a Skew Normal target distribution with $\lambda = -1$ and $\gamma = 10$.*

(a) $\kappa = 0.1$ **(b)** $\kappa = 0.2$



(c) $\kappa = 0.4$ **(d)** $\kappa = 0.8$

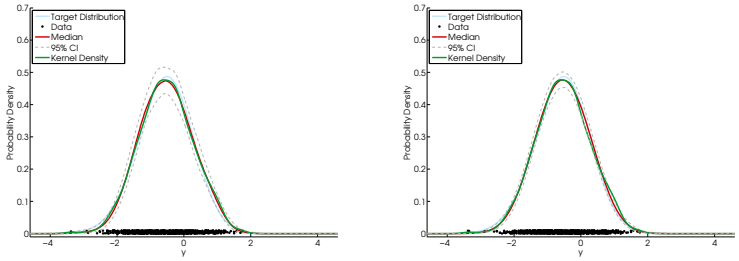
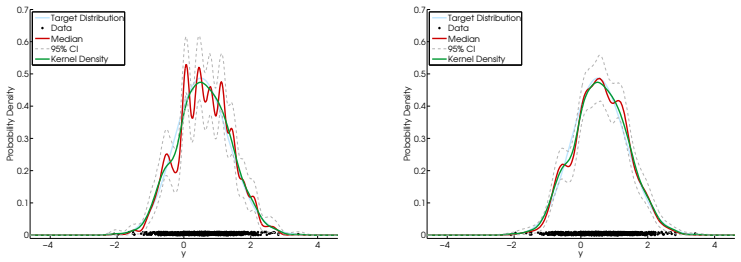


Figure B.10 – *Output of DPMN model using a Skew Normal target distribution with* $\lambda = 1$ *and* $\gamma = 10$.

(a) $\kappa = 0.1$ **(b)** $\kappa = 0.2$



(c) $\kappa = 0.4$ **(d)** $\kappa = 0.8$

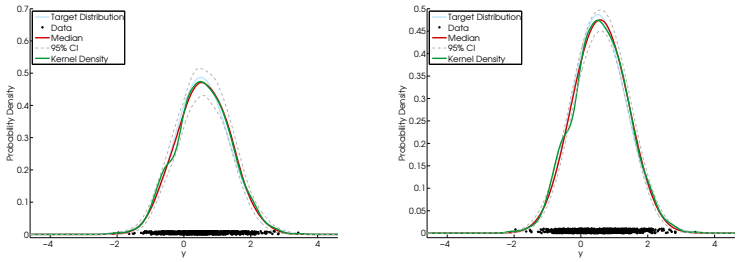
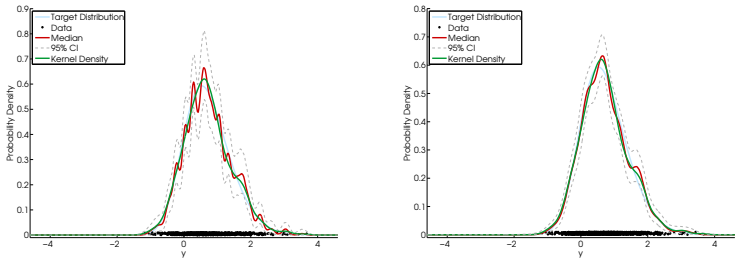


Figure B.11 – *Output of DPMN model using a Skew Normal target distribution with* $\lambda = 2$ *and* $\gamma = 10$.

(a) $\kappa = 0.1$ **(b)** $\kappa = 0.2$



(c) $\kappa = 0.4$ **(d)** $\kappa = 0.8$

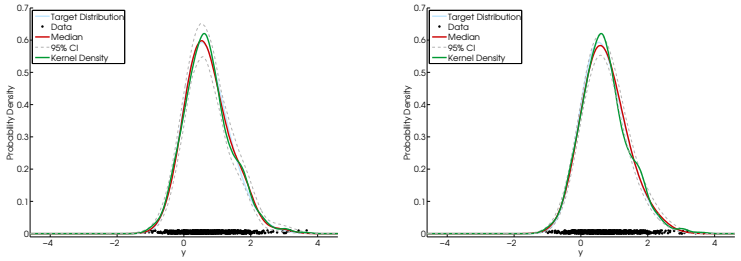
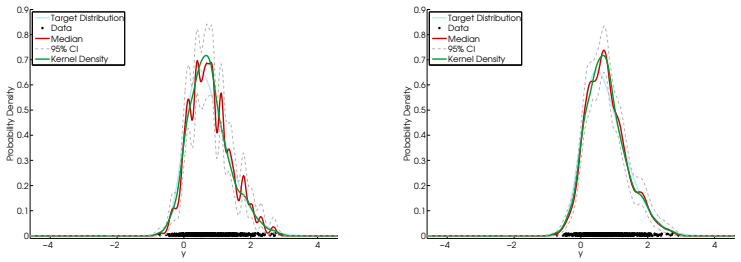


Figure B.12 – *Output of DPMN model using a Skew Normal target distribution with $\lambda = 3$ and $\gamma = 10$.*

(a) $\kappa = 0.1$ **(b)** $\kappa = 0.2$



(c) $\kappa = 0.4$ **(d)** $\kappa = 0.8$

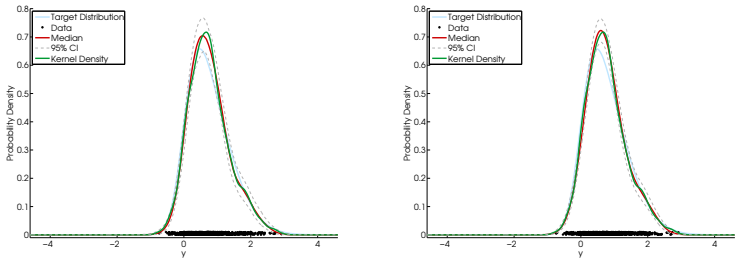
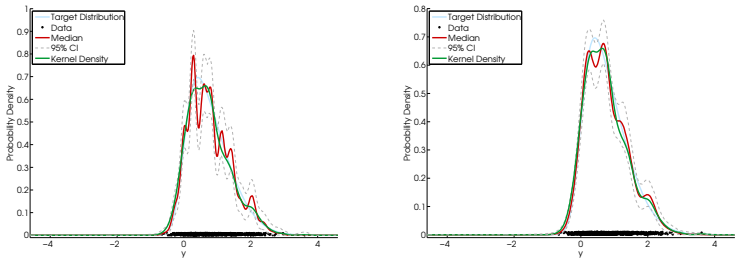


Figure B.13 – *Output of DPMN model using a Skew Normal target distribution with $\lambda = 4$ and $\gamma = 10$.*

(a) $\kappa = 0.1$ **(b)** $\kappa = 0.2$



(c) $\kappa = 0.4$ **(d)** $\kappa = 0.8$

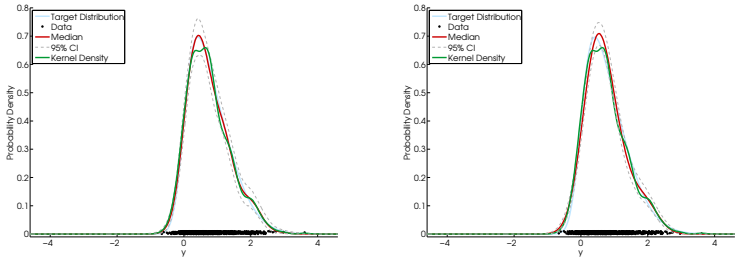
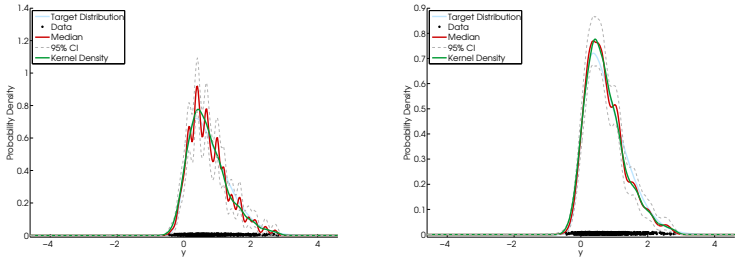
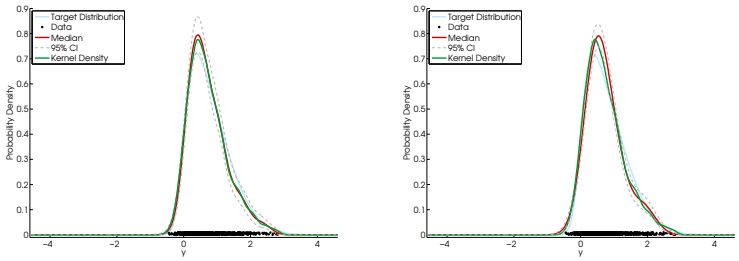


Figure B.14 – *Output of DPMN model using a Skew Normal target distribution with $\lambda = 5$ and $\gamma = 10$.*

(a) $\kappa = 0.1$ **(b)** $\kappa = 0.2$



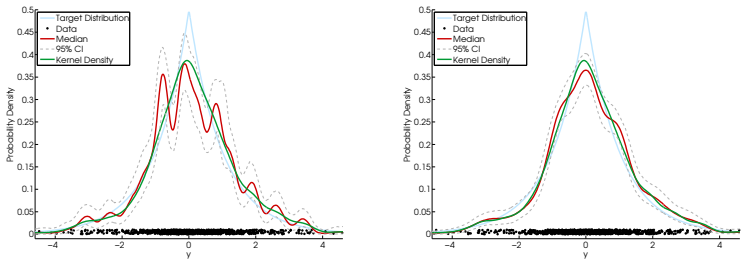
(c) $\kappa = 0.4$ **(d)** $\kappa = 0.8$



B.1.4 Exponential Power Distribution

Figure B.15 – *Output of DPMN model using an Exponential Power target distribution with* $\lambda = 1$ *and* $\gamma = 10$.

(a) $\kappa = 0.1$ **(b)** $\kappa = 0.2$



(c) $\kappa = 0.4$ **(d)** $\kappa = 0.8$

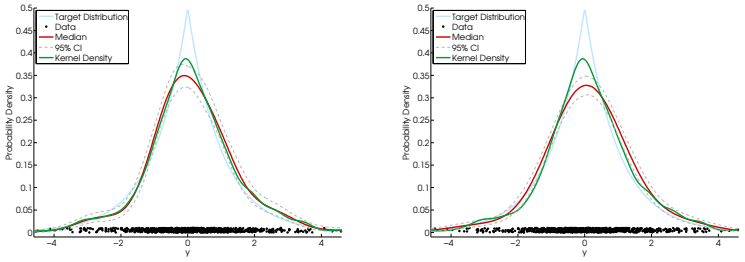
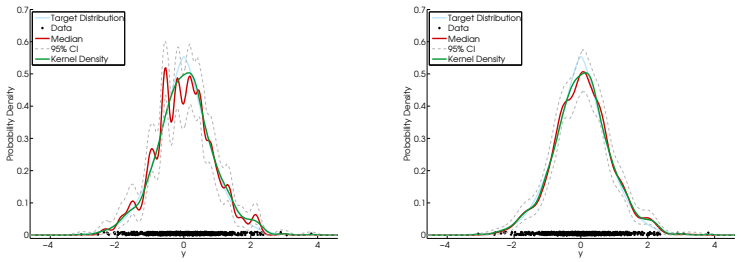


Figure B.16 – *Output of DPMN model using an Exponential Power target distribution with* $\lambda = 1.5$ *and* $\gamma = 10$.

(a) $\kappa = 0.1$ **(b)** $\kappa = 0.2$



(c) $\kappa = 0.4$ **(d)** $\kappa = 0.8$

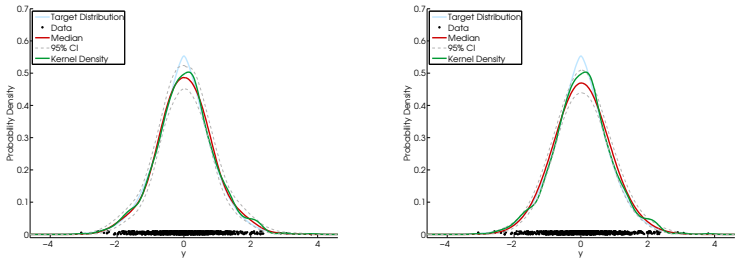
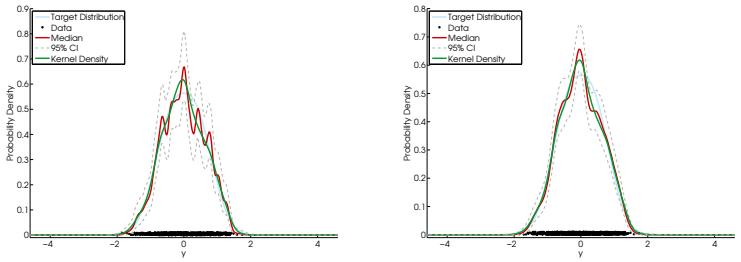


Figure B.17 – *Output of DPMN model using an Exponential Power target distribution with* $\lambda = 2.5$ *and* $\gamma = 10$.

(a) $\kappa = 0.1$ **(b)** $\kappa = 0.2$



(c) $\kappa = 0.4$ **(d)** $\kappa = 0.8$

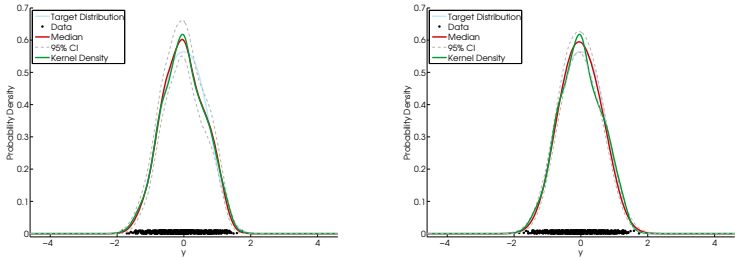
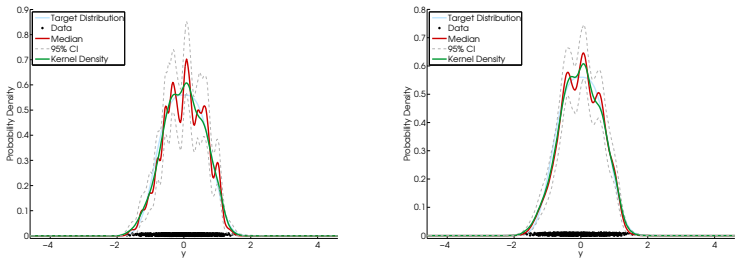


Figure B.18 – *Output of DPMN model using an Exponential Power target distribution with $\lambda = 3$ and $\gamma = 10$.*

(a) $\kappa = 0.1$ **(b)** $\kappa = 0.2$



(c) $\kappa = 0.4$ **(d)** $\kappa = 0.8$

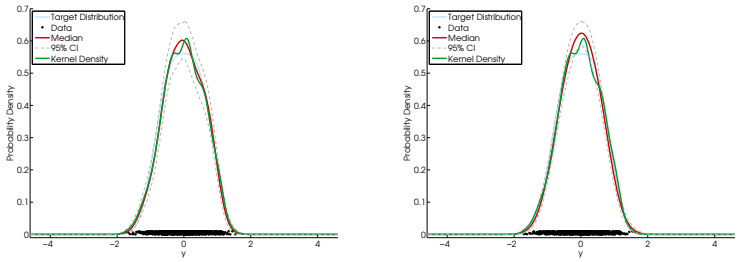
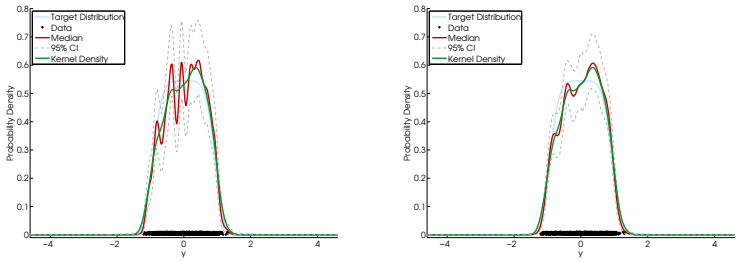
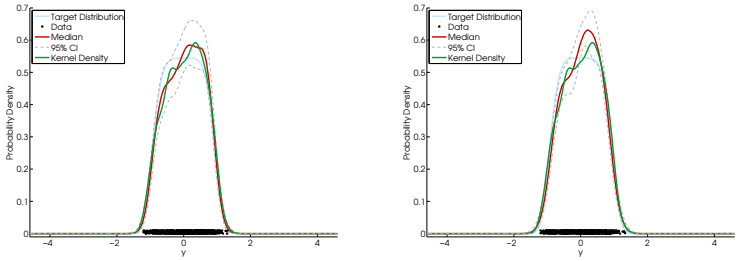


Figure B.19 – *Output of DPMN model using an Exponential Power target distribution with $\lambda = 5$ and $\gamma = 10$.*

(a) $\kappa = 0.1$ **(b)** $\kappa = 0.2$



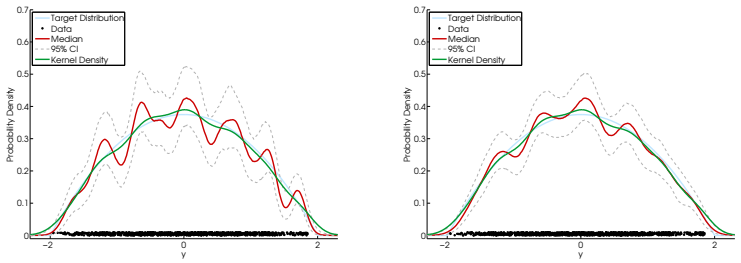
(c) $\kappa = 0.4$ **(d)** $\kappa = 0.8$



B.1.5 Beta Distribution

Figure B.20 – *Output of DPMN model using a Beta*$(2, 2)$ *target distribution with* $\gamma = 10$.

(a) $\kappa = 0.1$ **(b)** $\kappa = 0.2$



(c) $\kappa = 0.4$ **(d)** $\kappa = 0.8$

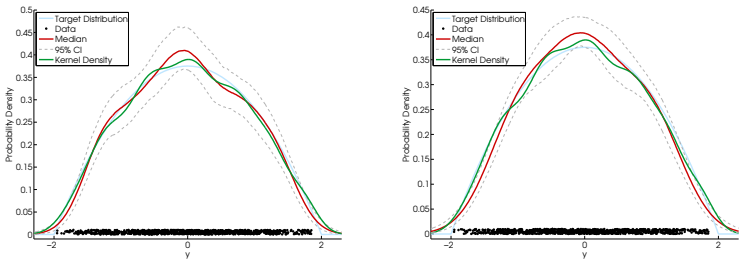
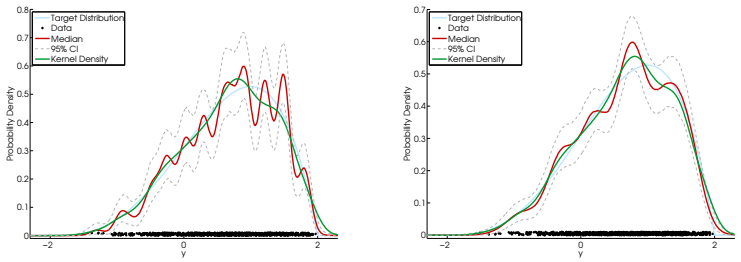


Figure B.21 – *Output of DPMN model using a Beta*$(4, 2)$ *target distribution with* $\gamma = 10$.

(a) $\kappa = 0.1$ **(b)** $\kappa = 0.2$



(c) $\kappa = 0.4$ **(d)** $\kappa = 0.8$

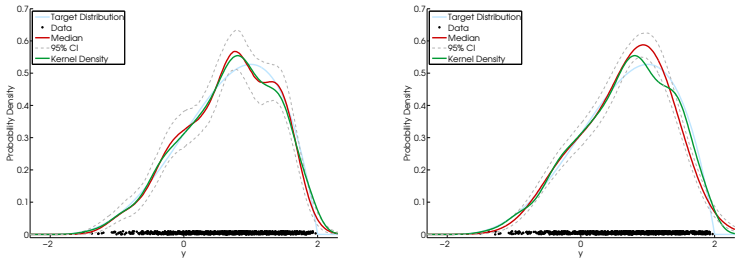
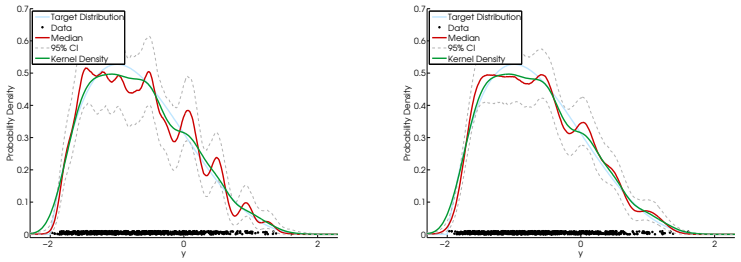
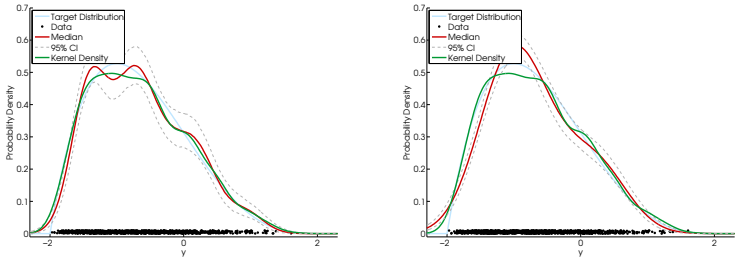


Figure B.22 – *Output of DPMN model using a Beta*$(2, 4)$ *target distribution with* $\gamma = 10$.

(a) $\kappa = 0.1$ **(b)** $\kappa = 0.2$



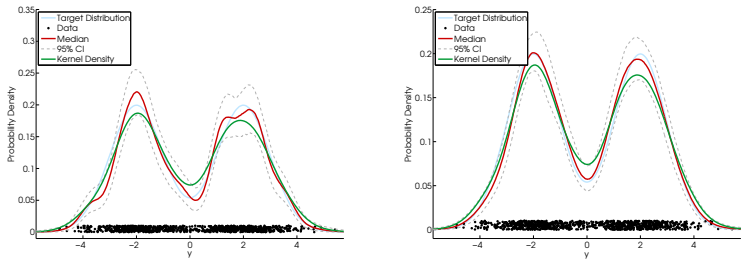
(c) $\kappa = 0.4$ **(d)** $\kappa = 0.8$



B.1.6 Mixture of Two Normal Distributions

Figure B.23 – *Output of DPMN model using a Normal mixture target distribution,* $p(y) = p \times \mathcal{N}(y; -1, 0.5^2) + (1 - p) \times \mathcal{N}(y; 0.5, 0.25^2)$, *with* $p = 0.5$ *and* $\gamma = 10$.

(a) $\kappa = 0.1$ **(b)** $\kappa = 0.2$



(c) $\kappa = 0.4$ **(d)** $\kappa = 0.8$

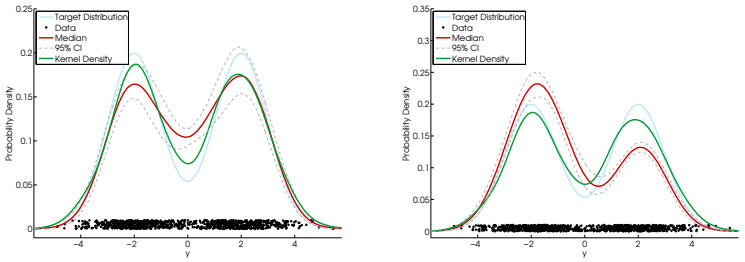
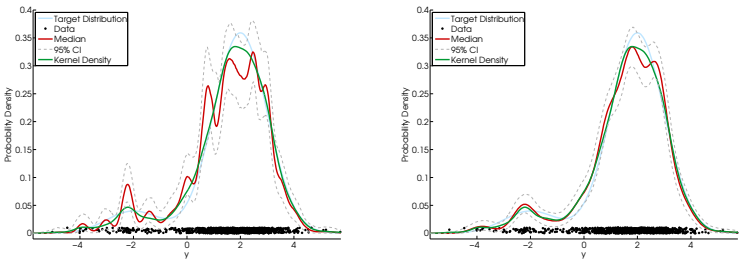


Figure B.24 – *Output of DPMN model using a Normal mixture target distribution,* $p(y) = p \times \mathcal{N}(y; -1, 0.5^2) + (1 - p) \times \mathcal{N}(y; 0.5, 0.25^2)$, *with* $p = 0.1$ *and* $\gamma = 10$.

(a) $\kappa = 0.1$ **(b)** $\kappa = 0.2$



(c) $\kappa = 0.4$ **(d)** $\kappa = 0.8$

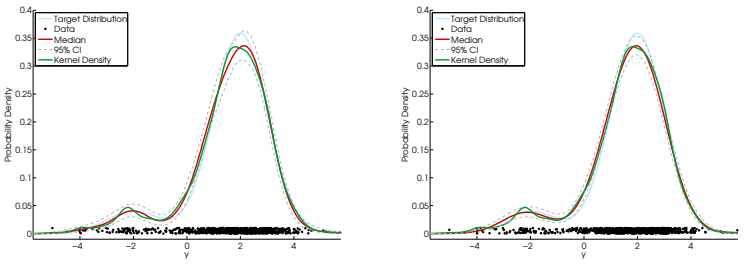
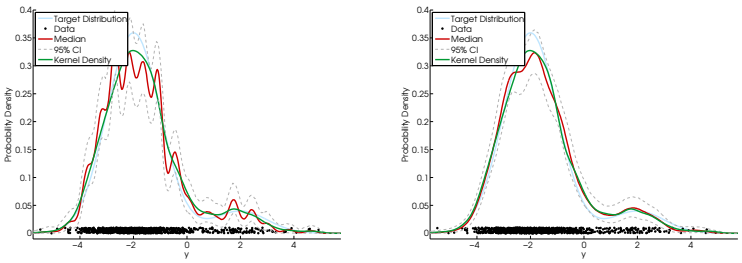


Figure B.25 – *Output of DPMN model using a Normal mixture target distribution,* $p(y) = p \times \mathcal{N}(y; -1, 0.5^2) + (1 - p) \times \mathcal{N}(y; 0.5, 0.25^2)$, *with* $p = 0.9$ *and* $\gamma = 10$.

(a) $\kappa = 0.1$ **(b)** $\kappa = 0.2$



(c) $\kappa = 0.4$ **(d)** $\kappa = 0.8$

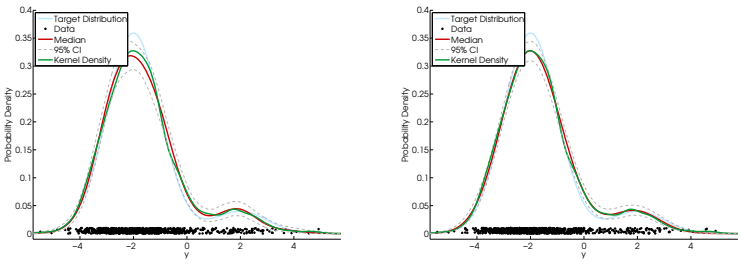
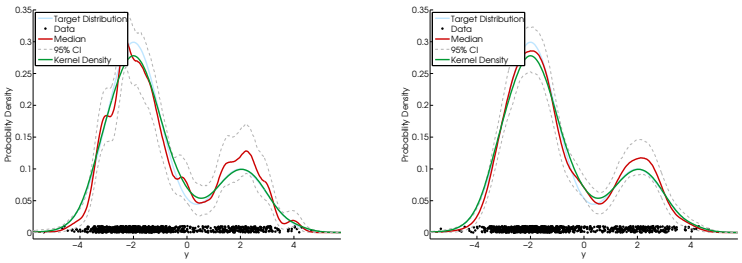


Figure B.26 – *Output of DPMN model using a Normal mixture target distribution,* $p(y) = p \times \mathcal{N}(y; -1, 0.5^2) + (1 - p) \times \mathcal{N}(y; 0.5, 0.25^2)$*, with* $p = 0.75$ *and* $\gamma = 10$*.*

(a) $\kappa = 0.1$ **(b)** $\kappa = 0.2$



(c) $\kappa = 0.4$ **(d)** $\kappa = 0.8$

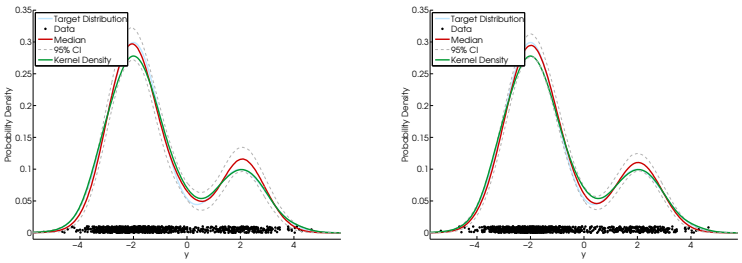
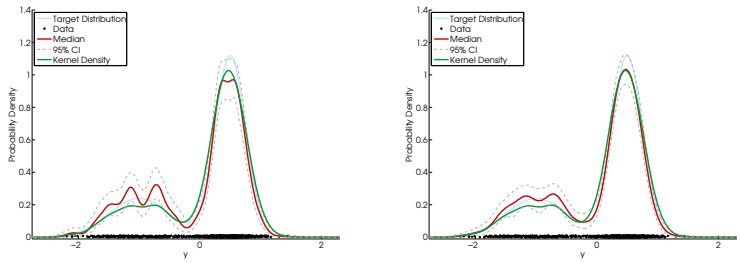
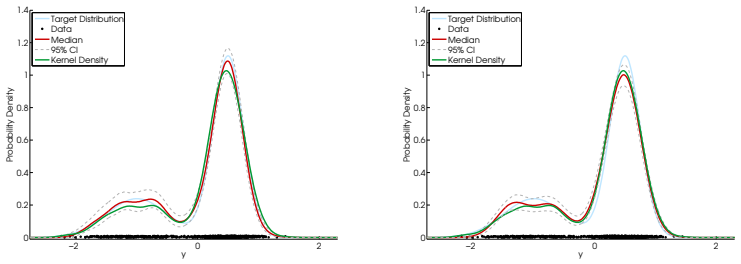


Figure B.27 – *Output of DPMN model using a Normal mixture target distribution,* $p(y) = p \times \mathcal{N}(y; -1, 0.5^2) + (1 - p) \times \mathcal{N}(y; 0.5, 0.25^2)$, *with* $p = 0.3$ *and* $\gamma = 10$.

(a) $\kappa = 0.1$ **(b)** $\kappa = 0.2$



(c) $\kappa = 0.4$ **(d)** $\kappa = 0.8$

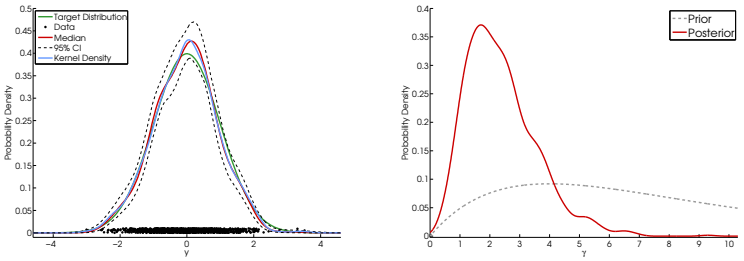


B.2 DPMN model output when γ is inferred from the data.

B.2.1 Normal Distribution

Figure B.28 – *Output of DPMN model using a Normal target distribution with uncertain γ.*

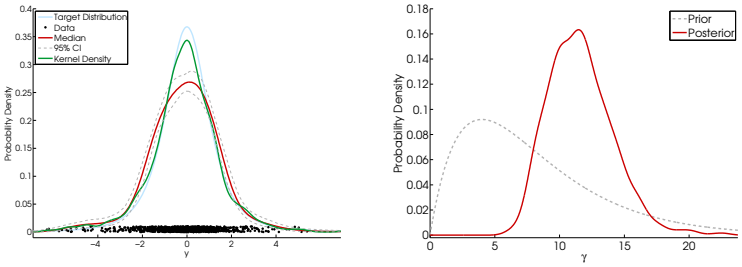
(a) Normal distribution **(b)** γ



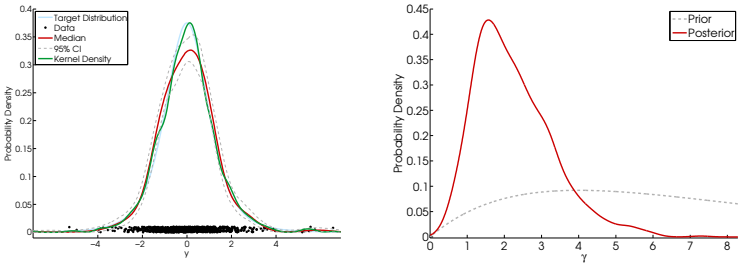
B.2.2 Student's t Distribution

Figure B.29 – *Output of DPMN model using a Student-t target distribution with uncertain γ.*

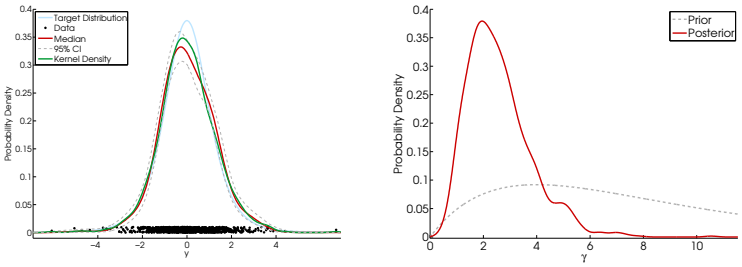
(a) Student-$t_{\nu=3}$ **(b)** Student-$t_{\nu=3}$



(c) Student-$t_{\nu=4}$ **(d)** Student-$t_{\nu=4}$



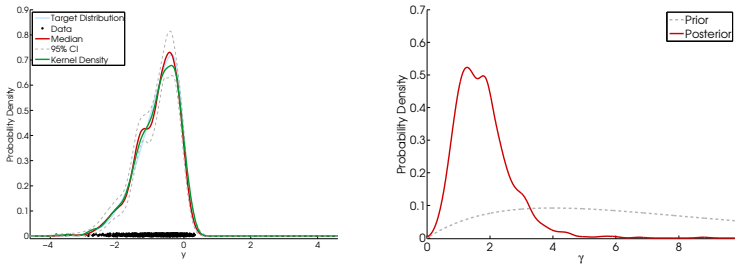
(e) Student-$t_{\nu=5}$ **(f)** Student-$t_{\nu=5}$



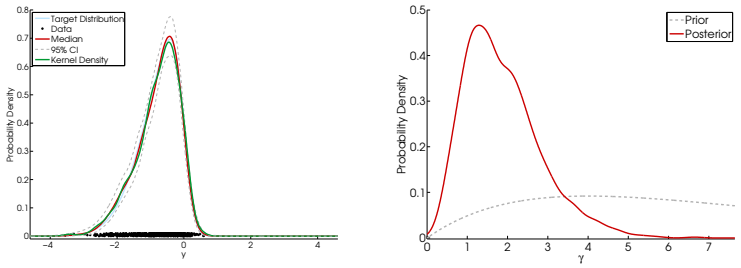
B.2.3 Skew-Normal Distribution

Figure B.30 – *Output of DPMN model using a Skew Normal target distribution with uncertain γ.*

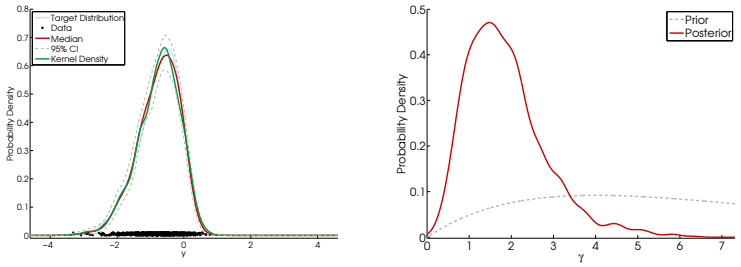
(a) $\lambda = -5$ **(b)** $\lambda = -5$



(c) $\lambda = -4$ **(d)** $\lambda = -4$



(e) $\lambda = -3$ **(f)** $\lambda = -3$



(g) $\lambda = -2$ **(h)** $\lambda = -2$

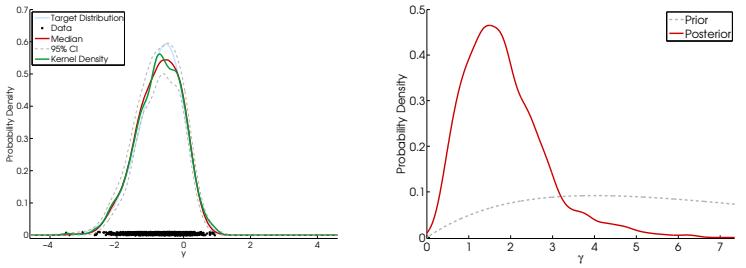
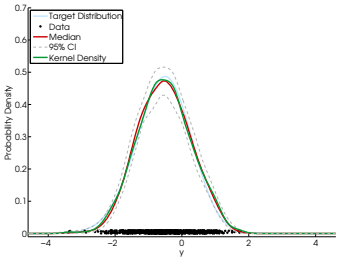
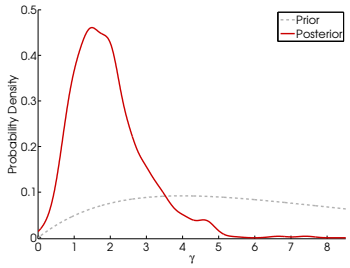


Figure B.30 – *Output of DPMN model using a Skew Normal target distribution with uncertain γ - Continued.*

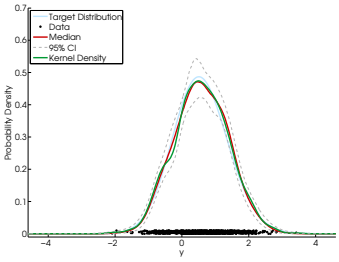
(i) $\lambda = -1$



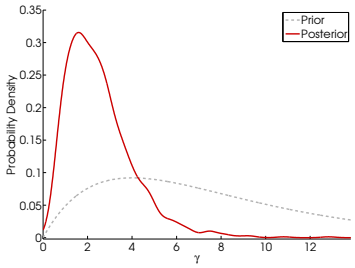
(j) $\lambda = -1$



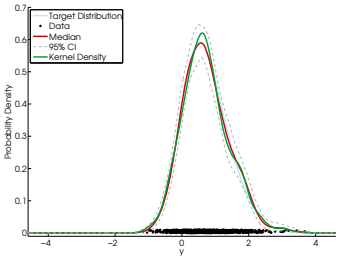
(k) $\lambda = 1$



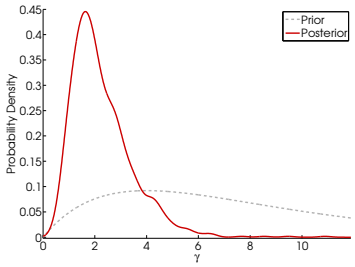
(l) $\lambda = 1$



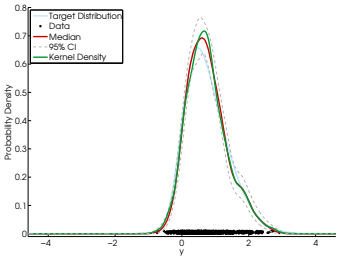
(m) $\lambda = 2$



(n) $\lambda = 2$



(o) $\lambda = 3$



(p) $\lambda = 3$

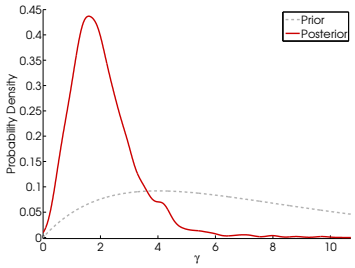
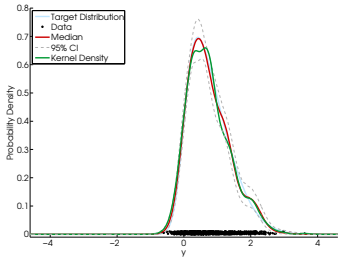
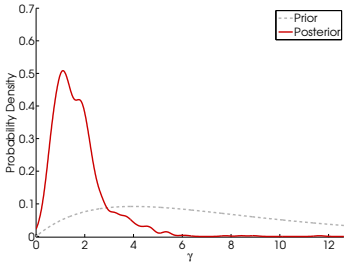


Figure B.30 – *Output of DPMN model using a Skew Normal target distribution with uncertain γ - Continued.*

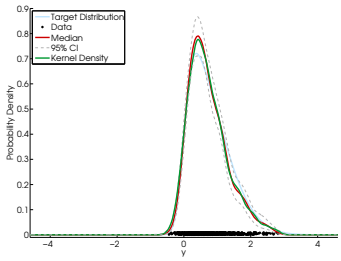
(q) $\lambda = 4$



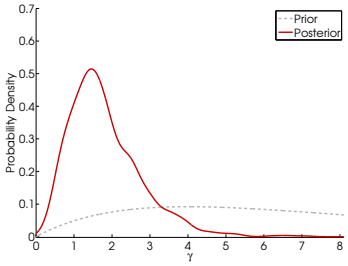
(r) $\lambda = 4$



(s) $\lambda = 5$



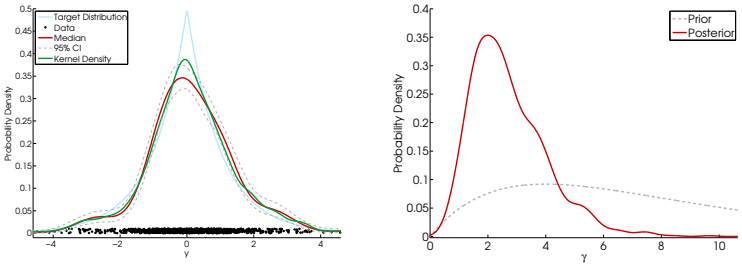
(t) $\lambda = 5$



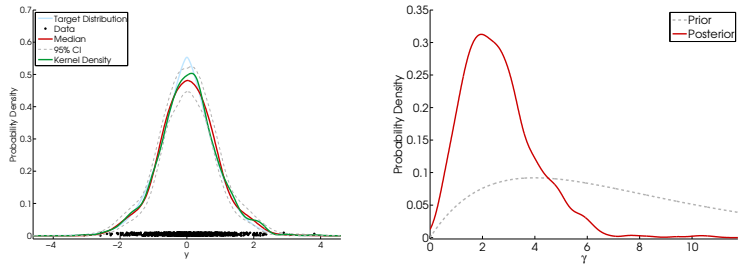
B.2.4 Exponential Power Distribution

Figure B.31 – *Output of DPMN model using a Exponential Power target distribution with uncertain γ.*

(a) $\lambda = 1$ **(b)** $\lambda = 1$



(c) $\lambda = 1.5$ **(d)** $\lambda = 1.5$



(e) $\lambda = 2.5$ **(f)** $\lambda = 2.5$

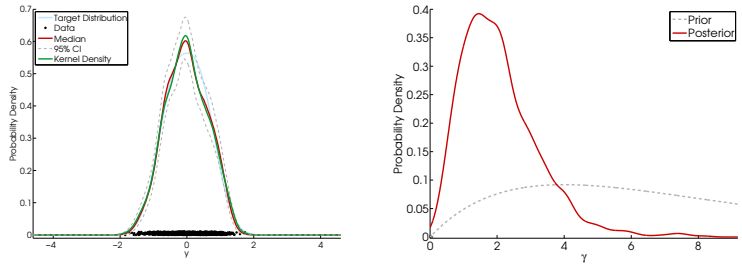
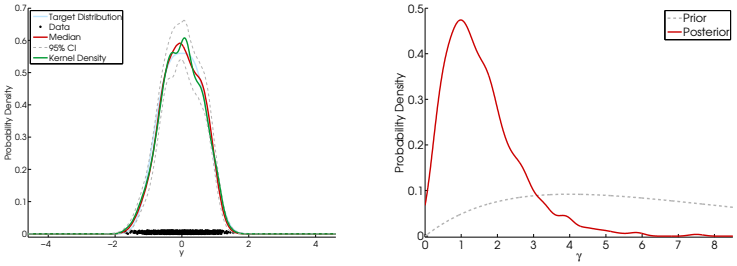
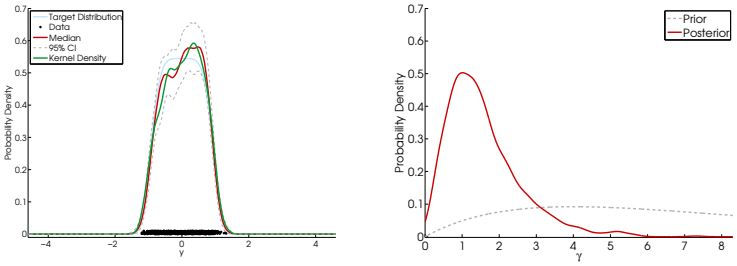


Figure B.31 – *Output of DPMN model using a Exponential Power target distribution with uncertain γ - Continued.*

(g) $\lambda = 3$ **(h)** $\lambda = 3$



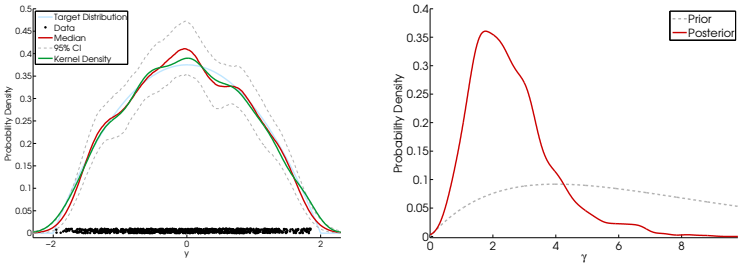
(i) $\lambda = 5$ **(j)** $\lambda = 5$



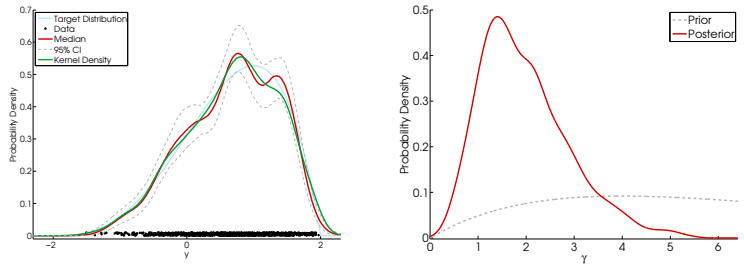
B.2.5 Beta Distribution

Figure B.32 – *Output of DPMN model using a Beta target distribution with uncertain γ.*

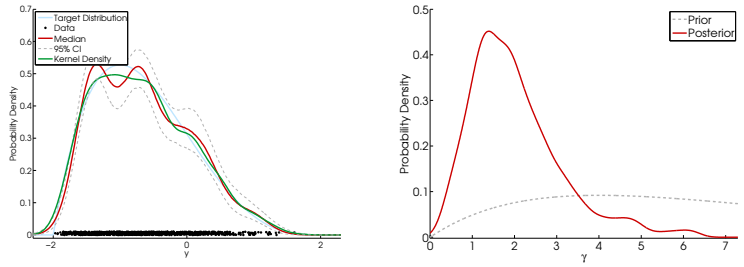
(a) Beta(2,2) **(b)** Beta(2,2)



(c) Beta(4,2) **(d)** Beta(4,2)



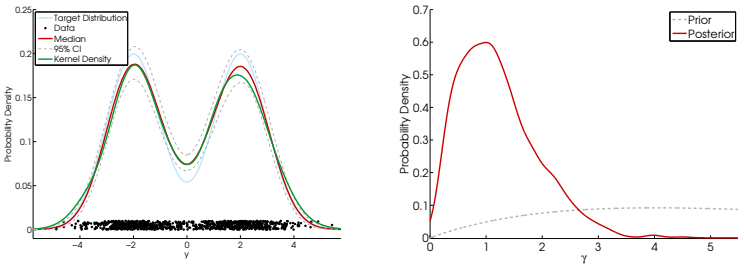
(e) Beta(2,4) **(f)** Beta(2,4)



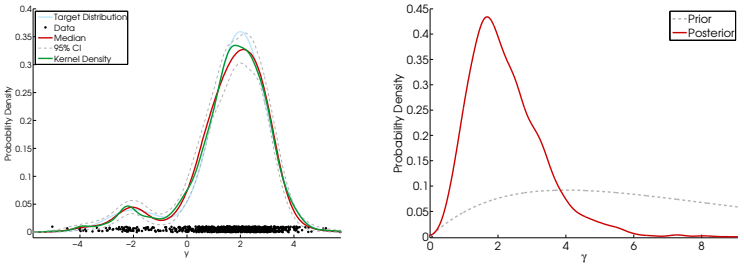
B.2.6 Mixture of Two Normal Distributions

Figure B.33 – *Output of DPMN model using a Normal mixture target distribution,* $p(y) = p \times \mathcal{N}(y; -1, 0.5^2) + (1-p) \times \mathcal{N}(y; 0.5, 0.25^2)$*, for various values of* p *and with uncertain* γ*.*

(a) $p = 0.5$ **(b)** $p = 0.5$



(c) $p = 0.1$ **(d)** $p = 0.1$



(e) $p = 0.9$ **(f)** $p = 0.9$

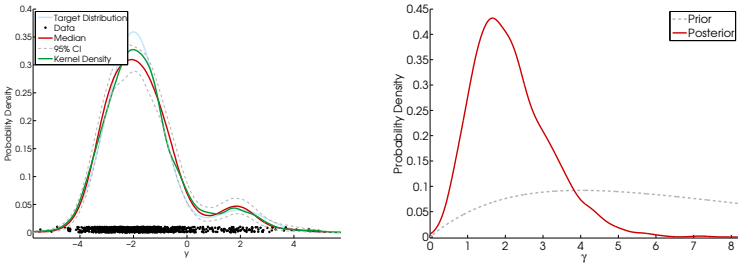
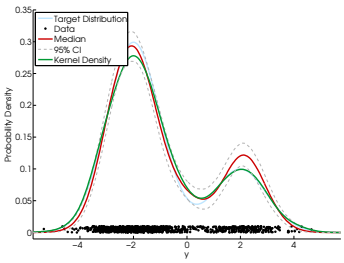
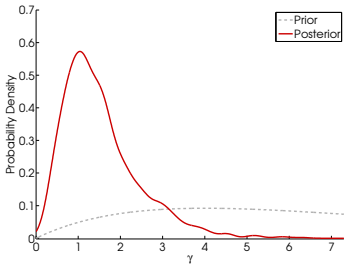


Figure B.33 – *Output of DPMN model using a Normal mixture target distribution,* $p(y) = p \times \mathcal{N}(y; -1, 0.5^2) + (1-p) \times \mathcal{N}(y; 0.5, 0.25^2)$*, for various values of* p *and with uncertain* γ *- Continued.*

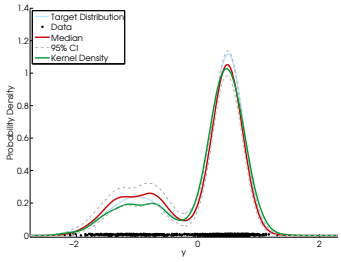
(g) $p = 0.75$



(h) $p = 0.75$



(i) $p = 0.3$



(j) $p = 0.3$