UNIVERSITY OF STRATHCLYDE
DEPARTMENT OF MANAGEMENT SCIENCE

On Taking a Moment to Learn from Experts

Bram Willem Wisse

A Thesis submitted for the degree of Doctor of Philosophy

Glasgow, 2012

Declaration of authenticity and author's rights

This thesis is the result of the authors original research. It has been composed by the author and has not been previously submitted for examination which has led to the award of a degree.

Acknowledgements

This research was made possible by the support of TNO Defence, Safety and Security. I would like to thank my supervisors, Tim Bedford and John Quigley, for their continuous support and incredible patience, and Ana Barros for the support from TNO. Special thanks also go out to my girlfriend, Joëlle, for having to miss me (mentally) for so many years, and the sacrifices she had to make to make this research possible. And to my family and friends, for whom I have barely been available but could always enjoy continuous support and understanding. Finally, I would like to thank Roger Cooke for providing access to his expert data sets for use in this research.

Je gaat het pas zien als je het doorhebt

Johan Cruijff

Contents

List of Figures

List of Tables

Abstract

Subject matter experts have become increasingly important as sources of valuable information in the support of decision making for the Dutch Defence. Yet, the Defence methodology toolbox is lacking a methodology for dealing with quantitative subject matter expert judgements. In this thesis we evaluate a methodology that reflects the discrete character of quantitative expert judgements and is flexible in the amount of detail that can both be specified by the experts and is needed for the decision problem at hand: the Bayes linear methodology. This entails that the methodology can be applied within a relatively short time frame, leading to a short response time. The methodology evaluated in this thesis also provides a vehicle to gradually switch from expert judgement to actually observed data when this becomes available.

To date little guidance is available as to how to obtain the assessments from experts necessary to populate a Bayes linear model. In this thesis we have evaluated (a bivariate extension of) the extended Pearson-Tukey method for the derivation of the second order moment assessment needed to quantify a Bayes linear model, by evaluating its performance for a wide variety of bivariate distributions. We found this method to perform very well when variables are not strongly skewed.

By means of simulation studies we show that the Bayes linear adjustment of moments can be inaccurate for not joint Normally distributed variables. Yet, we find that the use of higher order moment information can greatly increase the accuracy. For the distributions considered in this thesis the increase is between five and eleven orders of magnitude when third and fourth order moment information is used as well in the adjustment. For distribution with a poor performance of the regular adjustment of moments this increase in accuracy is sustained when this higher order moment information is to be obtained from expert assessments, leading to increased accuracy between one and two orders of magnitude. Finally we develop a performance based method to combine sets of (product) moment assessments from different experts into one set of assessments that represents a rational consensus of the experts' assessments, so that multiple experts can be consulted for a Bayes linear study.

Based on the results presented in this thesis we strongly advise to complement the Defence methodology toolbox with the Bayes linear methodology.

Chapter 1

Introduction

1.1 Introduction

The circumstances under which the Dutch Armed Forces have to operate have rapidly changed in the past decade. The Armed Forces are faced more and more with a broad spectrum of (new) operational theatres and irregular operating opponents. Together with the effect of rapid technological developments, this places new challenges to Defence (Barros 2009). Since e.g. the operational environments are increasingly unfamiliar prior to deployment, often little is known and little data is available for analysis and decision support. Therefore subject matter experts have become increasingly important as sources of valuable information. Yet, as (the elicitation of) expert judgments are fundamentally different from ordinary data (collection), previously proven methods for data collection and analysis do not suffice anymore. In this thesis we aim to complement current methods available with a methodology especially tailored to work with quantitative expert judgement.

We hold the viewpoint that when sufficient data is available, this data should be preferred over expert judgements. However, when data is lacking, we will have to resort to the only alternative available: the expert opinion. Therefore we focus in this thesis on cases in which quantitative assessments are desired from subject matter experts (SMEs). These quantitative assessments could e.g. be related to properties of (future) systems (of systems), like development and whole life costs, reliability and other performance indicators for a broad spectrum of operational circumstances. When SMEs are queried for assessments of magnitude, it is good practice not to rely on point estimates only but to ask for judgments about the uncertainty in these estimates as well. Probabilistic methods are therefore commonly applied to represent quantitative expert opinion. Starting from quantitative expert opinion, we would like to gradually switch to actual observed data when this becomes available. Bayesian statistical methods provide an excellent vehicle for this purpose. With Bayesian methods, priorly expressed quantitative beliefs can be revised in a rational and coherent way when data becomes available. Bayesian methods have the nice property that the more data is available, the less the revised beliefs will rely on the initial (expert) assessments.

Full probabilistic Bayesian methods however are very involved. Usually strong assumptions are required about the probabilistic distribution of, and the relationships between variables in the model to be able to conduct the calculation of revised beliefs. The specification of the prior beliefs needed to operationalise Bayesian models can easily be beyond the ability of subject matter experts, who are typically not experts in probabilistic methodology. Moreover, this process can be very time consuming which poses immediately a restriction to its use in military practice given the increasingly short time of response required in current military decision making. So there is a need for a methodology that possesses the benefits of the full probabilistic Bayesian approach, but suffers less from the disadvantages just described. In this thesis we propose and evaluate a methodology that can potentially fulfil this need.

1.2 The Bayes Linear Methodology

In this thesis we will focus on the applicability of the Bayes linear (BL) methodology to decision support and analysis problems, wherein quantitative assessments are desired from subject matter experts, that can be revised when data becomes available. The BL methodology was developed by Michael Goldstein in a series of papers (Goldstein 1981, Goldstein 1986, Goldstein 1988*a*, Goldstein 1988*b*, Goldstein 1991, Goldstein 1994). This methodology takes expectation rather than probability as the fundamental concept and is based on the following four principles (Goldstein 1994):

Principle 1 Specify only those aspects of their beliefs that assessors are both willing and able to quantify honestly.

Principle 2 Use coherent probabilistic guidelines for revising beliefs.

Principle 3 Base statistical models on judgements about observable quantities.

Principle 4 Use theory to interpret the underlying structure of beliefs.

In the base case, when we wish experts to assess their beliefs about magnitudes of quantities of interest and wish to learn more about these magnitudes by observing other quantities, Goldstein argues that the bare minimum aspects that must be considered are:

1. some quantitative judgements as to the magnitudes of the various quantities,

2. some expression of the degree of confidence in the judgements of magnitude,

3. some expression of the extent to which the prior judgements about the various quantities are interrelated (so that observation on some of the quantities may be used to modify judgements on other quantities).

In the Bayes linear methodology assessments of respectively means, variances and covariances are chosen to quantify these aspects. All three can be derived from first and second order (product) moments, so Bayes linear models are thus fully specified by a second order moment specification. Since belief revisions in light of new observations in this methodology also are operations on moments only, the methodology can be viewed as a method of moments. The assessment of moments can be given a strong foundation by the use of De Finetti's definition of coherent previsions: having the expert state their assessments such that no bet can be made based on the assessments that would make the expert a sure loser. Once a Bayes linear model is quantified by a set of coherent (product) moment assessments, the revision of these moments in the light of new observations will also result in a coherent set of revised moments.

The methodology however is by no means restricted to the assessment and revision of magnitudes only. As beliefs of magnitudes of uncertain quantities are specified, so can beliefs about functions of these uncertain quantities be, for example the square or the cube of the same quantity. By including these functions in the model, beliefs about e.g. variability and asymmetry can be specified and revised as well. To quote Goldstein (1994), the Bayes linear belief specification "may be viewed as reducing the full probabilistic approach to whatever level of detail we feel is both within our ability to specify and adequate to the problem at hand".

The methodology thus reflects the discrete character of quantitative expert assessments and is flexible in the amount of detail that can both be specified by the experts and is needed for the decision problem at hand. Furthermore, the methodology is assumption free as in that it does not require the quantities to have a probability distribution from a certain family of distributions. The only requirements needed are that the second order (product) moments for the quantities in the model are finite, and coherently specified.

Although the methodology does not employ the concept of probability, Bayes linear belief revisions can be compared with full probabilistic belief revision via the moments, since these are also defined in a probabilistic setting, as probability weighted averages. In fact, in the base case in which Bayes linear belief specification and revision are performed for only the quantities themselves and not functions of these quantities, the belief revisions are identical to those in the full probabilistic case using the same second order specification and the assumption of joint Normality as the probability distribution for the quantities of interest (referred to as variables in a probabilistic context). In this case the Bayes linear belief revision more or less reduces to an old trick in a new jacket, although it must be acknowledged that the methodology provides us with interpretative and diagnostic tools to analyse the specified beliefs as well, and, again, does not require the concept of probability.

Although the BL methodology is designed to model and revise quantitative (expert) assessments, little guidance has been provided to date on how the belief specifications needed to quantify the model can be provided by subject matter experts, and how well the methodology performs compared to the full probabilistic alternative. Therefore, to assess whether the Bayes linear methodology can fulfil the need identified in the introduction, the questions formulated in the next section need to be answered.

1.3 Research Questions

The Bayes linear methodology has the potential to meet the need specified in the introduction. But before one could comfortably apply the methodology, an (affirmative) answer to the research questions stated in this section is required.

To start, without experts being able to provide us with the moment assessments necessary we will have no specified beliefs to analyse and revise, so:

1. Can experts provide the beliefs necessary, i.e. can they assess (product) moments confidently and reliably?

Recall from the Principles 1 and 3 above that experts should be willing and able to make the judgements required, and that experts should be asked only for judgements about (in principle) observable quantities. Since we consider the Bayes linear approach as an alternative to full probabilistic modelling, as a moment based approximation to it, we need to be able to assess to what extent the Bayes linear belief revision corresponds to the belief revision we would have found using full probabilistic modelling based on the same (product) moment assessments:

2. How accurate are the Bayes linear adjustment rules when considered an approximation

to full probabilistic updating, when the quantities (variables) are not assumed to be joint Normally distributed?

We anticipate deviations between Bayes linear and full probabilistic belief revision when variables are not joint Normally distributed. Since Goldstein states that in his moment based method we can use whatever level of detail is within the ability of the expert to specify, we wonder whether these deviations can be reduced by increasing the level of detail:

3. Can the accuracy of Bayes linear belief adjustment be improved by using higher order information, and is this improvement sustained when deriving the moment assessments necessary from expert assessments?

Finally, it is typical for expert judgement based studies and commonly advised that multiple experts are consulted. What should a decision maker do with multiple sets of expert assessed (product) moments:

4. How can we aggregate moment assessments of multiple experts?

We will now proceed to discuss how this thesis is set up to address these research questions.

1.4 Structure of the Thesis

The thesis is set up as follows. In Chapter 2 we will discuss the measurement of uncertainty with probability and expectation. We will describe how these measures can be operationalised and thereby given meaning when they are based on judgements (of experts), and how they can be revised when new information becomes available. In Chapter 3 we discuss the difficulties, pitfalls and dangers that might rise when we actually try to conduct measurements of uncertainty with experts; the elicitation of quantitative expert judgements. We treat different approaches to evaluating whether an elicitation exercise can be considered successful. We review the literature on methods to elicit means, variances and covariances, thereby answering research question 1 for first and second order moment assessments.

We then proceed in Chapter 4 to formally introduce the methodology under evaluation in this research: the Bayes linear methodology. We describe how a Bayes linear model can be constructed from first and second order moment assessments, and the Bayes linear adjustment rules for means and variances, core to the methodology. We discuss the interpretations that can be given to the belief adjustments, and the interpretative and diagnostic tools provided by the methodology to analyse the specified beliefs and (potential) revisions of these by observations.

In Chapter 5 we consider the Bayes linear approach as an approximation to full probabilistic updating. We investigate how accurate the Bayes linear adjustments rules are when variables are not joint Normally distributed. We select a set of bivariate distribution families and evaluate the difference between the Bayes linear adjusted mean and variance and the conditional mean and variance for these distribution families. In this chapter we thus formulate an answer to research question 2.

We then evaluate the possible benefits of using higher order (product) moment information in Bayes linear belief adjustment, in Chapter 6. First we analyse the benefits for the situation in which exact knowledge of the moments is available, part one of research question 3. We then evaluate the performance of a methodology for the assessments of higher order (product) moment assessments, completing the answer to research question 1. Finally we evaluate the possible benefits of using higher order (product) moment information in Bayes linear belief adjustment when the (product) moments necessary are derived from assessments experts can provide, part two of research question 3.

In Chapter 7, finally, we address research question 4 by developing a performance based aggregation method for sets of expert assessed (product) moments.

Throughout this thesis we will make use of the designations 'subject matter expert', 'expert' and 'assessor'. We will use the term subject matter expert when referring to expert assessors in a Defence context. Expert will simply stand for any expert, someone's knowledge we wish to use and we will use assessor to refer to someone in general assessing something. To improve readability (expert) assessors will be referred to in masculine form and decision makers in feminine form.

Chapter 2

Uncertainty

Uncertainty about the value of a quantity can be studied by applying statistical methods to observed values for that quantity. In this thesis however we focus on the case in which we are interested to learn about the value of a quantity for which we do not (yet) have observations available. We might be interested e.g. in the acquisition costs or performance of an aircraft that is still only in the design phase, the reliability and accuracy of a new type of missile that is too expensive to test in real trials, or the added operational value of self protective measures yet to be developed. Even in the case that we could be informed by data that is available for in some way comparable systems, subject matter expert judgement is needed to relate this data to the system of interest. In this chapter we will therefore discuss the *subjective* assessment of uncertainty via expectation and probability. We start with defining the concept and discuss different categorisations that can be found in literature that distinguish between different types of uncertainty. We will then discuss how uncertainty can be measured, introducing the concepts of probability and expectation. Special attention will be given to the different interpretations that can be given to these measures. Finally we will treat different ways in which measurements of uncertainty can be revised when more information becomes available.

2.1 Uncertainty

Two topics are of main concern in this thesis: the representation of subjective uncertainty and the revision of this subjective uncertainty in the light of new information/observations. So let us first describe what we mean by uncertainty. At a basic level, something is uncertain whenever it is not completely certain. According to the dictionary, certainty is “something that is clearly established or assured”. In (Bedford & Cooke 2001) this ‘something’ is defined as a declarative sentence, which can be ‘established’ or ‘assured’ by determining (a) whether truth conditions exist for it and (b) the conditions for the value ‘true’ hold.

An evaluation of uncertainty will depend on the information (set of observations) that is available. We will demonstrate what we mean by this using a slightly modified example from (Winkler 1996). Suppose we toss a coin. When only being told that the coin is fair, most people will be very uncertain about whether the coin will land ‘heads’ or ‘tails’. But, if we have information/observations on all conditions surrounding this toss (like e.g. initial side facing up, height, velocity, wind, nature of surface the coin is bound to land on, etc.) the laws of physics could be used to predict the outcome with certainty, or close to certainty. Therefore, the uncertainty about the outcome of the coin toss depends on the available information/observations. In fact, uncertainty is that which is reduced or removed by observation (Bedford & Cooke 2001).

Different types (or: classifications) of uncertainty are mentioned in the literature. The type of uncertainty can have implications for how it can be measured. We will therefore first discuss these types before treating the quantitative measurement of uncertainty.

2.2 Types of Uncertainty

In the coin tossing example of the previous section, probably the most frequently mentioned categorisation of uncertainty was implicitly introduced: the distinction between aleatory (intrinsic) and epistemic (lack of knowledge) uncertainty. This classification is discussed in the first subsection, followed by parameter and model uncertainty and finally volitional uncertainty.

Not to be confused with uncertainty is ambiguity (Bedford & Cooke 2001). The authors point out that verbal and written language can often be explained in different ways. They state that this ambiguity, i.e. the lack of well-defined truth conditions, must be removed to be able to discuss uncertainty in a meaningful way. Where uncertainty is that which is reduced or removed by observation, ambiguity is that which is removed by linguistic convention (Cooke 1991).

2.2.1 Aleatory and Epistemic Uncertainties

Aleatory uncertainties, derived from the Latin word for dice ‘alea’, arise through variability intrinsic to a system. Epistemic uncertainties refer to the uncertainties due to lack of knowledge of a system of interest (see e.g. (Winkler 1996, Bedford & Cooke 2001, O’Hagan & Oakley 2004)). So epistemic uncertainty could be reduced by gaining more knowledge about the system of interest. Aleatory uncertainty, in contrast, is that uncertainty we cannot or do not make the effort to reduce.

When reconsidering the toss of the coin in the example in Section 2.1 (or maybe even better, replace it with the throw of an ‘alea’), it should become clear that it is not always apparent whether an uncertainty should be classified as aleatory or epistemic. Depending on the context both classifications could be appropriate. Winkler (1996) therefore mentions this distinction between types of uncertainty to be ’fundamentally flawed’. Though, in concrete situations the distinction will usually be quite apparent: when considering to engage in a game of chance in which winning or losing depends on the throw of a dice, it is usually not feasible to base that decision on a physical model of the throw as suggested in Section 2.1.

Aleatory and epistemic uncertainties are in literature also referred to as irreducible/reducible, stochastic/subjective, Type A/Type B and variability/state of knowledge.

2.2.2 Parameter and Model Uncertainty

Parameter uncertainty is described in (Bedford & Cooke 2001) as uncertainty about the ‘true’ value of a parameter in a mathematical model. Often no ’real-life’ interpretation of this parameter is available. The authors state that in this case parameter uncertainty can only be given a meaning, and be measured, if this uncertainty is taken to represent the uncertainty of an observer about the accuracy of model predictions on observable quantities. O’Hagan & Oakley (2004) regard parameter uncertainty to be generally epistemic, since, they state, one commonly just does not know what the correct values for input parameters are.

Quantifying model uncertainty is even more problematic, since “every model is definitely false” (Morgan & Henrion 1990). Describing model uncertainty as ‘uncertainty about the truth of the model’ would therefore not be useful (Bedford & Cooke 2001). O’Hagan & Oakley (2004) speak of ‘uncertainty about model inadequacy’ and state that to be ‘unequivocally epistemic’. They provide the following reasoning behind this. Consider the case in which a real process has been modelled, but in which there still is some residual variability in the value of the process when model conditions are repeated. This variability can be regarded as aleatory, if it is considered to be natural to the process. At the same time adding more conditions to the process could eliminate or reduce this residual variability. The removed variability then was due to the lack of knowledge about these extra conditions, and therefore must have been epistemic.

By introducing a discrete variable to indicate which model is used, model uncertainty can be seen as a special case of parameter uncertainty and can thereby be given a meaning (Bedford & Cooke 2001). That is, as much meaning as can be given to parameter uncertainty.

2.2.3 Volitional Uncertainty

Volitional uncertainty is uncertainty about so called first person events. A first person event is an event whose defining conditions involve decisions of the acting subject (Cooke 1986). The problem of first person events is that they cannot be measured via preference behaviour (the measurement of uncertainty using preference behaviour will be introduced in the next section). An example of the problem of measuring preference for a first person event, from (Bedford & Cooke 2001), is the following: suppose a person P is asked for his preference between the following two options:

(a) receive \$1,000,000 if P cleans his cellar next weekend, or receive \$0 otherwise,

(b) receive \$1,000,000 if the Dow-Jones is lower at the end of the week, or receive \$0 otherwise.

Considering most people in the position of P will prefer (a) to (b), one could say that (a) is more likely to occur then (b). But when the stake of \$1,000,000 is lowered to \$1, most people will probably prefer (b) to (a), implying (b) to be more likely to occur than (a). Thus, the act of measuring uncertainty about first person events by observing preference behaviour influences this uncertainty. This problem does not occur if another person than person P expresses his preference for the above described options.

2.3 Measurement of Uncertainty

Uncertainty can be attributed in many ways. In everyday life words like ‘probably’, ‘unlikely’ and ‘rarely’ are used to express uncertainty, by which different degrees of uncertainty can be distinguished. In the current and following two sections we will focus on how uncertainty can be measured quantitatively, i.e. how we can express uncertainty as a finite number.

We will discuss the measurement of uncertainty in relation to either events or random quantities. An event is something that either occurs or not, which we might not know with certainty. By random quantity we will simply mean any well-defined quantity about the value of which we might be uncertain. The concepts of event and random quantity can be easily related using the indicator function of the event: the indicator function that takes the value 1 if the event occurs and 0 if not, is a random quantity.

Before we can measure uncertainty quantitatively, we first need to describe of course what it exactly is we want to measure and how we intent to measure it. Lindley (2000, p.295) reminds us that some sort of a standard is needed, since ‘all measurement is based on a comparison with a standard'. Cooke (2004) argues that this standard should have an operational definition. Without clearly defining in empirically observable terms what a measurement of uncertainty is representing, he states, it is not possible to assess whether a specific representation of uncertainty is appropriate. This is of course especially the case with subjective representations of uncertainty. What questions are to be asked to someone when a quantitative assessment of his uncertainty about an event or quantity is desired? How are these questions to be understood by the assessor, how should the answers given to them be interpreted and what meaning can be given to them?

Various approaches to an operationally defined measure of uncertainty can be found in the literature. In the remainder of this section we will introduce one of these: Bruno de Finetti's betting approach (De Finetti 1974). Starting from a description of the observable phenomenon, betting behaviour, used in this measurement approach, we will show how two well-known quantitative measures of uncertainty, probability and expectation, can be operationally defined using this behaviour. The reason for discussing De Finetti's approach here before the other approaches is that it is a representation of subjective uncertainty and that *both* probability and expectation can be defined *directly* from the observable phenomena in this approach. So next to probabilistic methods, also moment-based methods like the Bayes linear methodology that take expectation as the primitive notion can be directly operationally defined with De Finetti's approach. In the Sections 2.4 and 2.5 the mathematical definition and properties of resp. probability and expectation are described. Section 2.6 treats other (operational) definitions of probability, more commonly referred to as interpretations of probability. We will discuss the implications these interpretations have on the types of settings in which probability can be applied as measurement of uncertainty, and the different ways in which probabilities can be revised when more information becomes available.

2.3.1 De Finetti's Betting Approach: Coherent Previsions

De Finetti's approach relates the uncertainty people have about a random quantity of interest with their betting behaviour. His ideas were first published in the 1930s. The standard reference to the De Finetti's approach has become his *Theory of Probability*, first published in Italian in 1970 and in English in two volumes in 1974 and 1975. An extensive historical and philosophical introduction into De Finetti's approach can be found in (Lad 1996). This work also updates De Finetti's Theory of Probability with work done in the 25 years following its initial publication (Schafer 2002).

In De Finetti's approach the 'measurement standard' for an individual's uncertainty is his *prevision* for a random quantity. A prevision is the value a person chooses when he is engaged in a bet in which his loss depends on the difference between his stated value and the true value of the random quantity. The bigger the difference, the bigger the loss. Formally, prevision is defined to be the value $\overline{x}$ that an individual chooses in either one of the two following, equivalent, criteria (De Finetti 1974, pp. 87-88):

- First criterion. Given a random quantity (or random magnitude) X, you are obliged to choose a value $\overline{x}$, on the understanding that, after making this choice, you are committed to accepting any bet whatsoever with gain $c(X - \overline{x})$, where c is arbitrary (positive or negative) and at the choice of an opponent.

- Second criterion. You suffer a penalty L proportional to the square of the difference (or deviation) between X and a value $\overline{x}$, which you are free to choose for this purpose as you please: $L = \left(\frac{X-\overline{x}}{k}\right)^2$ (where k, arbitrary, is fixed in advance, possibly differing from case to case).

It follows directly from both criteria that a person who does not want to be a sure loser, should not state a prevision that is smaller than the minimum possible value of the quantity or lager than its maximum possible value. A prevision smaller than the minimum value will result in a positive loss with certainty, and stating the minimum value as prevision instead would result in a smaller loss with certainty. A prevision that results in a sure loss is called incoherent. A prevision for the indicator function of an event is therefore incoherent if it is not in the closed interval $[0, 1]$.

But even if previsions are coherent individually, it is possible that a set of previsions is not. Let X for example be the indicator function for event A, and Y the indicator function for event A^c, indicating event A *not* happening. If someone's previsions for X and Y are resp. 0.75 and 0.5, then both previsions individually are coherent. But we *can* construct set of bets using both previsions in which this individual will be a sure loser: if, using De Finetti's first criterion, we bet c on both quantities, this individual's loss will be $c(X - \overline{x}) + c(Y - \overline{y}) = c((X + Y) - (\overline{x} + \overline{y})) = c(1 - 1.25) = -0.25c$. So for any $c > 0$, this individual will be a sure loser. Such a combination of bets leading to a sure loss is often referred to as a Dutch book. In the current example no Dutch book can be made against an individual who makes sure that both his individual previsions are coherent and that the sum of his previsions for X and Y is equal to 1. Individuals who do not wish to engage in bets that will result in a sure loss should thus state coherent previsions.

One of the major implications of the definition of coherent previsions is that it assumes an individual to be risk neutral. For example this would mean that if an individual is indifferent to receiving 0.50 for certain or 1.00 if event A occurs, he would also have to be indifferent to receiving 5,000,000 for certain or 10,000,000 if A occurs. Risk neutrality is not assumed a generally valid assumption (French 1986). To make the assumption more realistic minimum and maximum monetary values (often referred to as 'stakes') can be introduced between which the assumption of risk neutrality seems reasonable. Or so called units of utility, defined especially in such a way to ensure an individual's risk neutrality in regard of it, could be used instead of monetary units.

We shall now formalise De Finetti's notion of coherency. With respect to the first criterion, a set of previsions is said to be coherent if there is no linear combination of the bets for each of the previsions with a negative supremum. In the context of the second criterion a set of previsions is said to be coherent if there is no other possible choice that would certainly lead to a uniform reduction in penalty L. Coherency thus requires an individual not to have a preference for a given penalty if he has the option of another penalty that is certainly smaller.

De Finetti has shown that coherency of previsions for any two quantities X and Y, denoted here as $Pv(X)$ and $Pv(Y)$ respectively, is equivalent to the following two restrictions (De Finetti 1974, p.74):

$$\begin{aligned} &\text{(i)} \qquad \min \mathcal{R}(X) \leqslant Pv(X) \leqslant \max \mathcal{R}(X), \\ &\text{(ii)} \qquad Pv(X+Y) = Pv(X) + Pv(Y), \end{aligned} \tag{2.1}$$

where $\min \mathcal{R}(X)$ is the smallest member of the realm of X and $\max \mathcal{R}(X)$ the largest member. The necessity of restriction (i) for avoiding a sure loss was already explained in this section. The combined bet example from this section treats a special case of restriction (ii), which states that previsions are linear.

If we restrict X to be an indicator function of some event A, we find from (i) that $Pv(X) \geq 0$. When A is the certain event the realm of X reduces to the value 1. So, again from (i), we have that $Pv(X) = 1$ when A is the certain event. Let Y be the indicator function of event B. If we consider the eventuality of either of the events A and B occurring, we find from (ii) in terms of previsions of their indicator functions $Pv(X \text{ OR } Y) = Pv(X + Y - XY) = Pv(X) + Pv(X) - Pv(XY)$. If A and B cannot occur at the same time, i.e. $XY \equiv 0$, then $Pv(X + Y) = Pv(X) + Pv(Y)$. As we shall show in Section 2.4.1, these properties define previsions of indicator functions of events to be probabilities.

Previsions also satisfy the axioms of expectation, which will be described in Section 2.5.1. To see this, we need to add that coherency implies that every linear relation between the random quantities, $\sum_{i=1}^{n} c_i X_i = c$, must be satisfied by the corresponding previsions for these quantities, so that $\sum_{i=1}^{n} c_i Pv(X_i) = c$ (De Finetti 1974, p.89). So for the simplest case $n = 1$ we find that $Pv(cX) = cPv(X)$.

We conclude this section with a geometric interpretation that can be given to a set of coherent previsions, equivalent to the two restrictions from (2.1). Let $\mathbf{X}$ be a vector of random quantities. Then the set of all coherent previsions for $\mathbf{X}$ is the closed convex hull of the realm of $\mathbf{X}$. The convex hull representation of coherent previsions will be of interest to us in Section 7.2.2, where we will discuss the aggregation of sets of coherent assessments.

2.4 Probability

A probability is a normalised measure of uncertainty that obeys certain mathematical properties. We will use the notation $P(A)$ for the probability of an event A occurring. The probability of an event occurring is a value between 0 and 1, where $P(A) = 0$ means that it is impossible that A will occur. On the other extreme, a probability of 1 refers to the case that the event is certain to occur. A probability of 0.5 means that an event is just as likely to occur than not. We will first introduce the notion of a (σ-)field, a collection of events that can be assigned probabilities, and then proceed to define probability formally with Kolmogorov's axioms.

Events are defined here using set theory. Let Ω be a non-empty set of outcomes, or possible worlds, and let $\varnothing$ be the empty set. Set B is called a subset of set A, notation $B \subseteq A$, if and only if $\forall \omega \in \Omega : (\omega \in B \Rightarrow \omega \in A)$. If in addition to B being a subset of A, there also exist an element $\omega \in \Omega$ that is an element of A but not of B, then B is called a proper subset of A, with notation $B \subset A$. A set $A \subset \Omega$ is defined here as an event. For these events, the operations union, intersection and complement be defined as:

$$\begin{aligned} A \cup B &:= \{\omega \in \Omega : \omega \in A \text{ OR } \omega \in B\} \text{ is called the union of } A \text{ and } B. \\ A \cap B &:= \{\omega \in \Omega : \omega \in A \text{ AND } \omega \in B\} \text{ is called the intersection of } A \text{ and } B. \\ A^c &:= \{\omega \in \Omega : \omega \notin A\} \text{ is called the complement of } A. \end{aligned}$$

A set of events $\mathcal{F}$ is called a field if:

i. $\varnothing \in \mathcal{F}$,

ii. If $A \in \mathcal{F}$, then $A^c \in \mathcal{F}$

iii. If $A, B \in \mathcal{F}$, then $A \cup B \in \mathcal{F}$

If *iii* is augmented to account for infinite unions of events as in iii'.

iii'. If $A_1, A_2, \ldots \in \mathcal{F}$, then $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$,

then $\mathcal{F}$ is called a σ-field.

2.4.1 Kolmogorov's Axioms

Let $\mathcal{F}$ be a field over Ω, then the number $P(\cdot)$ obeying the following three axioms of Kolmogorov is called a probability:

I. For each set $A \in \mathcal{F}$, $P(A) \geq 0$.

II. $P(\Omega) = 1$.

III. If $A \cap B = \varnothing$, then $P(A + B) = P(A) + P(B)$.

The axioms state that probability (I) is a nonnegative number, that (II) the certain event is assigned the probability 1 and (III) that if two events cannot occur at the same time, the probability that either one of these events occurs is equal to the sum of the probabilities of each of the events. The axioms *I*-*III* form a finitely additive positive normalized measure. The third axiom can also be extended to account for infinite sequences of events, requiring $\mathcal{F}$ now to be a σ-field:

III'. If $A_1, A_2, \ldots$ are such that $A_i \cap A_j = \varnothing$ for $i \neq j$, then $P\left(\bigcup_{i=1}^{\infty} A_i\right) = \Sigma_{i=1}^{\infty} P(A_i)$.

The measure defined by the axioms I, II and III' is referred to as countable additive or σ-additive.

Although Kolmogorov's axioms are introduced here in the context of (σ-)fields of events, their application is by no means restricted to this interpretation. For as Kolmogorov (1956 (1933), p.1) states: "Every axiomatic (abstract) theory admits, as is well known, of an unlimited number of concrete interpretations besides those from which it was derived". In the previous section it was shown that a probability can be interpreted as someone's prevision of an indicator function, in Section 2.6 other interpretations will be discussed.

2.4.2 Conditional Probability and Bayes Rule

The probability that event A occurs given that (hypothetically) it is certain that event B has occurred is called the conditional probability of A given B, and is defined as:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}.$$

Since division by zero is not allowed, $P(B)$ needs to be bigger then zero. The probability of an event cannot be taken conditional on something that cannot occur.

Instead of this definition via unconditional probabilities, many authors that adhere a subjective interpretation of probability (e.g. Jeffreys, Jaynes) take conditional probability as primitive. They argue that someone's probability for an event is always conditional on his knowledge or state of information about the event. Returning to the betting approach from Section 2.3.1, De Finetti (1974) also introduces the notion of conditional prevision, prevision conditional on event occurring. If X is any quantity and E any event, then your conditional prevision for X given E, denoted by $P(X|E)$, is the number you specify with the understanding that you are thereby asserting your willingness to engage in any transaction that would yield you a net gain of the amount $s[XE - EP(X|E)]$, as long as $|s[XE - EP(X|E)]| \leqslant S$ for every pair of numbers (e, xe) in the realm of (E, XE), where S is the scale of your maximum stake (Lad 1996, p.123).

From Kolmogorov's axioms and the definition of conditional probability the following well known expression can by fairly easily derived (see e.g. (French 1986)):

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}. \tag{2.2}$$

This equation is known as the simplest form of Bayes' rule, or Bayes Theorem, and will be of particular interest to us throughout this thesis. If we define $A_1, A_2, \ldots, A_N$ to be any partition of the certain event, that is $\Omega = \bigcup_{i=1}^{N} A_i$ and $A_k \cap A_l = \varnothing$ for $k \neq l$, and let B be any event such that $P(B) > 0$, then we can define the general form of Bayes' rule as:

$$P(A_l|B) = \frac{P(B|A_l)P(A_l)}{\sum_{i=1}^{N} P(B|A_i)P(A_i)}. \tag{2.3}$$

Bayes rule plays an important role in the revision of probabilities in the light of new information/observations. This will be discussed in Section 2.6.1.

2.5 Expectation

The expectation of a quantity is also referred to as mean or as the first moment of that quantity. The expectation of the n-th power of a quantity X, $E(X^n)$, is called the n-th moment of X. Product-moments refer to the expected values of products of uncertain quantities.

One of the earliest celebrated works on games of chance is Christian Huygens' *De Ratiociniis in Ludo Aleae* (1657). Huygens was the first to axiomatise a measure of uncertainty. Rather than providing axioms for probability he started from an axiom of the 'value' of a fair game and derived three theorems on expectations (Whittle 1992). In this section we will repeat the axioms that mathematically define expectation, and show how expectation can be defined using probability. We will use the notation $E(X)$ to refer to the expectation (or expected value, or mean) of a random quantity X.

2.5.1 Axioms

Expectation is a normalised positive linear operator, which is realised by satisfying the following four axioms (see e.g. (Whittle 1992)):

$A1$**.** If $X \geq 0$ then $E(X) \geq 0$.

$A2$**.** If c is a constant then $E(cX) = cE(X)$.

$A3$**.** E(X+Y)=E(X)+E(Y).

$A4$**.** E(1)=1.

From (2.1) (Section 2.3.1) it is clear that previsions satisfy the axioms $A1$, $A3$ and $A4$. We have also described in Section 2.3.1 that coherency of previsions implies that $Pv(cX) = cPv(X)$, hence coherent previsions satisfy all expectation axioms.

We have defined expectation now directly from coherent previsions. A more common approach is to derive expectation as a probability-weighted average of the possible values of a quantity. For the discrete case in which the outcome set Ω is a countable set of values $\{\omega_1, \omega_2, \ldots, \omega_k\}$, the expectation of a random quantity, or random variable, X is defined as:

$$E(X) = \sum_{i=1}^{k} P(\omega_i) X(\omega_i). \tag{2.4}$$

For the continuous case that Ω is the real line and ω a scalar, the expectation of X is defined as:

$$E(X) = \int_{-\infty}^{\infty} X(\omega) f(\omega) dx, \tag{2.5}$$

for all $X(\omega)$ for which the integral is defined and absolutely convergent and with $f(\omega)$, referred to as probability density on Ω, obeying

$$\begin{aligned} f(\omega) &\geq 0, \\ \int_{-\infty}^{\infty} f(\omega) dx &= 1. \end{aligned}$$

2.6 Interpretations of Probability and their Implications

In this chapter we have shown how personal betting behaviour can be operationalised as measurement of uncertainty, with probability and expectation as mathematical consequences of coherent bets. Historically the development has been in the opposite direction: Kolmogorov (1956 (1933)) has been the first to formally define probability by the axioms given in Section 2.4.1, and since many attempts have been made to (operationally) define probability by giving this mathematical concept an appropriate interpretation. In this section we will discuss the most predominant of these interpretations, in the order in which they historically have been introduced.

Classical interpretation: the classical interpretation is associated with Laplace (1820) and is predominately discussed in the context of games of chance, like dice and card games. Laplace related the probability of an event to the ratio of the number of outcomes favourable to the event to the total number of possible outcomes, each assumed equally possible. So the probability of the event of a dice landing either '1' or '6' would therefore be related to the ratio #{'1','6'}/#{'1', '2', '3', '4', '5', '6'}=2/6=1/3.

The major weakness in the classical definition of probability is formed by the fact that each outcome is assumed 'equally likely'. Since 'equally likely' could just as well be described with 'equally probable', this definition of probability could be seen as making use of the notion of probability itself and therefore as circular. An attempt to overcome this circularity is formed by the principle of insufficient reason (according to Barnett (1982) first formally treated by Bayes, although Laplace's definition is usually referred to in literature), which specifically aims to define the concept of 'equally likely' without using the notion of probability. But all attempts to such a definition have encountered paradoxes (French 1986).

Apart from the foundational problems the ‘equally likely’ concept incurs, it also limits the application of probability: it cannot be deployed in situations in which possible scenarios are clearly not equally likely to occur.

Relative frequency interpretation: the frequentist approach is largely based on the work of Venn, Von Mises and Reichenbach (Venn 1876, Reichenbach 1949, Von Mises 1957). The basic concept of this approach is that of infinitely repeatable experiments. The probability of an event A is related to the long-run relative frequency of the occurrence of the event in under identical circumstances repeated trials of such an experiment:

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} 1_A(\omega_i),$$

where ω_i is the outcome of experiment i and $1_A(\omega_i)$ the indicator function of the occurrence of event A in experiment i. The expectation of a random quantity X can in the frequency approach also be directly related to an observable phenomenon: the long-run arithmetic average:

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} X(\omega_i).$$

axiom of convergence, axiom of randomness.

The frequentist approach hypothesises a limit to exist for the relative frequency of the occurrence of an event and that this limit is unique, independent of whoever repeats the experiment and whenever it is repeated (French 1986). It thus is an objective approach in the sense that “It describes any view of probability which does not depend on the personal feelings or actions of an individual” (Barnett 1982). Yet, the application of the frequency interpretation is limited: events or propositions must be possible outcomes of repeatable trials.

The justifiability of the frequency approach is questioned on this reliance on the possibility of an experiment to be repeated infinitely under identical circumstances. In practice it will not be possible to repeat any experiment an infinite number of times. French (1986, p.234) states that it is impossible to keep all circumstances identical when repeating an experiment. He therefore argues that this must be interpreted as ‘identical in all aspects relevant to the probability of an event’, leading to a circular definition of probability.

Logical interpretation: the logical approach to probability “[...] seeks to encapsulate in full generality the degree of support or confirmation that a piece of evidence E confers upon a given hypothesis H" (Hájek 2010). Early proponents of this approach are Johnson (1921), Keynes (1921) and Jeffreys (1939). A recent logical approach to probability is (Jaynes 2003). The most systematic study of logical probability was by Carnap (1950). This approach considers uncertainty about propositions rather than about events, and can be regarded as generalising deductive logic. A degenerate case of this approach, in which we would have either impossibility or certainty for each of the events, reduces to Aristotelian logic. The approach is 'objective' in the sense that it does not attribute probability to persons. Instead, it is normative in aiming to prescribe what one's measurement of uncertainty should be given a certain 'state of knowledge'. It is therefore also referred to as 'interpersonal'. The logical approach however fails to relate the object of measurement to an (in principle) observable phenomenon, and thus lacks an operational definition.

The logical interpretation suggests that people with the same knowledge should have the same probabilities, referred to as the 'necessary' view. Lindley (2000) describes two difficulties with this view. The first difficulty lies in defining exactly what is meant by 'two people having the same knowledge'. Secondly, this view entails probability to be independent from the observer, and thus it should be possible to evaluate it without reference to a person, which is not in agreement with the subjective definition. Furthermore, the situation in which two people have exactly the same knowledge would be hard to realise in practice.

Frank Ramsey is with De Finetti one of the founders of the subjective interpretation of probability. He developed his ideas at the same time as but independently from De Finetti. He criticised Keynes' logical interpretation by stating that "[...] there really do not seem to be any such things as the probabilities he [Keynes] describes" ((Ramsey 1926, p.161), cited from (Gillies 2000)). Where in the logical, interpersonal interpretation probability is considered to be the same for all rational individuals, Ramsey abandons the view that rationality should lead to a consensus on probability. Like De Finetti he views the probability of an event as the degree of belief someone has in the event occurring, and proposed to measure the strength of this belief by examining the character of some action to which it lead. And again like De Finetti, he concludes that betting behaviour is a suitable action to measure the strength of this type of belief.

Savage (1972 (1954)) proposes an alternative derivation of probability as degree of belief, using rational preferences instead of coherent bets.

Rational preference interpretation: the basic concept in the rational preference interpretation (Savage 1972 (1954)) is personal preference between events, determined via preference in bets on these events. From these preferences subjective relative likelihoods are derived. Three types of relative likelihoods are identified:

$A \succsim B$, event A is believed to be as least as likely to occur as event B.

$A \succ B$, event A is believed to be strictly more likely to occur as event B.

$A \sim B$, events A and B are believed to be equally likely to occur.

Fishburn (1986) surveys the development of the axiomatic foundations for the rational preference approach from the pioneering era of Ramsey, De Finetti, Savage and Koopman (1940) to the mid-80's. We will repeat here the outcome of these development efforts. The properties suggested in the literature for (measures of) these relative likelihoods that are generally considered uncontroversial are (Fishburn 1986):

- asymmetry: If $A \succ B$, then NOT $(B \succ A)$,
- nontriviality: $\Omega \succ \varnothing$,
- nonnegativity: $A \succsim \varnothing$,
- monotonicity: If $A \supseteq B$ then $A \succsim B$,
- inclusion monotonicity: If $(A \supseteq B, B \succ C)$ OR $(A \succ B, B \supseteq C)$ then $(A \succ C)$.

The following five axioms for the relative likelihoods, especially additivity, have been more challenged in the literature for their desirability:

- transitivity: If $A \succ B$ AND $B \succ C$, then $A \succ C$),
- additivity: If $A \cap C = \varnothing = B \cap C$, then $A \succ B \Leftrightarrow A \cup C \succ B \cup C$,
- complementarity: If $A \succ B$, then NOT $(A^c \succ B^c)$,
- comparability: A and B are comparable if either $A \succsim B$, $B \succsim A$, or both hold.
- consistency: $(A \sim B \Leftrightarrow A \succsim B$ AND $B \succsim A)$, and $(A \succ B \Rightarrow B \not\succ A)$.

Now, if for all pairs of events a weak preference $\succsim$ is given, and all pairs of events are comparable, transitive and consistent, this is referred to as a weak order. A weak order that is nontrivial, nonnegative and additive entails a measure of uncertainty that is unique up to rescaling. When normalised to unity this measure obeys the axioms of probability. The rational preference approach to probability thus defines someone's preference ordering to be rational if all the properties needed to derive a probability measure from the ordering are satisfied.

The assumption of comparability has met with considerable debate. It requires that all events can be meaningfully compared, even if they appear very different. Many authors believe this cannot always be done (French 1986).

Propensity interpretation: the interpretation of probability as a physical 'propensity', or disposition or tendency of a given type of physical situation to yield an outcome of a certain kind was developed by the philosopher Karl Popper (Popper 1957, Popper 1959). Like the relative frequency interpretation, probability as propensity views probability as a physical property 'in the world'. Yet, in the propensity interpretation also single, non-repeatable events can be evaluated.

2.6.1 Learning from Observations

There are two main approaches to learning (probabilistically) from observations today (Hacking 2001, p.190), one based on 'laws of large numbers' and one based on Bayes' rule from Section 2.4.2. The first draws inferences using the way in which relative frequencies stabilise as the number of trials increases. Typical notions in this approach are significance and confidence (Lindley 2000, Hacking 2001). Significance expresses how surprising observed data is if a particular hypothesis about the state of the world is considered to be true. For this an appropriate probability model linking the observations with the hypothesis is needed. If the data is very surprising (e.g. the probability of observing data which is that extreme or more extreme is very low), one will have little faith in the hypothesis being true.

Confidence regards the reliability of the probabilistic method used, rather than the specific inference made. Suppose, based on the underlying probabilistic model, a 95%-confidence interval is derived for a specific quantity of interest. This then is *not* to be interpreted as a 95% probability that the true value of the quantity lies in that interval. Rather in 95% of the cases for which a 95%-confidence interval is determined the true values for these quantities should lie in those intervals specified. Ronald A. Fisher and Jerzy Neyman have contributed greatly these 'law of large number' based inference methodologies.

For subjective or (inter)personal interpretations of probability Bayes rule is used to update the 'degrees of belief' in the light of new observations. In this approach the probability of a hypothesis H of interest after having observed E (for evidence), $P(H|E)$, can be directly derived from the probabilities $P(H), P(E|H)$ and $P(E|H^c)$, assessed before learning E (Equation (2.3) Section 2.4.2). $P(H)$ is commonly referred to as prior probability in this context, $P(E|H)$ and $P(E|H^c)$ as 'likelihoods' and $P(H|E)$ as posterior probability of H after learning E. For 'degree of belief' interpretations of probability the prior probabilities and the likelihoods can be directly assessed (e.g. no infinite amount of observations is needed to determine the limiting relative frequencies), and Bayes rule follows either from the probability axioms or from first principles (see Section 2.4.2). Hence Bayes rule then prescribes how to adjust the probability of H after observing E. This type of inference is therefore also referred to as Bayesian, and its users as Bayesians.

So where 'relative frequentists' cannot assess the probability of the hypothesis being true given the observed data, subjectivists can. That is, when an additional assumption is made. When someone states his prior probability and likelihoods at some time t, and learns E at some later time $t + 1$, then his beliefs about H at time $t + 1$, $P_{t+1}(H|E)$ are not necessarily the same as his posterior probability $P_t(H|E)$, which is assessed at time t. By adopting the Reflection Principle (Fraassen 1984, Fraassen 1995), which states a certain demand for 'diachronic coherence' imposed by rationality, the two can be equated.

A second difference between frequentist and subjective methods of learning from data is that in subjective methods prior beliefs and observed data are weighted, whilst in frequentist methods only the data is evaluated. It has been pointed out though (**?**) that there is subjectivity in the way the data is evaluated, such as judgements about an appropriate reference population for the data.

Frequentist and subjective learning from observations can also be linked, through the concept of exchangeability; a (finite or infinite) sequence of random variables is said to be exchangeable if the probability of any finite vector of outcomes taken from that sequence is unchanged when the order of outcomes is altered. De Finetti (1974) shows that in the case of a long sequence of exchangeable events, an individual will revise his subjective probability such that it will converge to the probability distribution found when the random variables are regarded to be independent and identically distributed. A subjectivist that considers the outcomes of sequential throws of a coin as exchangeable and who conditions his beliefs of the probability of 'heads' in a throw conditional on outcomes of previous throws, will find his beliefs to track the observed relative frequency of 'heads'. Lindley (2000) points out that exchangeability is a subjective judgement. A person that does not regard sequential throws of a coin to be fully exchangeable will therefore thus learn less from observed outcomes than someone who does.

In this section we have discussed the most predominant interpretations of probability. In the relative frequency, prevision and rational preference interpretations probability is related to observable phenomena. For the logical en propensity interpretations this is not the case. In the Sections 2.3.1 and 2.5 we have seen that expectation can be both operationally defined from first principles, and indirectly using probability.

2.7 Summary

In this chapter we have discussed the notion of uncertainty. We have introduced two measures of uncertainty, probability and expectation, and have shown how these measures can be operationalised using subjective assessments. The way these measures are operationalised, or interpreted, determines (and thus limits) the situations/problems to which they can be applied. We have also shown how expectation can be defined using probability as the primitive notion, and vise versa. We conclude with an observation made by Whittle (1992, p.47) about the duality of the two measures: "[...] the recognition by pure mathematicians that the linear functionals constituted by integrals with respect to a measure are objects which are technically dual to that of a measure, and offer all the advantages as the prime concept that we are claiming".

Chapter 3

Elicitation of Quantitative Expert Judgements

In the previous chapter we have discussed how uncertainty can be measured. Regarding judgements, we have described what 'behaviour' can be employed to measure someone's uncertainty about quantities of interest. Yet, this behaviour is susceptible to all sorts of influences, some of which emerging from the interaction between the analyst and the person who's judgements are queried for. In this chapter we will discuss the difficulties, pitfalls and dangers that might rise when we actually try to conduct measurements of uncertainty with experts; the elicitation of quantitative expert judgements. First, in Section 3.2, we will discuss the role human memory plays when experts formulate the judgements they are asked to provide. In Section 3.3 the heuristics and biases are discussed that might be encountered in expert judgement elicitation. Section 3.4 treats the problem of how we can evaluate whether expert judgement elicitation has been successful. Different methods for the elicitation of probabilistic summaries are discussed in Section 3.5, focussing on the elicitation of means, variances and covariances; the assessments needed to quantify a Bayes linear model. The last section will treat the literature on how different experts' assessments can be aggregated.

First we will define what we mean by elicitation in this thesis.

3.1 Elicitation

In this thesis we shall simply write elicitation when referring to the elicitation of expert judgement. A broad definition of elicitation is given by Meyer & Booker (2001). They define elicitation as 'the process of gathering the expert judgement through specially designed methods of verbal or written communication'. Garthwaite, Kadane & O'Hagan (2005) review the literature on the subjective assessment of uncertain quantities, and are consequently more restrictive when defining elicitation as 'the process of formulating one's knowledge about uncertain quantities in the form of a (joint) probability distribution for those quantities'. We will also discuss elicitation with regard to the subjective assessment of uncertain quantities. Since we consider both probability and expectation as primitive concepts for measuring uncertainty we will take a slightly broader view on elicitation, regarding it as the process of formulating one's knowledge about uncertain quantities in the form of an operationally defined measure of uncertainty for these quantities.

3.2 Personal Knowledge

In the previous chapter we have described what 'behaviour' of experts can be employed to measure their uncertainty about quantities of interest as probabilities or expectations. When eliciting either probabilities or expectations from experts, the implicit presumption is of course that these probabilities and expectations capture some of the knowledge the experts have about the uncertain quantities they are assessed for. In this section we will first briefly discuss different views on what constitutes this knowledge and how it can be 'accessed'. The aim here is not to treat these views in detail, but to focus on the implications that can be derived from these views about how to perform the elicitation. We will then proceed to define what we mean by 'expert' and how we regard and treat the judgements experts can provide us with.

3.2.1 The Role of Memory

Much has been written about human memory but the process of human cognition, the mental activity when a person processes information, is not well understood yet (Meyer & Booker 2001). Two current models from cognitive psychology that describe human memory are the 'fixed image' model and the 'recategorisation' model. The fixed image model, the more traditional, distinguishes two types of memory: the long-term memory (LTM) and the short-term memory (STM). The STM refers to information that has just been received and is being processed and is heavily restricted by its limited capacity. The large capacity LTM is thought of as the repository of one's knowledge. To demonstrate the working of the LTM, Hogarth (1987) gives the example of memorizing the following sequence of letters: NBRRYNLGPTVC. Many people would find it difficult to perfectly memorise this sequence in a short time, say 10 seconds. But knowing a 'code' for this sequence, e.g. that this sequence can be reproduced by taking the third letters of the months of the year, would make it considerably easier to memorise the sequence. In this view people are considered to construct their own informal 'codes' for remembering information. So in this model memory does not work by remembering what is actually recalled, but memories are rather *reconstructed* from fragments of information and the use of these codes.

Hogarth gives an interpretation of the fixed model that has important implications for the elicitation of expert judgement. He writes that people are often capable of memorising far more information that is personally relevant to them, than information that is novel to them. Their perception of information is therefore selective rather than comprehensive. As an example of this Hogarth refers to an experiment in which words were briefly flashed on a screen in front of subjects. The subjects who were disallowed to eat for some time before the experiment, were reported having seen more food-related words than subjects who had eaten as usual. A second hypothesis in this model is that information that involves personal experience or observation of incidents remains more important than less concrete information in memory. From these first two hypotheses we can derive that it might be wise to consult different experts, to diminish possible biases due to these personal preferences and imbalanced experiences. The use and aggregation of multiple experts' assessments is discussed in Section 3.6.

Thirdly, because of the limited capacity of the STM (Miller 1956, Cowan 2005), people can often not perform some sort of 'optimal' information processing but are bound to make use of heuristics and cognitive simplification mechanisms. It is important to be aware of these for they can lead to biases in the elicited judgements. These heuristics and biases are more extensively treated in Section 3.3.

Rosenfield (1988) introduces the newer recategorisation model. In this model the process of recollection is represented as the recategorisation of groups of brain cells with temporarily strengthened connections (Meyer & Booker 2001). How people perceive so called stimuli depends on how these have been categorised. Meyer and Booker give the following example to demonstrate this idea. When someone says 'glay shrip' in an experimental setting an English speaking subject may hear the words 'gray chip'. But when the subject subsequently learns the words were spoken by a sea captain, she might recollect hearing 'gray ship'. Thus, also in this model memory does not function as some sort of exact recollection process, but “it helps us reconceptualise the world according to our beliefs, needs or desires” (Meyer & Booker 2001).

Meyer & Booker (2001) give two implications of the recategorisation model to be aware of when making use of expert judgement. Firstly this model helps explaining the changes in peoples thinking process in time through the recategorisation of their knowledge. Secondly, the difference between the thinking process of an expert and a nonexpert is explained as a difference in the number and interacting of associations.

Both models of human memory discussed here imply that when we ask an expert to give her assessment for an uncertain quantity, we do not ask for a quantity that the expert has ready ‘in his head’ to be ‘extracted’. We will discuss the different views on this from the subjective probability literature in the next subsection.

3.2.2 Expert Judgements

We have already been using the term ‘expert’, but have yet to describe for whom we use this designation. Ideally an expert is someone who is deemed knowledgeable and experienced in the subject area of the quantities to be elicited. We will follow Garthwaite et al. (2005) however in using the term expert here for someone who is not necessarily more than just the person whose knowledge we wish to elicit. For the elicitation we distinguish two types of expertise: normative and substantive expertise (notions introduced by Winkler & Murphy (1968)). Normative expertise refers to the ability of the assessor to express his opinion in the desired (probabilistic) response mode, while substantive expertise refers to his knowledge about the subject of interest. So the subject matter experts on e.g. ballistic missile interception selected for a study on the properties of a new type of missile might very well be primarily selected based upon their substantive expertise, and thus require us to be aware that these experts do not necessarily have sufficient normative expertise, and might need training in the desired response mode.

Adherents of the Bayesian approach (see Section 2.6.1) to learning from expert opinion regard expert assessments as observations for which a prior distribution needs to be constructed. In this approach a decision maker who is informed by an expert’s assessments acts as a ‘supra Bayesian’ by expressing her beliefs about the expertise and intentions of the expert in the form of a prior distribution for the expert’s judgement. The problem with this approach is, especially when multiple experts are queried, that “it is thus the seemingly impossible task of this supra Bayesian Decision Maker to evaluate the individuals, their prior information sets, the interdependence of these information sets, the experts' [...] honesty, etc" (Genest & Zidek 1986).

Two approaches to avoid this task have been proposed by Jeffreys, the invariance approach, and by Jaynes, the maximum entropy (or: minimum information) approach. Jeffreys suggest a parametric prior distribution that is invariant under reparametrisation, commonly referred to as Jeffreys prior. The maximum entropy approach aims to incorporate as minimal information as possible to the prior distribution, while complying with the information that has been specified. The maximisation of entropy is always taken relative to some background or reference distribution however. The choice of the background distribution thus has influence on the resulting distribution and thereby adds information to it. A critique to these two approaches can be found in (Seidenfeld 1979).

Finally we return to the debate in the subjective probability literature that is related to this: the problem whether personal probabilities and expectations are precise. If probabilities and expectations are not numbers 'in one's head' waiting to be elicited, does there exist such a thing as a 'true' personal probability for an event or expectation for an uncertain quantity?

Winkler (1967) argues that there is not. He holds that an expert, or any other individual for that matter, does not have a built-in probability distribution for uncertain quantities of interest. Winkler views the assessor as having certain prior knowledge which is not easy to assess quantitatively. Therefore probabilistic expert judgement elicitation would not be about extracting some 'true' subjective probability distribution from the expert, but rather about expressing the expert's beliefs in the form of a probability distribution. An opposing view is that of O'Hagan (1988), who defines 'true' probabilities as those that would result if the expert were capable of perfectly accurate assessments of her own beliefs. O'Hagan regards differences that might be encountered in stated probabilities, e.g. as the result of different elicitation methods, as due to more or less inaccurate attempts to specify the expert's underlying 'true' probabilities. Though in (O'Hagan, Buck, Daneshkhah, Eiser, Gathwaite, Jenkinson, Oakley & Rakow 2006) he or his co-authors acknowledge that someone's estimate (not necessarily a probability estimate) is not 'sitting' in memory waiting to be retrieved: "It is something constructed from the ideas and associations that come to mind while the respondent thinks about how to answer the question".

As indicated in the previous subsection, research in the field of psychology has pointed out important limitations of human cognitive abilities we have to be aware of when conducting an elicitation.

3.3 Heuristics and Biases

Well-known research in the area of human limitations in assessing uncertain quantities is the work done by Kahneman and Tversky on heuristics and biases (see e.g. (Kahneman, Slovic & Tversky 1982), and (Hogarth 1987) for an extensive listing of heuristics and biases). These authors have identified several heuristics people might use when assessing uncertain quantities, and have shown that these heuristics can lead to biased assessments. More recent treatments of types of biases can be found in (Cooke 1991, Meyer & Booker 2001, Garthwaite et al. 2005). Williams (2010) provides an extensive treatment of how heuristics can lead to biases in military decision making. We will repeat some of the examples he gives in the current section to illustrate the effect the heuristics can have.

Meyer & Booker (2001) define bias as 'skewing of the expert judgement from some reference point'. They distinguish two different types of bias, with corresponding reference points. The first, cognitive bias, refers to the skewing from the standpoint of logical rules, the failure of the judgement to comply with specified logical rules. The second, motivational bias, occurs when the elicited judgement is skewed from the personal beliefs of the individual whose judgement is solicited for. This occurs when what someone says is different from what he truly believes. Social pressure and wishful thinking are two possible reasons for people to depart from stating their true beliefs. Motivational biases can also occur through misinterpretation of the elicited judgement by either an interviewer eliciting judgements or an analyst processing the judgements (Meyer & Booker 2001).

Kahneman and Tversky have performed much research on cognitive biases. One of these cognitive biases is the bias caused by the anchoring and adjustment heuristic. When estimating an unknown quantity this bias can occur when people use the heuristic of first fixing an initial value and then adjusting this value to arrive at a final estimate. Research has shown that this adjustment is typically too small (see (Garthwaite et al. 2005) for examples). The British Armed Forces exploited the effect of anchoring in World War II with the deception scheme called the Cyprus Defence Plan:

"Following the German seizure of Crete, the British were concerned that the 4,000 troops on Cyprus were insufficient to repel a German attack. Via the creation of a false division headquarters, barracks, and motor pools along with phony radio transmissions and telegrams, the British set out to convince the Germans that 20,000 troops garrisoned the island. A fake defensive plan with maps, graphics, and orders was passed via double agents [in] a lost brief-

case. The Germans and Italians fell for the ruse. This deception anchored the Germans on the 20,000 troop number for the remaining three years of the war. In spite of their own analysis that the number might be too high, intelligence intercepts and post-war documents revealed the Germans believed the number almost without question. This exposes another negative effect of anchoring: excessively tight confidence intervals. The Germans were more confident in their assessment than justified when considering the contradictory information they had. In summary, the Germans were anchored, made insufficient adjustments and had overly narrow confidence intervals." (Williams 2010)

A second judgemental heuristic identified by Kahneman and Tversky is the availability heuristic, which refers to the case when someone bases his estimate of the frequency or probability of an event on the ease with which he can recall information in favour of this event. As discussed in the previous section the ease with which people can recall information can be distorted by the way information is memorised. Personal and emotional involvement can thus lead to biased assessments via the availability heuristic:

"For example, the subjective probability assessment of future improvised explosive device (IED) attacks will most likely be higher from a lieutenant who witnessed such attacks than one who read about them in situation reports. Bias in their assessment occurs because the actual probability of future attacks is not related to the personal experience of either officer." (Williams 2010)

The representativeness heuristic is related to eliciting conditional probabilities like $P(A|B)$, the probability that event A will occur when it is given that B already has occurred. The representativeness bias arises when people base their assessment of a conditional probability on the degree of similarity between the events A and B. This heuristic fails to distinguish between $P(A|B)$ and $P(B|A)$ while these probabilities are typically not equal. Using Bayes Theorem: $P(A|B) = \frac{P(A)}{P(B)}P(B|A)$. People using this representativeness heuristic thus ignore the so called 'base rates' P(A) and P(B). Therefore this bias is also frequently referred to as the 'base rate fallacy'. As an illustration of bias due to a base rate fallacy, consider the following example:

"While on a platoon patrol, you observe a man near a garbage pile on the side of a major road. In recent IED attacks in the area, the primary method of concealment for the device is

in the numerous piles of garbage that lay festering in the street (trash removal is effectively non-existent due to insurgent attacks on any government employeeincluding sanitation workers). You immediately direct one of your squad leaders to apprehend the man. Based on S2 [intelligence] reports, you know that 90 percent of the population are innocent civilians, while 10 percent are insurgents. The battalion S3 [operations officer] recently provided information from detainee operations training—your platoon correctly identified one of two types of the population 75 percent of the time and incorrectly 25 percent of the time. You quickly interrogate the man. He claims innocence, but acts suspiciously. There is no IED in the trash pile. What is the probability that you detain the man and that he turns out to be an insurgent rather than a civilian? Most cadets answered between 50 percent and 75 percent. This estimate is far too high. The actual probability is 25 percent." (Williams 2010)

Two important biases not directly related to heuristics people use when making assessments are the overconfidence and the hindsight bias (see e.g. (Garthwaite et al. 2005)). The overconfidence bias refers to a too narrow assessment of central confidence intervals, i.e. when assessing a p% central confidence interval for quantities, less than p% of the post hoc observed realisations appear to be in the assessed intervals. The hindsight bias can arise when people are asked for the prior probability of an event for which it is already known whether it has occurred or not. Research has pointed out people's tendency to overestimate their prior probabilities of events they think to have actually occurred and underestimate their prior probabilities for event they think not to have occurred (see (Garthwaite et al. 2005) for references).

Cooke (1991) points out that probability and expectation assessments can also be distorted if the assessor wrongfully acts as if he has some sort of control of the situation. To explain this so called control bias, Cooke refers to an experiment described in (Langer 1975). In this experiment office workers in New York were given the opportunity to buy a ticket for an office lottery for \$1, which gave them a chance of winning \$50. A first group of 26 subjects was allowed to draw their tickets themselves from an urn; a second group of 27 were just given their tickets. When asked for which price the subjects would be prepared to sell their ticket to someone else, whose identity was unknown to the subjects, the median selling price of subjects in the group that was allowed to draw their own ticket was \$8.67, against \$1.96 in the other group. Clearly the subjects' selling prices are more an expression of their willingness to sell their ticket, influenced by e.g. curiosity, than of their subjective beliefs of actually winning the lottery.

Finally, Meyer & Booker (2001) also identify an inconsistency bias, referring to inconsistent statements subjects can give like contradictions. This type of bias can and should be identified and dealt with during the elicitation process.

Unlike the inconsistency bias, the other above discussed biases are very hard to identify, especially motivational biases. Many authors (e.g. (Hogarth 1975, Hogarth 1987, Meyer & Booker 2001, Garthwaite et al. 2005)) stress the influence of the elicitation method on the potential occurrence of biases. Therefore the elicitation process should be designed such as to avoid biases from appearing in judgement as much as possible. Meyer & Booker (2001) state that experts are subject to the same biases as others if the right preventive measures are not used in elicitation. Further, they observe that approaches to handling biases are rare and still in their early stages; two such approaches are given in (Cleaves 1986) and (Meyer & Booker 2001).

3.4 Evaluation of Elicitation

It has become clear from the previous sections that the elicitation of quantitative expert judgement is not a straightforward activity. The quality of the elicited judgements will often strongly depend on how the elicitation is conducted. But what exactly do we mean by the 'quality' of the elicited judgements, how can we assess this?

In first instance many people will probably regard the elicitation of assessments of uncertain quantities as successful if these assessments correspond to later observed values of these quantities. Yet, at the time of the elicitation we probably do not have these observations available, or else we would most likely have used these instead of the expert judgements. Also in our definition of elicitation we do not mention observations. Instead, we regard the elicitation as successful if it truthfully represents the expert's beliefs. Garthwaite et al. (2005) express this idea when they write for the elicitation of a probability distribution: "An elicitation is done well if the distribution that is derived accurately represents the expert's knowledge, regardless of how good that knowledge is".

O'Hagan et al. (2006) also hold that any evaluation of an elicitation should measure the extent to which an expert's knowledge and opinions are faithfully represented in probabilistic form. That is, how close the elicited probabilities are to the expert's 'true' probabilities. Whether these 'true' probabilities exist or not (see discussion in Section 3.2.2), there is no way of determining with certainty what they are. So in that sense, we cannot measure directly whether resulting assessments faithfully represent the expert's beliefs. But a prerequisite for a faithful representation of the expert's beliefs is that the expert should have a full and unambiguous understanding of the precise meaning of the assessments she is asked to give.

Coherency, introduced in Section 2.3.1, also constitutes a useful evaluation tool for the faithfulness of elicited values. Recall that a set of assessments is said to be coherent if no so called 'Dutch Book' can be made using these, i.e. if no set of bets using these assessments can be constructed such that you would be a sure loser if you would bet accordingly. If an expert agrees he would not engage in a bet that would result him a sure loss, then an incoherent set of assessments cannot be a faithful representation of the expert's beliefs. Thus, coherency checks of assessments can be used to evaluate elicited probabilities and expectations. Yet, coherency is by all means no guarantee for a faithful representation of the expert's beliefs.

Summarising, the elicitation is successful when the resulting assessments are coherent and faithfully represent the expert's knowledge and opinions. In Section 3.4.2 we will introduce the concept of proper scoring rules that can be employed to motivate experts to state their true beliefs. From a practical perspective the assessments are often the most useful if they correspond to (post hoc) observed values. In Section 3.6.2 we will discuss a model for aggregating expert judgements based on how well the experts' assessments of test quantities correspond to observed values for these quantities. Yet there is another, urgent, reason for desiring expert judgements to correspond to observations: scientific methodology. This will be discussed in the next subsection.

3.4.1 Cooke's Principles

Cooke (1991) suggests that any scientific study that uses (quantitative) expert judgement and aims to achieve a rational consensus should comply with the following five principles:

Reproducibility. It must be possible for scientific peers to review and if necessary reproduce all calculation. This entails that the calculational models must be fully specified and the ingredient data must be made available.

Accountability. The source of expert subjective probabilities (expectations) must be identified.

Empirical Control. Expert probability (expectation) assessments must in principle be susceptible to empirical control.

Neutrality. The method for combining/evaluating expert opinion should encourage experts to state their true opinions.

Fairness. All experts are treated equally, prior to processing the results of observations.

The requirement of reproducibility is standard to scientific practice. Accountability to the decision maker enhances the quality and credibility of the study, but might be quarrelsome to achieve in practice. Cooke gives the example of experts judging the reliability of a new technical system, that might be found in the employ of contracting firms designing the system in question. This example is particularly relevant for Defence. The decision maker, in this case Defence as the contractor, might wish to consult experts working for a contracting firm on e.g. the reliability of a system considered for acquisition. Yet the experts' assessments on the reliability of the system might be in conflict with the aim of their firm to sell the system to Defence, especially in the earlier stages of the acquisition process when the information gathered from the contractor is not binding. The experts therefore might insist on anonymity as a prerequisite to provide their genuine assessments. French (2011) acknowledges this issue of experts insisting on anonymity and mentions an additional one: when the opinions of experts are valued not only for their expertise, but also for who they are in the public's eyes, anonymity might also be desirable. Empirical control allows a study to be falsifiable in principle. Especially for Defence it is of high importance to identify as soon as possible when the assessments given by experts do not correspond to actually observed values. In the end they are interested in the actual values of e.g. reliability and performance of their systems. The principle of neutrality is fully in line with the aim of the elicitation to faithfully represent the expert's knowledge, argued earlier in this section.

The principle of fairness however is more controversial. The supra Bayesian approach mentioned in Section 3.2.2 for example requires a decision maker to express his beliefs about the expertise and intentions for each of the experts. Cooke (1991) acknowledges that some experts are always preferred to others by the act of selecting experts to participate in the study, but argues that "these decisions must be made initially on the basis of factors that cannot be meaningfully translated into numerical input" needed in the Bayesian approach.

French (2011) discusses the applicability of Cooke's principles and, being Bayesian, doubts the persuasiveness of the fairness principle. He distinguishes three situations in which opinions of groups of experts are employed which he refers to as (French 1985): the expert problem, the group decision problem and the textbook problem. The first type refers to the situation in which one decision maker wishes to be informed by a group of experts, the second type to the situation where there is a group of decision makers that are their own experts. Thirdly, in the textbook problem, a group of experts may just be required to give their judgements for others to use in future undefined circumstances. French (2011) points out that "the textbook problem is gaining in importance since data and expert judgements can be made available over the web to be used by many different individuals to shape their own beliefs in many different contexts". He argues that Cooke's principles need more discussion if expert judgement studies are to inform more public deliberation.

3.4.2 Scoring Rules

Cooke (1991) describes scoring as a numerical evaluation of probability assessments in the light of corresponding observations. Lad (1996) defines scoring rules in the broader context of assessments of expectations of uncertain quantities. In the literature two purposes can be identified for the use of scoring rules: to motivate people to make truthful and well-considered assessments and to evaluate the accuracy of their assessments (Cooke 1991, Morgan & Henrion 1990). Since better scores both result from people making more accurate assessments and having more knowledge (Garthwaite et al. 2005), a scoring rule evaluates both normative and substantive expertise.

Assessors can be motivated through the scoring system: the higher the score, the better their assessment is perceived. To satisfy the principle of neutrality described in the previous subsection and to comply Winkler's definition of successful elicitation (i.e. assessments representing 'true' beliefs of their originators), proper scores can be employed. A scoring rule is said to be proper if the assessor's expected score reaches its maximum when the assessor states his true beliefs.

Bolger & Wright (1993) identify two major approaches to appraising subjective probability judgement in their review of rival models of probability judgement, which are well rehearsed throughout the expert judgement literature: coherency (see Section 2.3.1) and calibration. Incoherent judgements lead to an inconsistency bias (Section 3.3). Calibration measures the extent to which an individual's stated probabilities correspond to observed relative frequencies and is traditionally measured in either of two situations: when eliciting discrete probabilities and when eliciting quantiles of a continuous distribution. Roughly speaking an assessor is called well-calibrated for probability value p if for the events that the assessor has assigned probability value p, the relative frequency of occurrence of the events is equal to p. Then an assessor is well-calibrated if he is well-calibrated for every probability value p. Through comparing subjective probabilistic judgement with observations, calibration can be seen as a form of empirical control over these judgements (Cooke 1991).

In addition to coherence and calibration, Cooke (1991) also discusses an information score to evaluate probabilistic assessments. Information is defined as the negative of entropy, and entropy is a measure for the spread of the probability mass for an event or proposition of interest. The more uniform the spread over all probabilities for an event, the less information can be obtained about whether or not the event is likely to occur. Therefore when all other things are equal, more informative assessments are preferred above less informative ones. A scoring system employed in practice that scores assessors on both calibration and information is the so-called classical model from Cooke (1991). The classical model will be introduced in Section 3.6.2.

A good mathematical treatment of (proper) scoring rules is given by Lad (1996), discussing the differing properties various scoring rules can have. By analysing these properties insight can be gained in the appropriateness of the use of a scoring rule in specific situations. Critical in the use of scoring rules in practice is that they must be sufficiently understood by the assessors (Hogarth 1975).

Though widely applied in evaluation (scoring) schemes, there has been a wide debate as to whether observations can be used to evaluate subjective probability assessments, at least dating back Bonferroni in 1925, that has still not been completely settled. The original argument focused around the question of whether relative frequencies could be used to evaluate subjective probability. Where some (e.g. Bonferroni, Fréchet) held that only subjective probabilities that are equivalent to relative frequencies can be regarded as valid, others (e.g. De Finetti) held that subjective probabilities regarding a sequence of quantities represent the state of uncertainty of the assessor about these. Therefore they considered making judgements about the correctness of these assessments in the light of observed outcomes to be meaningless, in the same way that resulting posterior probabilities should not be seen as corrections of previous judgements, but as conditional on a different state of information (Lad 1996).

3.5 Eliciting Quantitative Subjective Uncertainty

In this section we will discuss people's abilities to provide probabilistic assessments, i.e. assessments of probabilities or of summaries of probability distributions. After briefly treating more general results we will focus on the elicitation of means, variances and covariances; the assessments needed to quantify a Bayes linear model. A good evaluation of elicitation methods to quantify Bayes linear models is (Revie, Bedford & Walls 2010). We take the stance that people should be asked about quantities that are observable (in principle) only, Principle 3 from Section 1.1, with which many authors agree (see e.g. (Cooke 1991, Garthwaite et al. 2005)).

Much of the research on the elicitation of probabilistic quantities has been carried out in artificial laboratory settings. It is therefore questioned to what extent the results from many of these experiments are generalisable (see e.g. Winkler in his comments on (Hogarth 1975, p.290)), although these experiments are useful in anticipating which inconsistencies or biases might be encountered. Ferrell (1994) concludes that the procedures for the elicitation of probabilities are not based on a theory of judgement or on detailed knowledge of the process by which people 'produce' a subjective probability. At the same time however, he does regard the procedures for reducing bias and error to be sophisticated and well grounded in empirical observation.

Many experiments have pointed out that people often behave as 'conservative' information processors (see (Hogarth 1975) and (Garthwaite et al. 2005) for references). When people are asked to adjust their probabilities after receiving new information in the form of observations, these adjustments often appear to be 'conservative': the revised probabilities are closer to the prior probabilities than when Bayes rule (see (2.3), Section 2.4.2) is applied. This phenomenon could be explained with the anchoring and adjustment bias described in Section 3.3, where the prior probability is taken as 'anchor'. Hogarth (1975) references several researches that indicate that people process information in a fundamentally different way than Bayes rule.

People also appear to have difficulties with assessing extreme values. Garthwaite et al. (2005) point out that experiments show that subjects make poor assessments of extreme 'tails' of distributions, like when assessing e.g. 1%- and 99%-quantiles (the p%-quantile being the value $x_{\frac{p}{100}}$ for which $P(X \leqslant x_p) = \frac{p}{100}$). As a possible explanation for this they state that comparisons might possibly not come readily to mind when making assessments of unlikely events. A similar explanation is given by Hogarth (1975), who therefore suggest avoiding the use of the ends of the probability scale.

3.5.1 Elicitation of Means and Variances

Means and variances can both be derived from first and second order moments. The mean is simply the first moment of a quantity. The variance is equal to the second moment of a quantity minus the square of the first moment: $Var(X) = E(X^2) - E(X)^2$.

In their review of experiments on people's abilities to assess probabilistic summaries, Garthwaite et al. (2005) find that subjects' estimates of the mode, median and mean of samples show a high degree of accuracy when the distribution sampled from is approximately symmetric (a situation in which mode, median and mean are fairly similar). When sampling from a highly skewed distribution, estimates of modes and medians still showed reasonable performance, but assessments of means appeared to be biased toward the median.

In Section 2.5 we have discussed that the mean of an uncertain quantity can be directly assessed as the prevision for that quantity. Lad (1996) argues that variances can be assessed in a similar manner using previsions. By acknowledging that functions of uncertain quantities are uncertain quantities themselves, Lad defines the variance of a quantity X to be the prevision of the quantity $[X - Pv(X)]^2$, where $Pv(X)$ is someone's prevision for X. Yet, it is very questionable to what extent people will be able to provide these type of previsions. We do not regard variances in general to be observable quantities.

Many authors have found that in general it is not a good idea to have expert assess higher order moments of an uncertain quantity directly (Gokhale & Press 1982, Morgan & Henrion 1990, Kadane & Wolfson 1998), although also more positive experiences exist (Clemen, Fischer & Winkler 2000) (see Section 3.5.2). Even assessments of first moments (means) should be treated carefully, as Peterson & Miller (1964) have shown they can be biased towards the median for highly skewed distributions. Avoiding the direct assessment of higher moments is in agreement with the argument mentioned at the beginning of this section to ask experts only questions about observable quantities, considering people usually do not directly observe higher moments.

Keefer & Bodily (1983), Keefer & Verdini (1990), Smith (1993) and more recently Reilly (2002) provide excellent reading on how to derive expert assessments of means and variances indirectly from assessments of quantiles of the variables of interest. The methods evaluated in these articles do not require the assumption of a specific family of probability distributions nor do they require the assessment of its parameters. Surprisingly enough one of the simplest methods, the Pearson-Tukey method, is found to perform best for a large family of distributions.

3.5.1.1 Pearson-Tukey Approximations

The Pearson-Tukey method for estimating the mean and variance from quantile assessments is based on the observation made by C.P. Winsor (Pearson & Tukey 1965) that the ratio of the distances between suitable symmetrical quantiles to the standard deviation, the h%-distance:

$$h\%\text{-distance} = \frac{x_{(100-h)} - x_h}{\sqrt{Var(X)}},$$

where $x_{\frac{p}{100}}$ is the p%-quantile of X, is surprisingly constant for many well-known distributions. Pearson & Tukey (1965) suggest the following approximation of the mean using the 5%-, 50%-

and 95%-quantiles:

$$E(X) \approx x_{0.50} - 0.185\Delta, \qquad \text{with } \Delta = x_{0.95} + x_{0.05} - 2x_{0.50}.$$

The Δ-term is thus an approximation of the difference between the median and the mean of the distribution of X. For an estimation of the variance Pearson & Tukey (1965) developed the following iterative procedure, using 2.5%-, 5%-, 50%-, 95%- and 97.5%-quantiles:

$$Var(X) \approx max(\hat{\sigma}^2_{0.05}, \hat{\sigma}^2_{0.025})$$

where $\hat{\sigma}_{0.05}$ and $\hat{\sigma}_{0.025}$ are iteratively derived as:

$$\begin{aligned}\hat{\sigma}_{0.05} &= \frac{x_{0.95} - x_{0.05}}{max\left(3.29 - 0.1(\frac{\Delta}{\hat{\sigma}_{0.05}})^2, 3.08\right)},\\ \hat{\sigma}_{0.025} &= \frac{x_{0.975} - x_{0.025}}{max\left(3.98 - 0.138(\frac{\Delta}{\hat{\sigma}_{0.025}})^2, 3.66\right)}\end{aligned}$$

starting with the values:

$$\begin{aligned}\hat{\sigma}_{0.05} &= \frac{x_{0.95} - x_{0.05}}{3.25},\\ \hat{\sigma}_{0.025} &= \frac{x_{0.975} - x_{0.025}}{3.92}.\end{aligned}$$

This procedure was tailored such as to give good results across the Pearson and Johnson S_U systems of distributions. Keefer & Bodily (1983) simplified this procedure by not requiring iterations and eliminating the 2.5%- and 97.5%-quantiles:

$$Var(X) \approx \left(\frac{x_{0.95} - x_{0.05}}{3.29 - 0.1(\frac{\Delta}{\sigma_0})^2}\right)^2, \qquad \text{with } \sigma_0 = \frac{x_{0.95} - x_{0.05}}{3.25}.$$

Johnson (2002) explores the accuracy of the Pearson-Tukey approximation of the mean and the modified Pearson-Tukey approximation of the variance by Keefer and Bodily for a range of distributions that the authors assume plausible in a risk analysis context. The average error found for the mean for the sampled Beta1, Beta2, Gamma, Lognormal and Golenko-Ginzburg distributions (154 in total) was impressively small, 0.23%, and the maximum error found was only 0.77%, for a Beta2 distribution. For the variance the average error encountered was 2.8%. The maximum error in variance was 11.6%, for a Golenko-Ginzburg distribution.

Keefer & Bodily (1983) evaluate a large set of three- and five-point discrete approximations to Beta distributions, focussing on the approximation of the mean and variance from these. As one of the three-point approximations the authors evaluate the extended Pearson-Tukey approximation. For this approximation they take the 5%-, 50%- and 95%-quantiles with respectively the probability masses 0.185, 0.63 and 0.185. So in the extended Pearson-Tukey method the moments of X are approximated as:

$$E(X^n) = 0.185(x_{0.05})^n + 0.63(x_{0.50})^n + 0.185(x_{0.95})^n, \tag{3.1}$$

The Pearson-Tukey approximation of the mean outperforms the other 14 approximations considered in the evaluation, with an average absolute error of 0.02% and a maximum observed error of 0.07% for the 78 Beta distributions considered. Other approximations that perform well are the Swanson-Megill method with 0.05%, the extended Swanson-Megill also with 0.05%, the modified Davidson-Cooper method with 0.14%, the Perry-Greig method with 0.37% and the 5-point Bracket Median method with 0.74% as the average absolute error (see (Keefer & Bodily 1983) for references to these approximation methods). For comparison, the original PERT formula for the mean results in an average absolute error of 41.7%.

On the variance approximation both the modified Pearson-Tukey method and the extended Pearson-Tukey method outperform the other approximations even more clearly, with an average absolute error of 0.38% and 0.46% respectively and a maximum error of -1.7% and -1.6% respectively. Other methods have average absolute errors that are much larger, such as the extended Swanson-Megill approximation (2.7%), Moder-Rodgers (4.5%) and Davidson-Cooper (5.6%). The original PERT formula for the variance has an average absolute error of 549%.

3.5.2 Elicitation of Covariances

A covariance can be derived as the product-moment of two quantities minus the product of their means: $Cov(X, Y) = E(XY) - E(X)E(Y)$. Lad (1996) defines the covariance between the quantities X and Y to be the prevision of the quantity $[X - Pv(X)][Y - E(Y)]$, where $Pv(X)$ and $Pv(Y)$ are someone's previsions for X and Y. We repeat here that we do not think it is a good idea to have people assess higher moments directly, or covariances more specifically. Yet, in an experimental study Clemen et al. (2000) did have a positive experience with the direct assessment of correlations, which together with variances determine the covariance.

In this study the authors evaluate six different methods for the assessment of correlations. In these methods correlation is derived from

- an informal description of the strength of the relationship,
- a direct assessment of the correlation,
- a cumulative probability of one variable conditional on a quantile being hit by the other,
- a probability of concordance,
- a joint probability, and
- a probability for one variable conditional on a certain cumulative probability for the other.

The six methods were tested on two groups of MBA students to assess correlations between 'general knowledge' variables as the height and weight of fellow students, for which the true correlations were known. The direct assessment of correlation performed best in this study, and showed less variability than the other assessment method. In addition, the direct assessment method can not lead to infeasible answers and was judged less difficult than alternate methods.

Revie et al. (2010) have tested four different methods for the assessment of covariance on 23 postgraduate students: the Direct Calculation (DC) method, the direct elicitation of correlation (D), the Adjusted Expectation (AE) method and the Adjusted Uncertainty method (AU). In the DC method, the covariance is calculated from assessments of the means and variances, an assessment of the change of the expectation of one variable given that the expectation of the other has changed by a certain amount, with the amount to be chosen by the assessor himself and finally quantile assessments of one variable given that the other is known to be equal to a certain value with certainty, this value again chosen by the assessor. The AE method uses the means and variances, and a direct assessment of the Bayes linear adjusted expectation for a certain value, chosen by the assessor, for the other variable. The AU method, finally, uses means and variances and a direct assessment of the Bayes linear adjusted variance for one variable, given that the other variable is observed to be equal to its mean.

Revie et al. (2010) concluded that for the vast majority of cases, the AU method is unsatisfactory since it does not allow the variance of one variable to increase for any observation that is made for the other variable. The DC method led to the most acceptable results; no incoherencies were encountered, and this method resulted in the least cases in which negative correlations were found for relationship that the assessors had judged to be positive prior to the assessment exercise. The DC method was also judged the most often preferred method, the AU method the easiest. The authors do not report about the accuracy of the covariance assessments for the different methods, i.e. the extent to which the assessments correspond to the actual values of the covariances for the test questions.

In this research we will evaluate bivariate generalisation of the extended Pearson-Tukey method for the assessment of product moments, for three reasons. Firstly, the accuracy reported for the univariate case raises expectations about the performance of the method for product moments. Secondly, the bivariate generalisation uses marginal and conditional quantiles to estimate product moments. Since we consider the univariate method to be one of the best methods available for assessing means and variances, and this method uses quantile assessments, we consider it to be convenient for the assessor to be asked about a similar type of assessment, conditional quantiles, to derive product moment assessments. We note here that in our opinion quantile assessments relate to observable quantities, if the quantities being assessed are observable (in principle) themselves. Finally, in Chapter 6 we need assessments of higher order product moments as well, which can also be provided by the bivariate Pearson-Tukey method. We will introduce the bivariate generalisation of the extended Pearson-Tukey method in Section 6.3, and evaluate its performance there.

3.6 Aggregation of Experts' Assessments

When multiple experts are queried for their quantitative assessments, means of aggregating their opinions are needed. Depending on the desired level of interaction between the experts involved in the elicitation, different settings for eliciting judgement can be considered. On one extreme, no interaction between the experts, individual interviews can be chosen. On the other extreme, group sessions led by a so called moderator allow for full interaction between the experts. A great advantage of group sessions is that it allows for synergistic effects from interexpert discussions, which can lead to more accurate assessments and a greater amount of ideas (Meyer & Booker 2001, Garthwaite et al. 2005). On the other hand, dominant people (e.g. strong personalities, people taking higher hierarchical positions) and people adjusting their responses to what they believe will be acceptable to the group can introduce motivational biases when performing group elicitation. Garthwaite et al. (2005) also point out that judgements based on overlapping experience of the experts can be overweighed by being repeated in discussions. In a military setting this might occur when experts are selected that have been sent out on the same missions.

A further distinction between group sessions and individual elicitation is that group sessions seek for a consensus view, also referred to as behavioural aggregation. In the case where only one resulting view is desired from elicitation, the individually gathered judgements will need to be aggregated post hoc. Usually individual assessments will make use of some form of mathematical aggregation (one of the simplest forms of mathematical aggregation is e.g. taking the average of the different responses). Mathematical aggregation methods are the topic of the next subsection. A mathematical aggregation method that weighs the individual experts' opinions according to their performance on test questions is discussed in Section 3.6.2.

Apart from the already mentioned motivational biases, behavioural aggregation also has the obvious disadvantage that it can either appear impossible to reach a consensus or that the pressure to reach a consensus leads to experts suppressing dissenting opinions (Garthwaite et al. 2005). When a consensus is achieved, it is anonymous in the sense that it is the product of group interaction and can not be accounted to individual experts. Mathematical aggregation in principle allows for individual accountability. A criticism of mathematical aggregation is that it obscures the differences between different opinions and the reasons for these differences. Experts could e.g. have interpreted a question differently and therefore have reached different answers. Or experts could have based their answers on different, but both valid knowledge. Above that, mathematical aggregation could in principle lead to an aggregated answer none of the involved experts agrees with (Meyer & Booker 2001).

Finally, a third extensively used setting in expert judgement elicitation is the Delphi method. The Delphi method was developed by the RAND Corporation for the military during the cold war. The method is designed to mitigate the motivational biases, such as strong personalities or people with higher hierarchical ranks in the military dominating the discussions, typical to interactive group elicitation. Through controlled knowledge exchange the method still aims to benefit from synergistic effects. A typical Delphi elicitation could go as follows: first experts give their judgements separately. The moderator collects this data, makes it anonymous and distributes it back to the experts who are then asked to revise their judgements after receiving this new information (of course they are allowed to keep their judgements unchanged). More than one of these revision rounds are possible. The Delphi method thus forms a compromise between individual and group elicitation. The method does not completely eliminate the possibility of motivational biases to rise from the sharing of judgements of other experts, and due to the less efficient sharing of knowledge the synergy rising from this sharing will also be less than in group sessions (Meyer & Booker 2001, Garthwaite et al. 2005).

3.6.1 Mathematical Aggregation Methods and their Properties

Two of the most popular methods for mathematical aggregation of quantitative assessments are the linear and the logarithmic opinion pool, respectively taking a weighted average and a weighted geometric mean of the experts' assessments (Garthwaite et al. 2005). In both methods, higher weights (and thus a higher influence on the aggregated result) can be given to experts who are believed to make more accurate assessments. In performance based weighting, these weights are derived from the experts' performance on so called 'seed variables', quantities for which the true value is known to the analyst but not to the expert. The classical model of Cooke (1991), introduced in the next section, is an implementation of the linear pool that uses performance-based weighting to combine probability assessments.

Different aggregation methods of course have different properties. There has been an extensive debate in the literature about the different properties mathematical aggregation methods can have (see e.g. (McConway 1981, Genest & Zidek 1986, French 1985, French 1987, Cooke 1991), and about the desirability of each property. We will informally list the most predominant of these properties here.

Marginalisation Property: The same marginal probabilities are found whether (a) the assessors' distributions are first combined to form a single distribution, and then some marginalisation (i.e. restriction to some subspace of the outcome space) is performed on this, or (b) the individual assessors all perform the marginalisation separately, and the resulting individual marginal distributions are combined into a single distribution.

Zero Preservation Property: If all assessors judge an event A to have probability zero, then the combination of their probabilities for A is also zero.

Strong Setwise Function Property: The combined probability for an event A depends only on the probabilities given to A by the individual assessors.

Independence Preservation Property: If all assessors regard two events A and B as independent, then the combined probability for A is also independent of the combined probability for B.

External Bayesian Property: The result of first combining, and then processing the results of new observations via Bayes' theorem is the same as first letting the experts process the results of the new observations and then combining their updated probabilities.

Linear opinion pools have the marginalisation property, but fail to be externally Bayesian, whereas the opposite holds for logarithmic pools. The relative desirability of both these properties will therefore be important when choosing between these combination methods. The arguments in favour of the marginalisation property in a probability context are well rehearsed in (Cooke 1991), though not uniformly accepted (see e.g. (Lindley 1985, French 1985)).

Garthwaite et al. (2005) warn that when the knowledge of some experts substantially overlaps, these pooling methods can lead to what they call double counting of expertise, e.g. when experts having overlapping knowledge are assigned the same weights as other experts. At the beginning of this section it was already mentioned that this double counting of expertise can also occur in interactive group sessions. Garthwaite et al. (2005) therefore suggest to select the experts such that their knowledge is complementary.

Mathematical aggregation methods other than the above described opinion pools are the 'external', or 'supra Bayesian' (see Section 3.2.2), approach (Lindley, Tversky & Brown 1979) and the use of conjugate families of prior probability distributions. These methods are concerned with the aggregation of probability distributions. In the external approach, assessments are treated as 'data' to update prior beliefs. This requires substantial prior beliefs about the experts' opinions to be specified, also referred to as a supra Bayesian prior assessment. When no meaningful prior assessments about the experts' opinions can be given, non-informative priors can be used. In the conjugate family approach, it is assumed that each of the experts' opinions can be represented by a member of a specific family of probability distributions. The family of distributions is chosen such that the aggregation of members of this family will also be a member of the same family of distributions. Though computationally convenient, this method thus makes strong assumptions about the experts' opinions.

In their review Garthwaite et al. (2005) argue that it is not clear what a probability reached by consensus of a group means and whether it is representative for that group's behaviour. They also question whose opinion is represented by a pooling of experts' opinions. In the 'supra Bayesian'-type approaches the aggregated result simply reflects the updated opinion of the person who had stated his beliefs about the experts' opinions, e.g. the decision maker or the analyst conducting the elicitation.

3.6.2 Classical Model

The classical model (Cooke 1991) is an implementation of a linear pool of experts' assessments in which the weights used for each expert is derived from the performance of their assessments on 'seed variables'. Seed variables are variables for which the true value is known to the analyst (and of course not to the assessors), or will become known within a short time. These variables are thus 'seeding the performance based combination model'. Performance-based weighting is based on the assumption that the performance of the expert on the seed variables is in some way informative for how the expert will perform on assessing the variables of interest to the study. Therefore the seed variables should closely match these variables of interest. Cooke, Mendel

& Thijs (1988) indicate that an expert's performance on assessing *general knowledge* questions does not predict his performance on variables in the expert's field of expertise. Goossens, Cooke & Kraan (1998) provide evidence that performance-based weighting, using *application domain specific* test questions, leads to better results than when using equal weights for all the experts.

There is a version of the classical model for uncertain events, and a version for continuous variables. We will introduce the latter here, since this version will be of interest to us in Chapter 7. For the continuous version R quantiles for the cumulative probabilities $f_1, \ldots, f_R$ are elicited from the expert for each of the seed variables $X_1, \ldots, X_N$. The quantiles are ordered such that $0 \leqslant f_1 < \ldots < f_R \leqslant 1$, and in addition f_0 and f_{R+1} are taken to equal 0 and 1 respectively. We will use the notation x_{ire} here for these assessed quantiles, where index i denotes the seed quantity, index r the quantile and e the specific expert. From the definition of the quantiles the theoretical probability $p_r = P(x_i \in [x_{ire}, x_{i(r-1)e}]) = f_r - f_{r-1}$ can be derived for $r = 1, \ldots, R+1$, the probability that the true value x_i for seed variable X_i is between the f_{r-1}%- and the f_r%-quantile of expert e. The notation $\mathbf{p}$ is used here for the vector of these theoretical probabilities $(p_1, \ldots, p_{R+1})$. Finally, the relative frequency with which the true value of a seed question falls between an expert's stated f_{r-1}%- and f_r%-quantile is denoted with s_r, and $\mathbf{s}$ is taken to be the vector $(s_1, \ldots, s_{R+1})$ of these relative frequencies.

The performance of the experts is measured by a score that is a combination of a calibration and an information score. As discussed in Section 3.4.2, calibration measures the extent to which an experts stated probabilities correspond to observed relative frequencies; in the current case the extent to which the observed relative frequencies $\mathbf{s}$ correspond to the theoretical probabilities $\mathbf{p}$. To measure the (dis)agreement between $\mathbf{s}$ and $\mathbf{p}$, the relative information function $I(\mathbf{s}, \mathbf{p})$ is used, which is defined as:

$$I(\mathbf{s}, \mathbf{p}) = \sum_{r=1}^{R+1} s_i \ln \left(\frac{s_i}{p_i} \right).$$

Now, as the number of seed questions N increases, the probability distribution of the variable $T = 2NI(\mathbf{s}, \mathbf{p})$ will approach the chi-quare distribution with R degrees of freedom (Cooke 1991, p.188). The calibration score $C(e)$ of expert e in the classical model is taken to be the exceedance probability of T, where T is taken to be chi-square distributed with R degrees of freedom:

$$C(e) = 1 - \chi^2_R \left(2NI(\mathbf{s}, \mathbf{p}) \right), \tag{3.2}$$

where χ^2_R is the cumulative distribution function of the chi-square distribution with R degrees of freedom. The name 'classical model' stems from the close relation between this calibration scoring and hypothesis testing in classical statistics, and is contrasted with Bayesian aggregation methods.

The information score of the classical model expresses the degree to which the probability mass in $\mathbf{p}$ is concentrated on the possible values of the seed variables, relative to a selected background probability measure. Let f_{ie} be the minimal information probability density function of expert e for seed variable X_i satisfying the expert's quantiles x_{ire}. That is, f_{ie} is piecewise uniform with density $\frac{p_r}{x_{ire}-x_{i(r-1)e}}$ between the $(r-1)^{th}$ and the r^{th} quantile assessed by expert e for variable X_i. To fully determine f_{ie}, also a 0%- and a 100%-quantile, x_{i0} and $x_{i(R+1)}$, are needed. These bounds need to be finite to ensure that f_{ie} is a proper probability density, and need to lie outside the interval $[\underset{e}{min}(x_{i1e}), \underset{e}{max}(x_{iRe})]$ determined by the most extreme quantile assessments of all experts. Further, let f_i be the background probability measure for seed X_i. The classical model information score $I(e)$ is the average relative information of f_{ie} with respect to the background measure f_i over all seed variables:

$$I(e) = \frac{1}{N} \sum_{i=1}^{N} \left[\int_{x_{i0}}^{x_{i(R+1)}} f_{ie}(x) \ln \left(\frac{f_{ie}(x)}{f_i(x)} \right) dx \right]. \tag{3.3}$$

For both the calibration and information score better performance corresponds to higher scores. When the product of these scores is taken, this behaviour is preserved. The unnormalised weight w'_e for expert e is derived as the product of the calibration and information scores, multiplied by an indicator function that provides the opportunity to assign zero weights to poorly calibrated experts:

$$w'_e = C(e) \cdot I(e) \cdot 1_\alpha(C(e)). \tag{3.4}$$

The indicator function $1_\alpha(C(e))$ is zero when the exceedance probability $C(e)$ is below threshold α. So provided that the calibration score is above threshold α, the better either of the calibration or information score of an expert is, the greater the weight assigned to the expert is (unless all other experts have a calibration score below α and the expert is already assigned the maximum weight 1). The performance based weight w_e of the classical model for expert e can now simply be determined by normalisation:

$$w_e = \frac{w'_e}{\sum_e w'_e}, \tag{3.5}$$

trivially requiring that α in Equation (3.4) is chosen such that at least one expert has a nonzero weight. The classical model linear pool with weights from (3.5) for variable X_j, referred to as the decision maker's distribution f_{jDM}, can now be determined and is a function of cut-off value α:

$$f_{jDM}(x,\alpha) = \sum_e w_e(\alpha) f_{je}(x), \tag{3.6}$$

where f_{je} is experts e's minimal informative probability density function for X_j satisfying e's quantiles x_{jre} for this variable, which could either be a seed variable or a non-seed variable of interest for the study. Again, α is assumed to be chosen such that at least one expert has a nonzero weight (note that the expert weights then sum to unity due to normalisation (3.5)).

The proposed procedure for choosing α virtually adds the decision maker linear pool $f_{jDM}(x,\alpha)$ to the linear pool of the experts, and then seeks to maximise the decision maker's virtual weight over α. Let $w_{DM}(\alpha)$ be the weight for the decision maker in a new linear pool f_{jV} with both the decision maker and all the experts, where the experts have their weights derived from (3.5):

$$f_{jV}(x,\alpha) = \frac{\sum_e w_e(\alpha) f_{je}(x) + w_{DM}(\alpha) f_{jDM}(x,\alpha)}{w_{DM}(\alpha) + \sum_e w_e(\alpha)}. \tag{3.7}$$

The weight $w_{DM}(\alpha)$ is called virtual because the linear pool is not changed by adding the decision maker with this weight and then normalising the pool again: $f_{jV}(x,\alpha)$ is equal to $f_{jDM}(x,\alpha)$, as can be verified by substituting Equation (3.6) in (3.7). Now α is varied between 0 and $\max_e(C(e))$ and chosen such that the decision maker's virtual weight is maximised over α:

$$\alpha' = \underset{\alpha \in (0, \max_e(C(e)))}{argmax} \quad w_{DM}(\alpha). \tag{3.8}$$

Cooke however warns against the uncritical use of this procedure for determining the cut-off α. He points out that it might be imprudent to let a very poorly calibrated expert dominate other experts who are even worse (Cooke 1991, p.194), which might be the result of the procedure. On the other hand, the calibration scores tend to go down as well when the number of seed variables increases (Cooke 1991, p.193). So caution and careful consideration of the calibration scores are needed when determining an appropriate α.

3.6.2.1 Properties of the Classical Model Weighting Scheme

The classical model is designed to comply with the methodological principles for the use of expert judgement discussed in Section 3.4.1. Reproducibility and accountability can be achieved by providing scientific peers the unanonymised experts' assessments, although there might be circumstances in which this is less desirable as discussed in Section 3.4.1. Empirical control is achieved in the scoring scheme, in which the experts' indirectly stated probabilities are evaluated against observed relative frequencies. The calibration score (3.2) can also be used to investigate whether any of the experts is calibrated well enough for a decision maker to have faith in the performance of classical model linear pool based on the experts' assessments.

To satisfy the principle of neutrality experts should be motivated to state their true beliefs. The unnormalised weights (3.4) are designed such that an expert who understands this score, can maximise his expected unnormalised weight by stating his true beliefs for the quantiles he is asked to assess. At least, this would be the case if the calibration score would have an exact calculation. Since the distribution of $2NI(\mathbf{s}, \mathbf{p})$ approaches the chi-square distribution arbitrarily closely when N is taken large enough, but remains an approximation for finite N, the unnormalised weight w'_e is referred to as a 'weakly asymptotic strictly proper score' (see (Cooke 1991, Chapter 9) for details). The normalised weight however is not. If an expert expects all other experts to perform so poor that they all receive a zero calibration score, then he might expect small deviations from his true beliefs about the quantiles to be assessed for the seed variables not to matter for his normalised weight. Finally, since all experts are treated equally in the derivation of the weights, the classical model satisfies the principle of fairness.

The expert with the best calibration score (or experts with the best calibration scores if there is no unique best expert) always remains in the classical model linear pool (clmp), since α is always smaller than the best calibration score. This entails that the expert with the highest unnormalised weight also always remains in the cmlp, since this expert must have a nonzero calibration score.

There is no guarantee in the classical model that the cmlp performs better on the seed questions than the best expert, or the equal weights linear pool.

3.7 Summary

In this chapter we have discussed what we asked for when we ask people to make quantitative judgements. We have discussed the heuristics and biases people are susceptible to, and how to evaluate if our elicitation exercise has been successful. We have discussed peoples abilities to make probabilistic assessments, focussing on means, variances and covariances, the quantities that need to be assessed to quantify a Bayes linear model. Finally we treated the aggregation of experts' probabilities, which will serve as a foundation for the development of the performance based aggregation method of moment assessments in Chapter 7.

Chapter 4

The Bayes Linear Methodology

The Bayes linear methodology has characteristics that make it very suitable for expressing and revising quantitative expert judgements about uncertain quantities. The methodology reflects the discrete character of quantitative expert assessments and is flexible in the amount of detail that can both be specified by the experts and is needed for the decision problem at hand. The methodology is assumption free as in that it does not require the quantities to have a probability distribution from a certain family of distributions. In this chapter we introduce the methodology. We describe how a Bayes linear belief structure can be constructed in Section 4.2. The belief adjustment rules for the mean and variance, core to the methodology, are introduced in Section 4.3, together with the possible interpretations that can be given to these belief revisions and other occurrences of these rules in the literature. In Section 4.4 we summarise the interpretative and diagnostic tools available to analyse the specified beliefs and (potential) revisions of these by observations.

4.1 Introduction

> *"The essence of the belief structure construction is to allow us to make collections of belief statements which are much less detailed than those required for the usual Bayesian analysis but which still possess sufficient structure that they may be systematically analysed."*
>
> (Goldstein 1988*a*)

The Bayes Linear (BL) methodology was developed by Goldstein in a series of papers (Goldstein 1981, Goldstein 1986, Goldstein 1988*a*, Goldstein 1988*b*, Goldstein 1991, Goldstein 1994) and has been compiled into a comprehensive book (Goldstein & Wooff 2007). A brief overview over the methodology is given in (Goldstein 1998). Earlier considerations of linear Bayes methods and some of the key results can be found in (Stone 1963) and (Hartigan 1969). The methodology takes expectation rather than probability as the fundamental concept and is based on the following four principles (Goldstein 1994):

Principle 1 Specify only those aspects of their beliefs that assessors are both willing and able to quantify honestly.

Principle 2 Use coherent probabilistic guidelines for revising beliefs.

Principle 3 Base statistical models on judgements about observable quantities.

Principle 4 Use theory to interpret the underlying structure of beliefs.

In the base case, when we wish experts to assess their beliefs about magnitudes of quantities of interest and wish to learn more about these magnitudes by observing other quantities, Goldstein argues that the bare minimum aspects that must be considered are:

1. some quantitative judgements as to the magnitudes of the various quantities,
2. some expression of the degree of confidence in the judgements of magnitude,
3. some expression of the extent to which the prior judgements about the various quantities are interrelated (so that observation on some of the quantities may be used to modify judgements on other quantities).

In the Bayes linear methodology assessments of respectively means, variances and covariances are chosen to quantify these aspects. All three can be derived from first and second order (product) moments, so BL models are thus fully specified by a second order moment specification. The BL methodology can therefore be fully developed from De Finetti's concept of coherent previsions discussed in Section 2.3.1. By working only with (product) moment assessments BL avoids the use of probability distributions and offers a simpler approach to belief analysis and revision than full probabilistic methodology. BL avoids the need for distributional assumptions (unless assumptions are made in the derivation of the (product) moments) and involved posterior distribution calculations.

The methodology is by no means restricted to the assessment and revision of magnitudes only. As beliefs of magnitudes of uncertain quantities are specified, so can beliefs about functions of these uncertain quantities be, for example the square or the cube of the same quantity. By including these functions in the model, beliefs about e.g. variability and asymmetry can be specified and revised as well. To quote Goldstein (1994), the Bayes linear belief specification "may be viewed as reducing the full probabilistic approach to whatever level of detail we feel is both within our ability to specify and adequate to the problem at hand".

The methodology thus reflects the discrete character of quantitative expert assessments and is flexible in the level of detailed specified and reasoned with. Furthermore, the methodology is assumption free as in that it does not require the quantities to have a probability distribution from a certain family of distributions. The only requirements needed are that the second order (product) moments for the quantities in the model are finite, and coherently specified.

Hence, while being similar in spirit to full probabilistic analysis, more complex problems can be modelled with BL with a same amount of time and effort. BL models have a relatively limited level of detail, but analyses and belief revisions are performed directly on and only with the assessments given by the assessors.

The Bayes linear methodology has been applied in the water industry, in the analysis of computer simulators for complex physical systems and various other studies. For an overview of applications of the methodology and references to these, we refer to (Goldstein & Wooff 2007, p.94).

4.2 Bayes Linear Belief Structure

In introducing the elements of a BL model we assume that there is a decision maker (DM) who wishes to inform his decision by the BL model, and an analyst who helps him with building and analysing the model. The DM and the analyst can be the same person. Let $B = \{X_1, \ldots, X_n\}$ be a collection of quantities about the value of which the DM is uncertain but interested for her decision problem. We call B the base of the BL model. The BL model is fully specified by the first and second order (product) moments for the quantities in base B: $E(X_i)$ for $i = 1, \ldots, n$, and $E(X_k X_l)$ for $k, l = 1, \ldots, n$. Note that these moments specify the means $E(X_i)$, variances $Var(X_i) = E(X_i^2) - E(X_i)^2$ of the quantities in the base, as well as the covariances $Cov(X_i, X_j) = E(X_i X_j) - E(X_i)E(X_j)$ between them. Due to the linearity of expectations, this second order (product) moment specification for B also determines the second order (product) moment specification for the collection of all finite linear combinations of elements of B. We will denote this collection with $\langle B \rangle$.

The only requirement for the first and second order moment assessments of the quantities in base B is that these assessments are finite and coherent. Note that this requirement is often not fulfilled for fat tailed distributions like e.g. the Cauchy distribution, which does not even have a finite mean. We recall from Section 2.3.1 that a set of expectations is coherent if these expectations are in the convex hull of the realm of the quantities for which they are assessed. We will refer to a set of quantities B with coherent first and second order moments specification as a coherent belief structure on B, with notation: $[B]$. The requirement of coherency for belief structure $[B]$ entails that the covariance matrix for the quantities in B that can be calculated from the (product) moments in $[B]$ is nonnegative definite (see (Wisse, Bedford & Quigley 2008*a*) for simple demonstration).

4.3 Belief Adjustments

The quantities of a Bayes linear base B are commonly divided into two collections denoted with the vectors $\mathbf{B}$ and $\mathbf{D}$. The quantities in $\mathbf{D}$ are the quantities for which observations (data) will become available, which the decision maker will use to adjust his beliefs about the quantities in $\mathbf{B}$. The adjustment of means, variances of and covariances between the elements of $\mathbf{B}$ is done by linear fitting on the observations $\mathbf{D}$, hence the name Bayes linear methodology.

4.3.1 The Bayes Linear Adjustment Rules

The adjusted expectation of a random quantity X from $\mathbf{B}$, given observation of the quantities in $\mathbf{D}$, notation $E_{\mathbf{D}}(X)$, is the linear combination $E_{\mathbf{D}}(X) = \mathbf{h}\mathbf{D}^*$ which minimises the expected squared error with X, $E\left([X - \mathbf{h}\mathbf{D}^*]^2\right)$, over all $\mathbf{h}$, where $\mathbf{D}^* = (1, \mathbf{D})$, the vector $\mathbf{D}$ supplemented by the unit constant.

The BL adjusted expectation $E_{\mathbf{D}}(X)$ of X given $\mathbf{D} = \mathbf{d}$, resulting from this minimisation, is determined by the prior mean, variance and covariance specifications for X and $\mathbf{D}$, and the observations $\mathbf{d}$:

$$E_{\mathbf{D}}(X) = E(X) + Cov(X, \mathbf{D})Var(\mathbf{D})^{\dagger}(\mathbf{d} - E(\mathbf{D})). \tag{4.1}$$

where the matrix $Var(\mathbf{D})^{\dagger}$ is the Moore-Penrose generalised inverse. When $Var(\mathbf{D})$ is nonsingular, $Var(\mathbf{D})^{\dagger} = Var(\mathbf{D})^{-1}$ is the usual matrix inverse. The adjusted expectation is linear, Equation (4.2), and conglomerable, Equation (4.3):

$$E_{\mathbf{D}}(a_1X_1 + a_2X_2) = a_1E_{\mathbf{D}}(X_1) + a_2E_{\mathbf{D}}(X_2) \tag{4.2}$$

$$E(E_{\mathbf{D}}(X)) = E(X) \tag{4.3}$$

The difference between the quantity X and the adjusted expectation $E_{\mathbf{D}}(X)$ is referred to as the adjusted version of X given $\mathbf{D}$, $\mathbb{A}_{\mathbf{D}}(X)$:

$$\mathbb{A}_{\mathbf{D}}(X) = X - E_{\mathbf{D}}(X). \tag{4.4}$$

The adjusted version $\mathbb{A}_{\mathbf{D}}(X)$ has expectation zero and is uncorrelated with both the data $\mathbf{D}$ and the adjusted expectation $E_{\mathbf{D}}(X)$. The variance of the adjusted version $\mathbb{A}_{\mathbf{D}}(X)$, the expectation of the squared error $(X - E_{\mathbf{D}}(X))^2$, is called the adjusted variance of X given $\mathbf{D}$, notation $Var_{\mathbf{D}}(X)$. The adjusted variance is fully determined by the covariance matrix of X and $\mathbf{D}$:

$$Var_{\mathbf{D}}(X) = Var(X) - Cov(X, \mathbf{D})Var(\mathbf{D})^{\dagger}Cov(\mathbf{D}, X). \tag{4.5}$$

Note that the adjusted variance does not depend on the observations $\mathbf{d}$. The Equations (4.1) and (4.5) form the core of the BL methodology and are referred to as the Bayes linear adjustment rules. Since

$$X = E_{\mathbf{D}}(X) + \mathbb{A}_{\mathbf{D}}(X)$$

and $E_{\mathbf{D}}(X)$ and $\mathbb{A}_{\mathbf{D}}(X)$ are uncorrelated, we can split the variance into the two components

$$Var(X) = Var(E_{\mathbf{D}}(X)) + Var(\mathbb{A}_{\mathbf{D}}(X)). \tag{4.6}$$

The variance of the adjusted expectation $E_{\mathbf{D}}(X)$ is called the variance of X resolved by $\mathbf{D}$, $RVar_{\mathbf{D}}(X)$. From (4.5) and (4.6) we find that

$$RVar_{\mathbf{D}}(X) = Cov(X, \mathbf{D})Var(\mathbf{D})^{\dagger}Cov(\mathbf{D}, X). \tag{4.7}$$

The ratio of the resolved variance and the prior variance, called resolution $R_{\mathbf{D}}(X)$, provides a measure for how informative the observations $\mathbf{D}$ are for X:

$$R_{\mathbf{D}}(X) = \frac{RVar_{\mathbf{D}}(X)}{Var(X)} = 1 - \frac{Var_{\mathbf{D}}(X)}{Var(X)}. \tag{4.8}$$

4.3.2 Interpretations of Belief Adjustments

Goldstein & Wooff (2007) identify three different viewpoints to the Bayes linear approach. We can see the BL analysis as an approximation to full Bayes analysis, requiring less time and with simpler calculations. And, when considering BL as the more fundamental approach, as a generalisation of full Bayes analysis, where the constraint of the requirement of a full probabilistic prior specification is relaxed. Thirdly, BL can be seen as complementary to full Bayes analysis, offering a variety of interpretative and diagnostic tools to analyse the prior beliefs and the belief adjustments. Section 4.4 will treat the interpretative and diagnostic machinery of the BL methodology. We will now proceed to discuss the interpretations of the BL adjusted mean and variance for each of these three interpretations.

The BL adjusted mean can be seen as an approximation to the full Bayes conditional mean. In this view, the adjusted variance is a simple, easily computable upper bound on full Bayes preposterior risk, under quadratic loss, for any full prior specification consistent with the given mean and variance specifications (Goldstein & Wooff 2007).

When the quantities in $\mathbf{D}$ are indicator functions that together form an event partition, the BL adjusted mean is equal to the full Bayes conditional mean. Starting from this equality, Goldstein & Wooff (2007) argue that the adjusted expectation in the general case can also be viewed as a natural generalisation of conditional expectation, without the restriction that one must only perform the conditioning on indicator functions of a partition. The adjusted variance is then interpreted as a prior variance, but applied to the residual variation one would have for the quantity when the variation accounted for by the observations $\mathbf{D}$ is deducted from the prior variance.

Thirdly, the adjusted expectation can be seen as an estimator of the value of X, complementary to certain standard estimators in multivariate analysis. The adjusted variance is then simply the mean squared error of this estimator for the conditional mean.

4.3.3 Other Occurrences of the Adjustment Rules

The BL adjustment rules are also found in the literature as the linear least squares (LLS) solution to an overdetermined system of linear equations. The Extended Gauss-Markov Theorem (see e.g. (Whittle 1992)) states that when random variables are jointly Normally distributed, the LLS estimate $\widehat{\mathbf{X}}$ of a vector $\mathbf{X}$ for data vector $\mathbf{Y}$ coincides with the conditional mean of $\mathbf{X}$ given $\mathbf{Y}$, $E(\mathbf{X}|\mathbf{Y})$. Furthermore, the variance-covariance matrix of the estimation error, $Cov(\mathbf{X}-\widehat{\mathbf{X}})$, can then be identified with the conditional variance-covariance matrix $Cov(\mathbf{X}|\mathbf{Y})$. So when the BL adjustments are considered approximations to full probabilistic updating, the adjustment rules are exact when all variables are jointly Normally distributed.

Ericson (1969) has shown for variable X and data $\mathbf{D}$ and with finite prior variance that, when the posterior mean of variable X is linear in the data (i.e. the conditioning set), this posterior mean can be expressed in means and (co)variances of the distributions of X and $\mathbf{D}$.

The expression for the posterior mean is then equal to the BL adjusted mean. Ericson's result holds for example for conjugate prior models from the linear exponential family and is also used in credibility theory (see e.g. (Klugman, Panjer & Willmot 1998)). The Pearson system of bivariate distributions developed by Van Uven (Van Uven 1947*a*, Van Uven 1947*b*, Van Uven 1948*a*, Van Uven 1948*b*) has the property that the mean of one variable conditional on the other is linear in the condition. So for this family the approximation of the conditional mean with the BL adjusted mean is exact.

4.4 Analysis of Belief Structure and Diagnostic Tools

In this section different properties of the BL belief structure and the BL belief adjustment will be discussed. More information on the concepts introduced in this section can be found in (Goldstein & Wooff 2007). Via the so called canonical directions and corresponding resolutions we can investigate how much we expect to learn in advance from a specific data set. So when having the choice between different data sets, we can choose the one we expect to reduce our uncertainty the most. When observations are available, the *discrepancy* tells us how concordant (or discordant) with our prior expectations the observations are. The *bearing* of the adjustment gives information about the magnitude and the direction of the adjustment together, whilst the *size* measures the normalised magnitude of the adjustment. The measures help to determine whether the observations are so discordant with prior beliefs that one could reconsider these prior beliefs.

4.4.1 Canonical Resolutions

Whether we interpret the adjusted variance as an upper bound to preposterior risk, residual variation or as the mean squared error of the adjusted mean, we would like it to be as small as possible. That is, when we have the choice of observing either data $\mathbf{D_1}$ or $\mathbf{D_2}$, we would choose the data that leads to the smallest adjusted variance. We have already noted that the adjusted variance only depends on prior variances and covariances, and not on the value of the observations. This means that for a fully specified belief structure we can determine how much we expect to learn from the data before actually obtaining the observations.

The resolution $R_{\mathbf{D}}(X)$, Equation (4.8), provides a scale free measure for the extent in which the variance of a single quantity will be reduced by the observation of data $\mathbf{D}$. But we can also evaluate how much $\mathbf{D}$ tells us about all our quantities in $\mathbf{B}$ altogether. This is done by analysis of the canonical structure of the belief structure $[B]$.

Canonical direction: the linear combination $Z_j \in \langle B \rangle$ is called the jth canonical direction for the adjustment of $\mathbf{B}$ by $\mathbf{D}$ if it maximises the resolution $R_{\mathbf{D}}(Z)$ over all linear combinations $Z \in \langle B \rangle$ that have non-zero prior variance and that are uncorrelated a priori with $Z_1, \ldots, Z_{j-1}$. Each Z_j is scaled to have prior expectation zero and prior variance one.

Canonical resolution: the resolutions $R_{\mathbf{D}}(Z_i)$ of the canonical directions Z_i are called canonical resolutions, and are notated in short with λ_i.

The canonical directions and resolutions can be calculated from the resolution transform matrix $\mathbb{T}_{\mathbf{B}:\mathbf{D}}$

$$\begin{aligned} \mathbb{T}_{\mathbf{B}:\mathbf{D}} &= Var(\mathbf{B})^{\dagger} RVar_{\mathbf{D}}(\mathbf{B}) \\ &= Var(\mathbf{B})^{\dagger} Cov(\mathbf{B}, \mathbf{D}) Var(\mathbf{D})^{\dagger} Cov(\mathbf{D}, \mathbf{B}). \end{aligned} \tag{4.9}$$

If the normed eigenvectors of $\mathbb{T}_{\mathbf{B}:\mathbf{D}}$ are ordered from high to low, i.e. $1 \geq \lambda_1 \geq \lambda_2 \geq \ldots, \lambda_r \geq 0$, then λ_i is the ith canonical resolution and its corresponding eigenvector is the ith canonical direction Z_i.

We thus expect to learn most about the quantities in $\mathbf{B}$ that have strong correlations with the canonical directions with high resolutions.

4.4.2 Discrepancy, Size and Bearing

The discrepancy is a simple measure that can be used to assess the extent to which observations are in accordance with prior specifications. The discrepancy of a single observation d, $Dis(d)$, is defined as the square of the standardised observation:

$$Dis(d) = \frac{[d - E(D)]^2}{Var(D)}. \tag{4.10}$$

A large discrepancy can be an indication that either the prior expectation has been misspecified, the prior variance has been underestimated or the observation has been misrecorded. A very small discrepancy might indicate an overestimated prior variance. The discrepancy of a vector of observations $\mathbf{d}$ is defined as

$$Dis(\mathbf{d}) = [\mathbf{d} - E(\mathbf{D})]^T Var(\mathbf{D})^{\dagger} [\mathbf{d} - E(\mathbf{D})]. \tag{4.11}$$

This vector discrepancy is equal to the maximum discrepancy found over all linear combinations of elements of $\mathbf{d}$ that have positive variance.

The size of and the bearing for the adjustment help to understand quantitatively how prior beliefs are changed by the adjustment. The size of the adjustment of X by $\mathbf{D} = \mathbf{d}$, $Size_{\mathbf{d}}(X)$, is calculated as the normalised squared difference between the prior and adjusted expectation

$$Size_{\mathbf{d}}(X) = \frac{[E_{\mathbf{d}}(X) - E(X)]^2}{Var(X)}. \tag{4.12}$$

A large size of an adjustment indicates that the prior variance might be overly small in the light of the adjustment, or that the observed change might be too large in the light of the prior variability of X. The size of a vector $\mathbf{B}$ of quantities by $\mathbf{D} = \mathbf{d}$, $Size_{\mathbf{d}}(\mathbf{B})$, is defined to be the maximum size $Size_{\mathbf{d}}(F^+)$ found over all linear combinations $F^+ \in \langle B \rangle$ that have positive variance, and is calculated as

$$Size_{\mathbf{d}}(\mathbf{B}) = [E_{\mathbf{d}}(\mathbf{B}) - E(\mathbf{B})]^T Var(\mathbf{D})^{\dagger} [E_{\mathbf{d}}(\mathbf{B}) - E(\mathbf{B})]. \tag{4.13}$$

A property of the belief adjustment that expresses both the direction and magnitude of the change between prior and adjusted beliefs is the bearing. The bearing of the adjustment of $\mathbf{B}$ by $\mathbf{D} = \mathbf{d}$ is

$$\mathbb{Z}_{\mathbf{d}}(\mathbf{B}) = [E_{\mathbf{d}}(\mathbf{B}) - E(\mathbf{B})]^T Var(\mathbf{B})^{\dagger} [\mathbf{B} - E(\mathbf{B})]. \tag{4.14}$$

The bearing may be interpreted as the linear (normalised) likelihood (Goldstein & Wooff 2007). The bearing $\mathbb{Z}_{\mathbf{d}}(\mathbf{B})$ has the nice property that the magnitude of the adjustment of any linear combination $F \in \langle B \rangle$ can be derived as the covariance with the bearing

$$Cov(F, \mathbb{Z}_{\mathbf{d}}(\mathbf{B})) = E_{\mathbf{d}}(F) - E(F).$$

Any quantity from $\mathbf{B}$ that is uncorrelated with the bearing of $\mathbf{B}$ will thus have its prior expectation unchanged by the adjustment by $\mathbf{d}$. The bearing of the adjustment is closely related to the size of the adjustment; the latter is also calculated as the variance of the former:

$$Size_{\mathbf{d}}(\mathbf{B}) = Var(\mathbb{Z}_{\mathbf{d}}(\mathbf{B})).$$

The prior expectation of this size of the adjustment is the trace of the resolution transform matrix, $tr(\mathbb{T}_{\mathbf{B:D}})$, which is equal to the sum of the canonical resolutions (eigenvalues of $\mathbb{T}_{\mathbf{B:D}}$). So if the size of the adjustment is much larger than the sum of the canonical resolutions, then the adjusted beliefs are surprisingly discordant with our prior specifications. A much smaller size might indicate overstated prior uncertainty.

4.5 Summary

In this chapter we have introduced the Bayes linear methodology. We have discussed the BL adjustment of means and variances, core to the methodology, and interpretations of these adjustments. We treated the interpretative and diagnostic machinery available to analyse the second order (product) moments specified, and the adjustments of these by observations. In Section 3.5 we discussed how the second order (product) moments necessary to specify a Bayes linear belief structure can be derived from expert assessments. In Section 6.3 we will continue this discussion for the assessment of higher order (product) moments, and evaluate the bivariate generalisation of the extended Pearson-Tukey method for this purpose. In the next chapter we will consider the Bayes linear adjustment rules as approximation to full probabilistic updating, and evaluate the accuracy of this approximation. We are not aware of any research that has been performed up to date on this topic. In Chapter 6 we will evaluate the effect the inclusion of higher order (product) moment information has on the accuracy of this approximation, both when the (product) moments are exact and when the moments are derived using the bivariate Pearson-Tukey method.

Chapter 5

Bayes Linear Approximation

In this chapter we consider the Bayes linear approach as an approximation to full probabilistic updating. We will investigate how accurate the Bayes linear adjustment rules are when variables are not joint Normally distributed. We select a set of bivariate distribution families in Section 5.2 and evaluate the difference between the Bayes linear adjusted mean and variance and the conditional mean and variance. First analytically in Section 5.3, and secondly using Monte Carlo sampling in Section 5.4. The findings are summarised in the final section of this chapter.

5.1 Introduction

The Bayes linear adjusted mean, $E_Y(X)$, and variance, $Var_Y(X)$, of a single quantity X that is adjusted by a single quantity Y are calculated as:

$$\begin{aligned} E_Y(X) &= E(X) + \frac{Cov(X,Y)}{Var(Y)} \left(y - E(Y) \right), & (5.1) \\ Var_Y(X) &= Var(X) - \frac{Cov(X,Y)^2}{Var(Y)}. & (5.2) \end{aligned}$$

We shall simply write 'adjusted mean' and 'adjusted variance' when referring to these adjustment rules. In Section 4.3.2 we have discussed three different ways in which these adjustment rules can be interpreted. In this chapter we take the approximation interpretation, viewing the BL adjustment of the mean as an approximation to the full probabilistic conditional mean $E(X|Y)$. It is well known that expectation minimises squared error. In Chapter 4 we discussed that the Bayes linear adjustment of the mean of X by Y, $E_Y(X)$, is defined to be the linear function of Y that minimises squared error with X. Therefore, for any conditional expectation that is linear in the condition the conditional expectation will be identical to the Bayes linear adjustment rule for the mean, as was noted by Ericson in (Ericson 1969).

The adjusted variance is defined to be the expectation of the squared error of the adjusted mean with X, $E((X - E_Y(X))^2)$. The adjusted variance can thus be interpreted as an approximation to the expectation of the conditional variance, $E(Var(X|Y))$ (note in (5.2) that the adjusted variance also does not depend on Y):

$$\begin{aligned} E(Var(X|Y)) &= E\left(E\left[(X - E(X|Y))^2|Y\right]\right) = E\left[(X - E(X|Y))^2\right] \\ &\approx E\left[(X - E_Y(X))^2\right] = Var_Y(X). \end{aligned} \tag{5.3}$$

In case the conditional mean is linear in the condition and the adjusted mean thus is exact, the approximation (5.3) is exact as well, and the adjusted variance equals the expected conditional variance. Among continuous bivariate distributions with a linear conditional mean are the bivariate Normal, Filon-Isserlis beta, Kibble's Gamma, Cheriyan's Gamma, McKay's Gamma, a bivariate Pareto, Gosh's F, Rhodes, Pearson's bivariate Student and Pearson's Type II distribution (see e.g. (Mardia 1970) and (Balakrishnan & Lai 2009)), and the linear exponential family in general. Two examples of bivariate distribution with nonlinear conditional mean are Gumbel's bivariate exponential and the Farlie-Gumbel-Morgenstern bivariate Gamma distribution. Bivariate (and multivariate) Normally distributed variables do not only have a linear conditional mean, but for these variables the conditional variance is also constant. Hence, for joint Normally distributed variables the adjusted variance is also identical to the conditional variance.

In general conditional variances will not be constant however, and the adjusted variance might be a poor approximation to it. In this chapter we investigate the errors made using the adjusted variance as an approximation to the conditional variance for a set of known bivariate distributions. First analytically in Section 5.3 and by means of Monte Carlo sampling in Section 5.4. The set of bivariate distributions evaluated in this chapter is introduced in the next section.

5.2 Selection of the Distributions

The distributions for which the adjusted variance will be compared with the conditional variance have been selected to cover as wide a variety of behaviours of the variables as possible. To be able to perform the evaluations in Section 5.3 we need an analytical expression of:

- the conditional mean and variance,

- the means and (co)variances.

To be able to calculate the difference measures for evaluation of the variance approximations described in Section 5.4, and analyse these against properties of the marginal and joint behaviour of both variables we further need to be able to calculate with sufficient precision:

- the marginal 5%− and 95%−quantile,

- the joint moments of up to the fourth order.

The marginal quantiles are needed for the calculation of the difference measures for the variances described in Section 5.4.1. The additional moments are used to calculate the marginal skewness and kurtosis, as well as the higher order correlations $Corr(X^i, Y^j)$ with $i, j = 1, 2$, against which the differences will be analysed.

First we searched the literature for systems of bivariate distributions covering a broad spectrum of distributions meeting the above requirements. The only system we found that meets the above requirements is Van Uven' bivariate extension of the Pearson system of distributions (Van Uven 1947*a*, Van Uven 1947*b*, Van Uven 1948*a*, Van Uven 1948*b*). For this system the joint, marginal and conditional distribution, the conditional mean and variance and the unconditional (product) moments can be expressed in the parameters of the two differential equations defining the system. However, not all parameter combinations for the 17 parameters of this system lead to proper density functions. Since we were not able to constraint the parameter space such that each point in this space would correspond to a proper bivariate density function, we could not use this system.

Instead we searched the literature for known bivariate distributions that meet the requirements stated. This resulted in the following four distributions for which we can perform the desired calculations: Filon-Isserlis' bivariate Beta, Kibble's bivariate Gamma, Cheriyan's bivariate Gamma and a bivariate F distribution (see (Mardia 1970) and (Balakrishnan & Lai 2009) for details). With these distributions we can investigate bivariate distributions with marginal Type I (Beta), Type III (Gamma) an Type VI (F) distributions from the Pearson system of univariate distributions. Each of the four distributions has a conditional mean that is linear in the condition. For these distribution the Bayes linear adjusted mean thus is equal to the conditional mean, as is shown in Appendix A.1. So the attention in the remainder of this chapter will be turned to the accuracy of the adjusted variance. In the remainder of this section each distribution will be briefly introduced, together with the sampling strategy applied for the distribution in the evaluation in Section 5.4.

5.2.1 Filon-Isserlis' Bivariate Beta

The Filon-Isserlis surface is a bivariate Beta distribution, defined by the following probability density function (pdf):

$$h_{FI}(x,y) = \frac{\Gamma(p_1+p_2+p_3)}{\Gamma(p_1)\Gamma(p_2)\Gamma(p_3)} x^{p_1-1} y^{p_2-1} (1-x-y)^{p_3-1}, \qquad x, y \geq 0, x+y \leqslant 1,$$

and $p_1, p_2, p_3 > 0$. X has as marginal distribution a Beta distribution with shape parameters p_1 and p_2+p_3. Y is distributed as a Beta distribution with parameters p_2 and p_1+p_3. Both variables have a lower bound of 0, an upper bound depending on the value of the other variable with a maximum of 1. Each variable can be symmetric, positively skewed or negatively skewed. The kurtosis of the marginals can be smaller, equal to or larger than 3.

The joint moments of this distribution can be calculated as

$$E(X^r Y^s) = \frac{\Gamma(p_1+r)\Gamma(p_2+s)\Gamma(p_1+p_2+p_3)}{\Gamma(p_1)\Gamma(p_2)\Gamma(p_1+p_2+p_3+r+s)}.$$

The conditional mean and variance of X are given by

$$\begin{aligned} E(X|Y=y) &= \frac{p_1}{p_1+p_3}(1-y), && (5.4) \\ Var(X|Y=y) &= \frac{p_1 p_3}{(p_1+p_3)^2(1+p_1+p_3)}(1-y)^2. && (5.5) \end{aligned}$$

The 5%$-$ and 95%$-$quantile of Y can be derived from the inverse cumulative distribution function of the univariate Beta distribution with shape parameters p_2 and p_1+p_3.

Sampling from Filon-Isserlis' Beta. A Filon-Isserlis distribution is defined by the parameters p_1, p_2 and p_3. For the current analysis p_1, p_2 and p_3 are independently sampled from a Normal distribution with zero mean and a standard deviation of 10, of which the absolute value is taken. The marginal Beta distributions of the Filon-Isserlis bivariate Beta distribution are Pearson Type I distributions. In Figure 5.1a the cases from a sample of $10,000$ Filon-Isserlis distributions are displayed that have a squared skewness and kurtosis smaller than 15, where the sample was constructed as described in this paragraph.

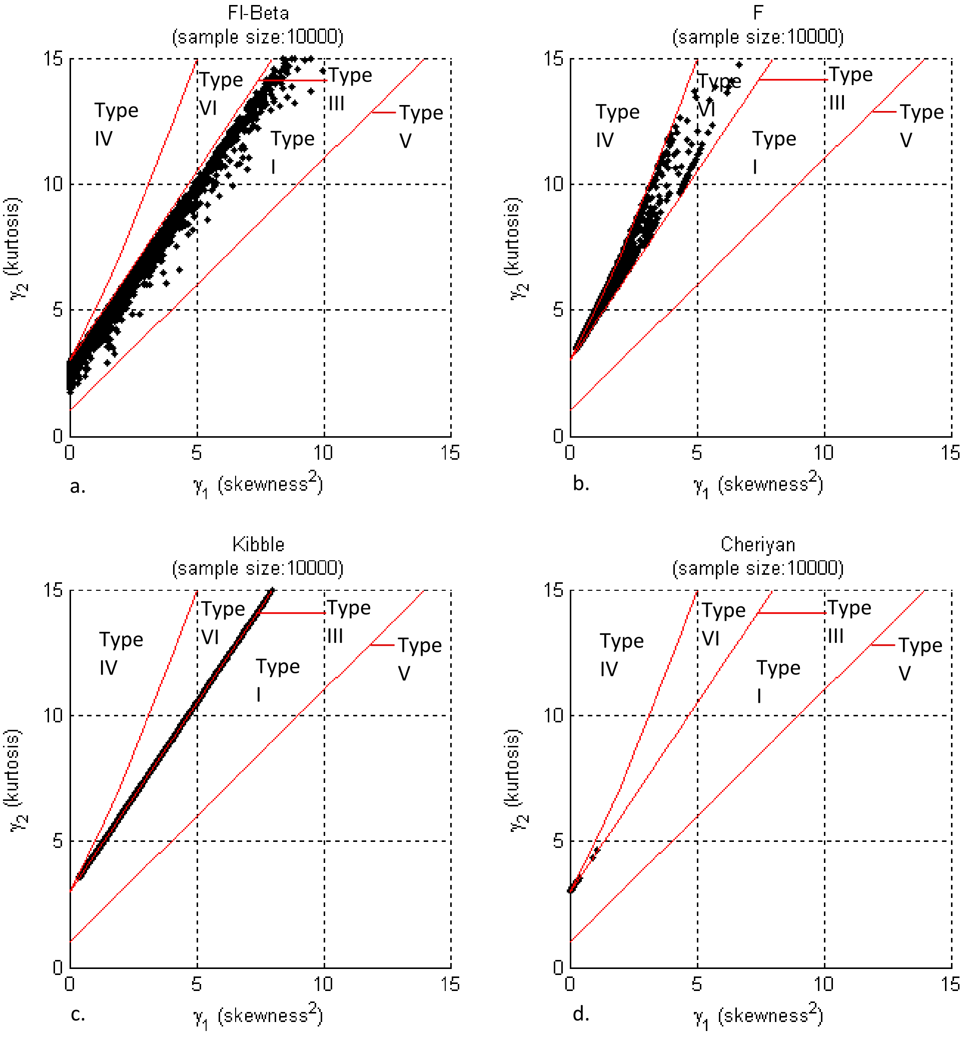

Figure 5.1: Diagram of the Pearson system of univariate distributions with for the a. Filon-Isserlis, b. F, c. Kibble and d. Cheriyan distribution the cases displayed of a sample of 10,000 that have a squared skewness and kurtosis smaller than 15.

5.2.2 Bivariate F

The bivariate F distribution used in this study is also known as the bivariate inverted Beta or the bivariate inverted Dirichlet distribution. Its probability density function $h_F(x, y)$ is

$$h_F(x,y) = Kx^{(\nu_1-2)/2}y^{(\nu_2-2)/2}\left(1+\frac{\nu_1 x+\nu_2 y}{\nu_0}\right)^{-(\nu_0+\nu_1+\nu_2)/2}, \qquad x,y \geq 0,$$

where the ν's are positive and the constant K is given by

$$K = \Gamma\left(\frac{\nu_0+\nu_1+\nu_2}{2}\right)\nu_0^{-(\nu_0+\nu_1+\nu_2)/2}\frac{\nu_0^{\nu_0/2}\nu_1^{\nu_1/2}\nu_2^{\nu_2/2}}{\Gamma(\nu_0/2)\Gamma(\nu_1/2)\Gamma(\nu_2/2)}.$$

Both X and Y have an F-distribution as marginal distribution, X with ν_1 and ν_0 degrees of freedom, Y with ν_2 and ν_0. So both variables have a lower bound of 0, are positively skewed and have a kurtosis bigger than 3 (the kurtosis of a Normally distributed variable).

The joint moments of this distribution are calculated as

$$E(X^rY^s) = \frac{\Gamma(\frac{1}{2}\nu_0 - r - s)\Gamma(\frac{1}{2}\nu_1 + r)\Gamma(\frac{1}{2}\nu_2 + s)}{\Gamma(\nu_0/2)\Gamma(\nu_1/2)\Gamma(\nu_2/2)(\nu_1/\nu_0)^r(\nu_2/\nu_0)^s}.$$

The conditional mean and variance of X are given by

$$\begin{aligned} E(X|Y=y) &= \frac{\nu_0 + \nu_2 y}{\nu_0 + \nu_2 - 2} && (5.6) \\ Var(X|Y=y) &= \frac{2(\nu_0 + \nu_1 + \nu_2 - 2)}{\nu_1(\nu_0 + \nu_2 - 2)^2(\nu_0 + \nu_2 - 4)}(\nu_0 + \nu_2 y)^2. && (5.7) \end{aligned}$$

The 5%− and 95%−quantile of Y can be derived from the inverse cumulative distribution function of the univariate F-distribution with ν_2 and ν_0 degrees of freedom.

Sampling from F. The univariate F distribution is defined by the parameters ν_0, ν_1 and ν_2. For the current analysis ν_1 and ν_2 are independently sampled from a Normal distribution with zero mean and a standard deviation of 10, of which the absolute value is taken. To ensure finiteness of the higher moment used in the analysis ν_0 is sampled in the same way as ν_1 and ν_2, then a value of 12 is added to this. The marginal distributions of this bivariate F distribution are Pearson Type VI distributions. See Figure 5.1b for the cases out of a sample of $10,000$ that have both the squared skewness and kurtosis smaller than 15, where the sample is constructed as described in this paragraph.

5.2.3 Kibble's Bivariate Gamma

Kibble's bivariate Gamma distribution is defined by the following pdf:

$$h_{Kibble}(x,y) = f_\alpha(x)f_\alpha(y)I_{\alpha-1}\left(\frac{2\sqrt{\rho xy}}{1-\rho}\right)\frac{\Gamma(\alpha)}{1-\rho}(\rho xy)^{(\alpha-1)/2}e^{\frac{-\rho(x+y)}{1-\rho}}, \qquad x, y \geq 0, 0 \leqslant \rho < 1,$$

where $f_\alpha(t) = \frac{1}{\Gamma(\alpha)}e^{-t}t^{\alpha-1}$ and $I_\alpha(\cdot)$ is the modified Bessel function of the first kind and order ν. Both X and Y have as marginal distribution a Gamma distribution with the scale parameter equal to 1 and shape parameter α. So both variables have a lower bound of 0, are positively skewed and have a kurtosis bigger than 3 (the kurtosis of a Normally distributed variable).

The joint moments of this distribution can be derived from the moment generating function

$$M_{Kibble}(s,t) = [(1-s)(1-t) - \rho st]^{-\alpha}.$$

The conditional mean $E(X|Y=y)$ and variance $Var(X|Y=y)$ of X are given by

$$\begin{aligned} E(X|Y=y) &= (1-\rho)\alpha + \rho y, && (5.8) \\ Var(X|Y=y) &= (1-\rho)^2\alpha + 2\rho(1-\rho)y. && (5.9) \end{aligned}$$

The $5\%-$ and $95\%-$quantile of Y can be derived from the inverse cumulative distribution function of the univariate Gamma distribution with scale parameter 1 and shape parameter α.

Sampling from Kibble's Gamma. To sample a member of the family of Kibble's bivariate Gamma distributions it is sufficient to sample a set of allowable values for the two parameters that fully determine the distribution. Thus sample a $\alpha > 0$ and a ρ from the interval $[0,1)$. For the current analysis α is sampled uniformly between 0 and 10, and, independently, ρ uniformly between 0 and 1. Univariate Gamma distributions are Pearson Type III distributions. In Figure 5.1c the cases from a sample of $10,000$ Kibble distributions are displayed that have a squared skewness and kurtosis smaller than 15, where the sample was constructed as described in this paragraph.

5.2.4 Cheriyan's Bivariate Gamma

The Cheriyan distribution is also a bivariate Gamma distribution, the pdf $h_{Cheriyan}(x,y)$ is

$$h_{Cheriyan}(x,y) = \frac{e^{-(x+y)}}{\Gamma(\theta_1)\Gamma(\theta_2)\Gamma(\theta_3)} \int_0^{min(x,y)} (x-z)^{\theta_1}(y-z)^{\theta_2-1} z^{\theta_3-1} e^z dz, \qquad x,y \geq 0,$$

with $\theta_1, \theta_2, \theta_3 > 0$.

So again both X and Y have as marginal distribution a Gamma distribution with a scale parameter of 1. The shape parameter of X's marginal distribution is $\theta_1+\theta_3$, the shape parameter for Y is $\theta_2+\theta_3$. Both variables have a lower bound of 0, are positively skewed and have a kurtosis bigger than 3.

The joint moments of this distribution can be derived from the moment generating function

$$M_{Cheriyan}(s,t) = (1-s)^{-\theta_1}(1-t)^{-\theta_2}(1-s-t)^{-\theta_3}.$$

The conditional mean and variance of X are given by

$$
\begin{aligned}
E(X|Y=y) &= \theta_1 + \frac{\theta_3}{\theta_2+\theta_3}y, && (5.10)\\
Var(X|Y=y) &= \theta_1 + \frac{\theta_2\theta_3}{(\theta_2+\theta_3)^2(1+\theta_2+\theta_3)}y^2. && (5.11)
\end{aligned}
$$

The 5%$-$ and 95%$-$quantile of Y can be derived from the inverse cumulative distribution function of the univariate Gamma distribution with scale parameter 1 and shape parameter $\theta_2+\theta_3$.

Sampling from Cheriyan's Gamma. To sample a member of the family of Cheriyan's bivariate Gamma distributions it is sufficient to sample a set of allowable values for the three parameters that fully determine the distribution. For the current analysis θ_1, θ_2 and θ_3 are independently sampled from a Normal distribution with zero mean and a standard deviation of 500, of which the absolute value is taken. A set of $10,000$ Cheriyan distributions sampled as here described is displayed in Figure 5.1d.

5.3 Analytical Evaluation

In this section we will evaluate the analytical expressions of the difference d_{var} between the regular adjusted variance and the conditional variance:

$$d_{var} = Var_Y(X) - Var(X|Y),$$

for each of the distributions described in the previous section. The derivation of the expressions evaluated in this section are given in Appendix A.2.

5.3.1 Filon-Isserlis' Bivariate Beta

For the Filon-Isserlis Beta distribution the difference between the regular adjusted and the conditional variance is:

$$
\begin{aligned}
d_{var,Filon-Isserlis}(y) &= \frac{p_1p_2p_3(1+2p_1+p_2+2p_3)}{(p_1+p_3)^2(1+p_1+p_3)(p_1+p_2+p_3)(1+p_1+p_2+p_3)}\\
&\quad + \frac{p_1p_3}{(p_1+p_3)^2(1+p_1+p_3)}(2y-y^2). \qquad (5.12)
\end{aligned}
$$

On the whole domain of observation Y the adjusted variance and the conditional variance are different. The difference is the lowest at the lower bound $y=0$, where it is equal to the constant from (5.12), and highest at the upper bound $y = 1$. If $y = 1$, X needs to be 0 since $X, Y > 0$ and $X + Y \leqslant 1$. So if $y = 1$ there is no uncertainty left about X, the conditional variance is 0 and the difference between the adjusted and the conditional variance is equal to the adjusted variance.

In Figure 5.2 the relative difference between the adjusted and conditional variance is displayed against the cumulative probability of the condition, for 5 F-I Beta distributions. For the case that all three parameters equal 6, we see that the relative difference is less than 50% only between the 5%− and 80%−quantile of the condition. When parameter p_1 is increased to 36 the relative difference decreases and is smaller than 50% on the whole domain of the condition. The same results are found when only parameter p_3 is increased to 36, which we might have anticipated since both parameters are interchangeable in $d_{var,Filon-Isserlis}$.

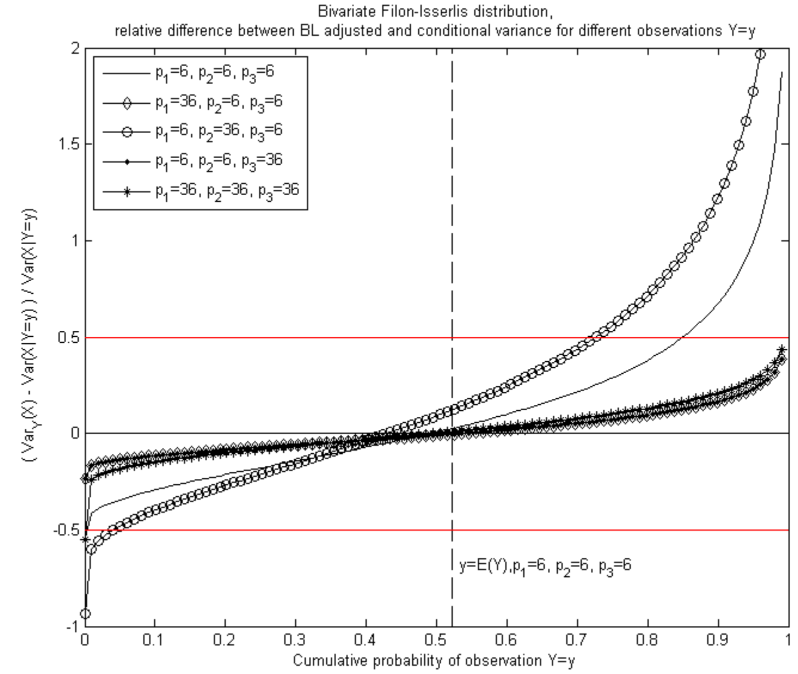

Figure 5.2: Relative difference between the adjusted and conditional variance over the cumulative distribution of the condition, for 5 F-I Beta distributions.

For parameter p_2 increased to 36 the opposite occurs, the relative difference increases on almost the whole domain of the condition. Increasing all three parameters to 36 resembles the effect of changing either p_1 or p_3 alone to 36. Finally, note that the adjusted and conditional variance are in general not equal when the condition is equal to its expectation (dashed line).

5.3.2 Bivariate F

The difference d_{var} between the adjusted and the conditional variance for the bivariate F distribution is:

$$\begin{aligned} d_{var,F}(y) &= \frac{2\nu_0^2[(\nu_0+\nu_1-2)(\nu_0+\nu_2-2)-\nu_1\nu_2]}{\nu_1(\nu_0-2)^2(\nu_0-4)(\nu_0+\nu_2-2)} \\ &\quad + \frac{2\nu_0^2(\nu_0+\nu_1+\nu_2-2)}{\nu_1(\nu_0+\nu_2-2)^2(\nu_0+\nu_2-4)} \\ &\quad + \frac{2(\nu_0+\nu_1+\nu_2-2)}{\nu_1(\nu_0+\nu_2-2)^2(\nu_0+\nu_2-4)}(-2\nu_0\nu_2 y-\nu_2^2 y^2). \end{aligned} \tag{5.13}$$

The relative difference between the adjusted and conditional variance gets smaller for the F distribution when all three parameters increase, see Figure 5.3. The difference does not depend (strongly) on parameter ν_1, increases with increasing ν_2 and decreases with increasing ν_0.

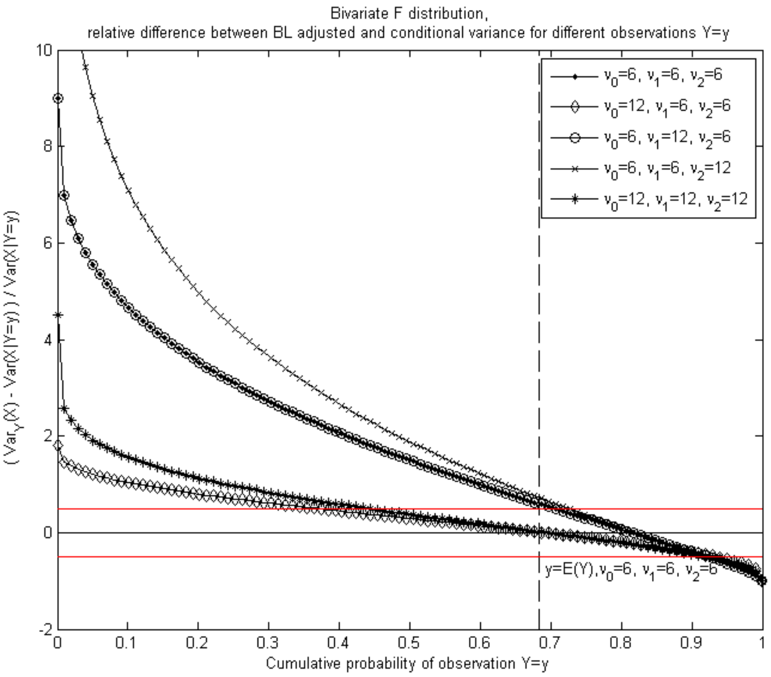

Figure 5.3: Relative difference between the adjusted and conditional variance over the cumulative distribution of the condition, for 5 F distributions.

5.3.3 Kibble's Bivariate Gamma

The difference d_{var} between the adjusted and the conditional variance for Kibble's bivariate Gamma distribution is:

$$d_{var,Kibble}(y) = -2\rho(1-\rho)(y - \frac{1}{2}E(Y)). \tag{5.14}$$

So for Kibble's Gamma the adjusted and the conditional variance are equal when observation y is equal to its prior mean $E(Y)$ and/or when the correlation ρ is zero (a correlation of 1 is not allowed for this distribution, see Section 5.2.3). For observations y smaller than the prior mean the difference is positive, for larger y the difference is negative. In Figure 5.4 we see that the difference increases with both an increasing correlation and α.

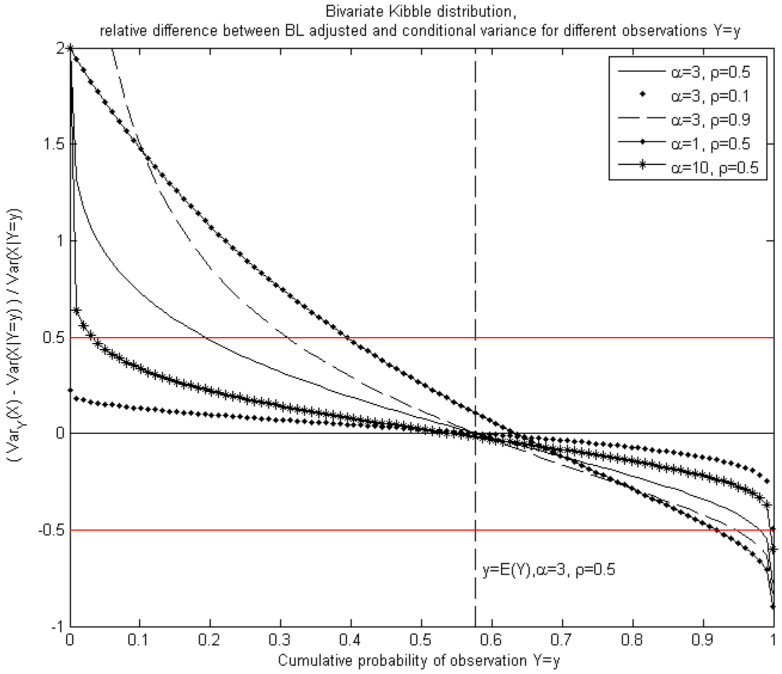

Figure 5.4: Relative difference between the adjusted and conditional variance over the cumulative distribution of the condition, for 5 Kibble distributions.

5.3.4 Cheriyan's Bivariate Gamma

The difference for Cheriyan's Gamma, $d_{var,Cheriyan}$, is:

$$d_{var,Cheriyan}(y) = \frac{\theta_2\theta_3}{\theta_2+\theta_3} - \frac{\theta_2\theta_3}{(\theta_2+\theta_3)^2(1+\theta_2+\theta_3)}y^2. \tag{5.15}$$

This difference is zero when $y^2 = (\theta_2+\theta_3)(1+\theta_2+\theta_3) = E(Y)(1+E(Y))$, since $E(Y) = \theta_2+\theta_3$. So for $y = \sqrt{E(Y)^2 + E(Y)}$ the BL adjusted variance is equal to the conditional variance. For observations $y < \sqrt{E(Y)^2 + E(Y)}$ the difference is positive, when y is larger the difference is negative. This negative difference for $y > \sqrt{E(Y)^2 + E(Y)}$ is not bounded. The relative difference between the adjusted and conditional variance decreases with an increasing θ_1 (Figure 5.5); the absolute difference is unaffected by θ_1, but the conditional variance itself increases with θ_1. When increasing either of the interchangeable θ_2 and θ_3 the relative difference becomes larger at the bounds of the domain of the condition, but will not necessarily become strictly larger or smaller in the center of the domain. When all three parameters are increased a mixed effect results.

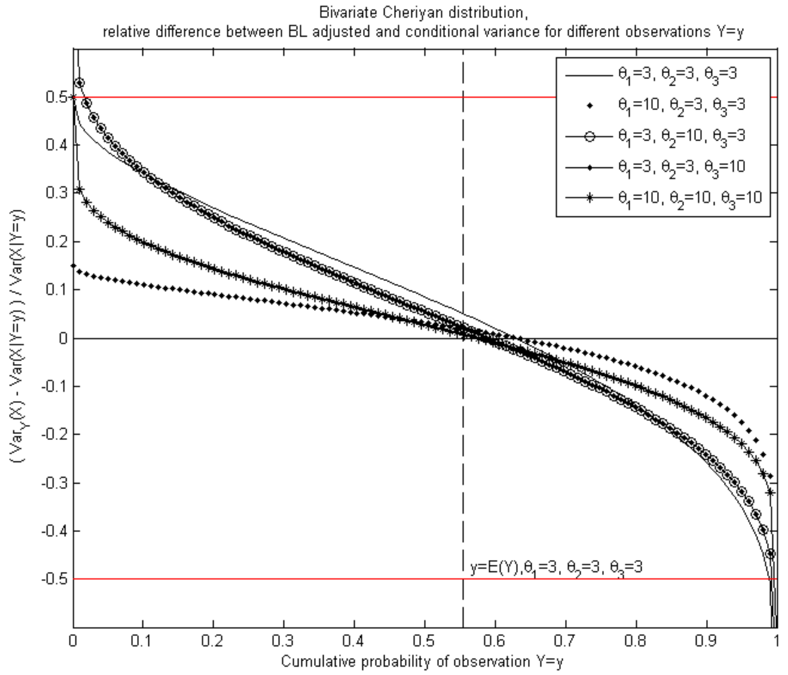

Figure 5.5: Relative difference between the adjusted and conditional variance over the cumulative distribution of the condition, for 5 Cheriyan distributions.

We now have an idea of the difference d_{var} between the adjusted variance and the conditional variance for the four bivariate distributions under consideration. Yet, the analytical expressions of this difference are in the parameters of each of the distributions. In the next section large samples from these four bivariate distribution families are taken to explore if there are common properties of a bivariate distribution via which we can learn about the accuracy of the adjusted variance as an approximation to the conditional variance. We will end this section with one property that might be an indication for the size of d_{var}: the ratio of the minimum and the maximum value of the conditional variance on the $5\% - 95\%$ interquantile range of the condition. The more constant the conditional variance is over the condition, the better the constant adjusted variance can in principle approximate it. In Table 5.1 the mean, standard deviation, minimum and maximum value of this ratio in a sample of $10,000$ F-I Beta, F, Kibble and Cheriyan distribution is given.

Table 5.1: The mean, standard deviation, minimum and maximum value of ratio of the minimum and the maximum value of the conditional variance on the $5\% - 95\%$ interquantile range of the condition. For a sample of $10,000$ F-I Beta, F, Kibble and Cheriyan distribution, in percentages.

	mean	st.dev.	min	max
F-I Beta	38.70	21.88	0.00	100.00
F	51.27	17.46	11.14	95.44
Kibble	41.95	23.14	0.00	99.96
Cheriyan	91.78	5.91	54.99	100.00

For the Cheriyan distribution the difference between the minimum and the maximum value of the conditional variance is by far the smallest on the $5\% - 95\%$ interquantile range of the condition. The minimum value is on average 92% of the maximum value. For the F-I Beta and the Kibble distributions the differences are the largest.

5.4 Monte Carlo Analysis

In this section we will analyse a sample of $10,000$ cases of each of the bivariate distributions described in Section 5.2. For each case, we have calculated the value of four measures of the difference between the adjusted and the conditional variance, and analysed these differences against marginal and joint properties of the bivariates. In the next subsection we introduce these four measures. In Section 5.4.2 we discuss how the analysis was set up in Matlab. The results of the analysis will be discussed in the remainder of this section.

5.4.1 Difference Measures

The conditional variance and thus difference between the adjusted and conditional variance depends on value of condition Y. In the measures describing this difference we therefore want to take the behaviour of this difference over the values of Y into account. To be able to calculate average differences over Y, we need a bounded domain of evaluation for Y (since it is not possible to calculate the average of a non-constant function over an unbounded domain in general). For these bounds the 5%- and 95%-quantiles of Y have been chosen, so that the observation can be 'surprisingly but not extremely far away from its prior expectation'. The following four measures are employed in this chapter to measure the difference between the adjusted and conditional variance:

RDE. Relative difference when condition y is equal to its prior expectation $E(Y)$:

$$RDE = \frac{Var_Y(X) - Var(X|y = E(Y))}{Var(X|y = E(Y))}. \tag{5.16}$$

The RDE tells us how big the difference between the BL adjusted variance and the conditional variance is when we are least surprised by the value of the observation, that is when the observation is exactly equal to what we had expected it to be.

MRD. Maximum relative difference for observation y on the interval $[y_{0.05}, y_{0.95}]$, where $y_{0.05}$ and $y_{0.95}$ are resp. the $5\%-$ and $95\%-$quantile of Y, calculated as:

$$MRD = \frac{Var_Y(X) - Var(X|y^*)}{Var(X|y^*)}, \tag{5.17}$$

where y^* is found as the value from the interval $[y_{0.05}, y_{0.95}]$ maximising:

$$\left|\frac{Var_y(X) - Var(X|y)}{Var(X|y)}\right|. \tag{5.18}$$

The MRD tells us about the maximum 'damage' we can expect, also when observation is surprisingly but not extremely far away from its prior expectation. So in a military context the MRD corresponds to the 'worst case scenario'.

ARD. Average relative difference for observation y on the interval $[y_{0.05}, y_{0.95}]$, calculated as:

$$ARD = \frac{\displaystyle\int_{y_{0.05}}^{y_{0.95}} \frac{Var_Y(X) - Var(X|y)}{Var(X|y)} dy}{y_{0.95} - y_{0.05}}. \tag{5.19}$$

With the ARD we can assess bias in the error, that is whether the BL adjusted variance is structurally higher or lower than the conditional variance on the evaluated interval for the observation.

$AARD$. Average absolute relative difference for observation y on the interval $[y_{0.05}, y_{0.95}]$, calculated as:

$$AARD = \frac{\displaystyle\int_{y_{0.05}}^{y_{0.95}} \left| \frac{Var_Y(X) - Var(X|y)}{Var(X|y)} \right| dy}{y_{0.95} - y_{0.05}}. \tag{5.20}$$

The AARD, finally, measures the average error, whether positive or negative, we make in the approximation of the conditional variance, on the evaluated interval for the observation.

Note that both the ARD and AARD are not probability weighted errors over the $5 - 95\%$ interquantile range, and thus do not correspond the expected relative difference in this interval.

5.4.2 Implementation of the Analysis in Matlab

For each of the bivariate distributions described in Section 5.2 a sample of 10,000 cases was taken. For each case of each of the bivariate distributions the following steps were undertaken:

Step 1. Sample the parameters defining the distribution according to the sampling strategies described in Section 5.2. Sampling from the Normal distribution was conducted by using the standard random number generator 'rand()' from Matlab which returns a value between 0 and 1 (uniformly) and using the inverse cumulative distribution function of the Normal distribution from a standard Matlab statistics package.

Step 2. Calculate the (product) moments of up to the fourth order of the distributions using the parameters from Step 1.

Step 3. Calculate the conditional mean and variance of the distribution as a function of the condition using the parameters from Step 1.

Step 4. Calculate the Bayes linear adjusted mean and variance of the distribution (as a function of the condition) using the moments calculated at Step 2.

Step 5. Calculate the values of the four difference measures RDE (5.16), MRD (5.4.1), ARD (5.19) and AARD (5.20) described in Section 5.4.1 for the difference between the conditional mean and variance calculated in Step 3. and the Bayes linear adjusted mean and variance calculated in Step 4. For this the marginal $5\%-$ and $95\%-$quantiles of the distribution are needed. These quantiles we derived using the parameters from Step 1., Matlab's 'rand()' function and using the inverse cumulative distribution function of the univariate Beta (F-I Beta), F, and Gamma (Kibble, Cheriyan) distribution from a standard Matlab statistics package.

Step 6. Store the product moment matrix calculated in Step 2. and the values of the difference measures at Step 5.

In the next section the stored difference measurements are analysed. The difference values are further analysed against properties of the marginal distributions in Section 5.4.4, calculated from the stored product moment matrices, and against joint properties in Section 5.4.5, again calculated from the stored product moment matrices.

5.4.3 Overall Results

In Table 5.2 the mean, the standard deviation and the maximum of each of the four difference measures is displayed for a sample of $10,000$ cases from each of the four distributions. The difference between the adjusted and conditional variance d_{var} is the largest in the Filon-Isserlis Beta sample, for all four difference measures. The average absolute difference in this sample is 75%. Even when the observed value is least surprising (equal to its expected value), the average difference observed is still 5%. The *average* absolute maximum difference is of order 10^5%, and on average the conditional variance is overestimated by 59% for the bivariate Beta distributions. In Section 5.3 it was already discussed that this distribution has the largest fluctuations of the conditional variance on the $5\% - 95\%$ interquantile range of the condition.

The Cheriyan distributions were shown to have by far the smallest fluctuations of the conditional variance. And indeed the adjusted variance is also by far the best approximation of the conditional variance for Cheriyan distributions. The average absolute difference is 2.2% and the maximum AARD in the sample is 15% for this distribution. The average bias (ARD) is 0.06% and the average difference for the condition equal to its expectation is about the same, 0.04%. For the F distribution we find an average AARD of 19%, for the Kibble distribution the average AARD is slightly higher with 27%. The variation of the AARD is also slightly higher for the Kibble distribution with a standard deviation of 18% against 11% for the F distribution, as are the average absolute maximum differences with 154% against 51%. The Kibble distribution on the other hand has a slightly smaller bias (ARD) and the adjusted variance is exact for this distribution when the condition is equal to its expectation, as shown in Section 5.3.3.

The results in Table 5.2 show that the adjusted variance is not a very close approximation of the conditional variance, for the four distributions considered. In the next two sections we will analyse the differences against marginal and joint properties of the distributions, to find out if there are properties that can indicate bad approximation and that can enable us to anticipate this.

Table 5.2: Differences between adjusted and conditional variance for sample of $10,000$ cases from each bivariate distribution family, differences in %.

F-I Beta	mean*	st.dev.*	max	**F**	mean*	st.dev.*	max
AARD	75.41	1431.95	$1.34\cdot10^5$	AARD	18.77	10.90	59.98
MRD	$1.16\cdot10^5$	$1.03\cdot10^7$	$1.03\cdot10^9$	MRD	51.19	42.15	256.03
ARD	58.56	1430.59	$1.34\cdot10^5$	ARD	1.02	3.25	18.83
RDE	4.91	10.38	199.51	RDE	1.87	2.36	13.95
Kibble	mean*	st.dev.	max	**Cheriyan**	mean*	st.dev.	max
AARD	26.53	18.47	315.11	AARD	2.20	1.68	15.16
MRD	153.58	804.72	$4.81\cdot10^4$	MRD	4.58	3.66	37.72
ARD	0.07	8.62	315.11	ARD	0.04	0.11	2.13
RDE	0.00	0.00	0.00	RDE	0.06	0.06	1.00

* For the MRD the mean and the standard deviation are calculated from the absolute value of the MRD.

5.4.4 Evaluation against Marginal Properties

The Bayes linear methodology is a method of moments. Firstly, a BL model is fully specified by first and second order (product) moments of the quantities in the base of the model, and secondly the belief adjustments are all operations on (product) moments only. Higher order marginal and product moments for the distributions under consideration, or functions thereof, are thus the obvious properties to characterise the marginal and joint behaviour of the bivariates.

To allow for cross-distribution comparisons normalised central moments are used. As the first two normalised central moments of a variable are always 0 and 1 respectively (when the first two moments are finite), the first two normalised central moments of interest are the third order, or skewness, and the fourth order, or kurtosis. A Normally distributed variable is not skewed, i.e. has a skewness of 0, and has a kurtosis of 3 (or: an excess kurtosis of 0). Since the adjusted variance is equal to the conditional variance for bivariate Normally distributed variables, we are especially interested if a deviation of either of the two variables from a skewness of 0 and a kurtosis of 3 has an impact of the difference d_{var}.

In Table 5.3 the average, standard deviation and extreme value of the marginal skewness and excess kurtosis are given for different percentile ranges of the AARD of the sample of Beta distributions. The average, standard deviation and maximum AARD are also included in the table. We find that the average skewness of X increases with an increasing AARD. The same holds for the average skewness of Y, but note that for a positive skewness of Y the average skewness is close to zero for high percentiles of the AARD, while for a negative skewness the average is close to zero for much lower percentile ranges. For the lower percentiles of the AARD the average excess kurtosis of Y is much larger than for the higher percentiles, while the excess kurtosis of X has a higher average for higher percentiles of the AARD.

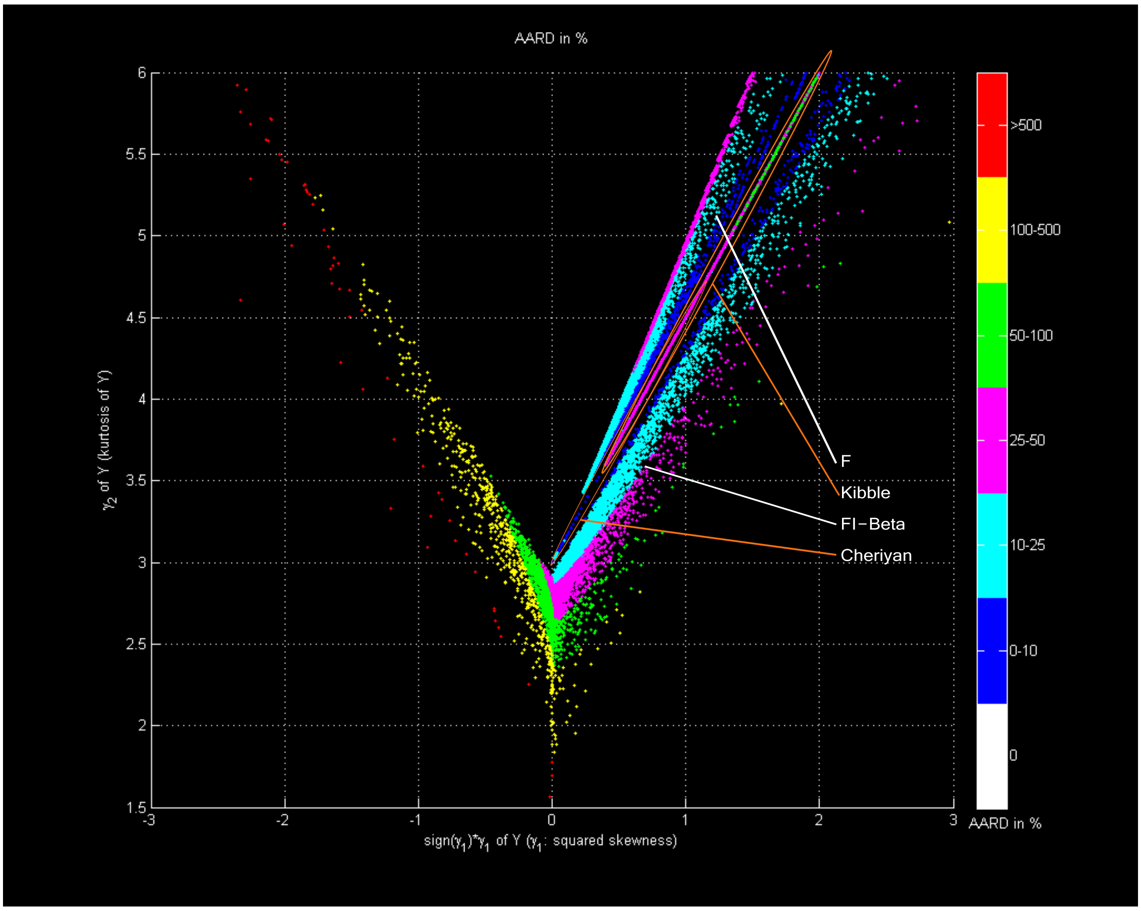

Figure 5.6: The AARD against the squared skewness and kurtosis for Beta, F and Gamma distributions.

For the F distribution the average skewness and excess kurtosis of X also increase with an increasing AARD, but for the average skewness and excess kurtosis of condition Y no monotonous relationship with the AARD is found. For the Kibble and Cheriyan distribution no monotonous relationship is found for either skewness and excess kurtosis, for both variables. The tables for these distributions are included in Appendix B.1. We find a clearer relationship between the skewness, kurtosis and the AARD when we plot the AARD on the Pearson diagram of Y, i.e. against the squared skewness and regular kurtosis of Y, see Figure 5.6. For the Beta and F distributions the AARD gets smaller when the (skewness, kurtosis)-value of the marginal distribution of condition Y approaches the Pearson Type III line. Yet in the limit, on the Type III line itself, the AARD is not necessarily small, as the AARDs for the Kibble distributions clearly show. Figure 5.6 is zoomed in on smaller skewness and kurtosis values to enable a clearer distinction of individual distributions.

Table 5.3: Mean, standard deviation and maximum value of skewness and excess kurtosis for different percentile ranges of AARD of $10,000$ FI-beta distributions. AARD in %.

F-I Beta		Percentiles of AARD											
		0-1%	1-10%	10-20%	20-30%	30-40%	40-50%	50-60%	60-70%	70-80%	80-90%	90-99%	99-100%
AARD	mean	1.412	7.658	13.148	17.192	20.841	24.734	29.490	35.543	44.739	62.368	139.987	$3.73{\cdot}10^3$
	st.dev.	1.043	2.095	1.246	1.077	1.013	1.212	1.579	2.068	3.283	7.912	69.058	$1.39{\cdot}10^4$
	max	3.041	10.875	15.263	19.072	22.631	26.843	32.215	39.372	50.834	79.018	435.572	$1.34{\cdot}10^5$
skew $X < 0$	mean	-0.631	-0.391	-0.290	-0.244	-0.221	-0.222	-0.208	-0.176	-0.206	-0.310	-0.260	-0.124
	st.dev.	0.752	0.568	0.297	0.227	0.226	0.253	0.204	0.177	0.197	0.265	0.369	0.000
	min	-3.360	-7.771	-1.949	-1.589	-1.477	-2.160	-1.066	-1.205	-0.919	-0.993	-1.661	-0.124
skew $X > 0$	mean	0.553	0.596	0.684	0.654	0.747	0.710	0.850	1.002	1.079	1.324	1.835	2.928
	st.dev.	1.092	0.860	1.921	1.493	2.112	0.946	1.503	1.676	1.460	1.641	3.034	2.391
	max	7.268	7.961	41.465	25.828	45.226	13.104	23.774	19.242	18.089	20.829	67.854	16.370
skew $Y < 0$	mean					-0.012	-0.024	-0.038	-0.071	-0.117	-0.216	-0.512	-1.397
	st.dev.					0.010	0.026	0.028	0.051	0.075	0.118	0.256	0.444
	min					-0.020	-0.081	-0.120	-0.221	-0.322	-0.538	-1.316	-2.618
skew $Y > 0$	mean	8.990	1.955	0.963	0.644	0.481	0.413	0.319	0.295	0.306	0.319	0.304	0.053
	st.dev.	5.053	1.100	0.594	0.445	0.346	0.399	0.297	0.329	0.344	0.369	0.290	0.061
	max	25.272	7.852	4.826	5.055	2.507	3.577	3.048	3.855	2.709	2.998	1.723	0.140
excess kurtosis X	mean	1.104	0.921	3.548	2.503	4.974	1.369	3.405	4.526	4.066	5.676	15.879	18.767
	st.dev.	6.199	5.201	76.766	35.143	97.064	9.450	32.301	30.319	24.257	27.631	193.970	44.700
	max	58.125	87.330	2416.228	939.397	2784.723	244.327	801.192	499.866	467.316	560.231	5632.619	355.271
excess kurtosis Y	mean	144.968	6.898	1.580	0.638	0.271	0.204	-0.013	-0.074	-0.149	-0.198	-0.003	2.321
	st.dev.	162.023	8.628	2.630	1.477	0.748	1.185	0.547	0.656	0.438	0.441	0.525	2.071
	max	819.904	78.011	29.371	25.931	7.717	15.041	10.887	16.324	7.949	9.126	2.245	9.522

5.4.5 Evaluation against Joint Properties

Like we used marginal moments to describe the marginal properties of the distributions we characterise the joint behaviour of the variables by their product moments. To enable cross-distribution comparison (higher order) correlations $\rho_{i,j} = Corr(X^i, Y^j)$ are calculated from these product moments. To also be able to determine the difference of the joint behaviour from the joint behaviour of joint Normal variables we analyse the differences d_{var} against $\delta_{i,j}$

$$\delta_{i,j} = Corr(X^i, Y^j) - Corr_{Normal}(X^i, Y^j),$$

where $Corr_{Normal}(X^i, Y^j)$ are the higher order correlation of a bivariate Normal distribution with the same first and second order (cross-)moments as the distribution under consideration. Note that by definition $\delta_{1,1} = 0$.

When the correlation is close to zero, the adjusted variance will barely be different from the prior variance. Since the correlation is an expression of (linear) dependence between the bivariates, a correlation close to zero might indicate that one will not learn much about the one variable by observing the other, so the conditional variance might not be very different from the prior variance even though in principle it could be. The *relative* difference between adjusted and conditional variance could thus be very small when correlation is close to 0, especially when the conditional variance is not close to 0.

In Table B.4 the average, standard deviation and maximum of the absolute correlation $|\rho|$ are given for different percentile ranges of the AARD, for all four distributions. The average absolute correlation clearly increases with the percentile ranges of the AARD, for all distributions. The same is found for the absolute value of $\delta_{1,2}$. For both the correlation and $\delta_{1,2}$ however, the average absolute values differ strongly within the same AARD percentile range for the four distributions. For $|\delta_{2,1}|$ and $|\delta_{2,2}|$ no monotonous relationship with the AARD percentile ranges was observed. The tables for $|\delta_{2,1}|$ and $|\delta_{2,2}|$ are included in Appendix B.2.

5.5 Summary and Conclusions

In this chapter we have investigated whether the Bayes linear adjusted variance can be good approximation to the conditional variance. We have analysed the difference between the adjusted and conditional variance for samples from four bivariate distributions that have marginal distributions that cover wide range of Pearson Type I, III and VI distributions: the Filon-Isserlis Beta, an F, Kibble's Gamma and Cheriyan's Gamma distribution.

Table 5.4: Mean, standard deviation and maximum value of the absolute value of the correlation, $|\rho|$, and higher order correlation difference $|\delta_{1,2}|$ for different percentile ranges of AARD of $10,000$ F-I Beta, F, Kibble and Cheriyan distributions. AARD in %.

		Percentiles of AARD											
		0-1%	1-10%	10-20%	20-30%	30-40%	40-50%	50-60%	60-70%	70-80%	80-90%	90-99%	99-100%
F-I Beta $\lvert\rho\rvert$	mean	0.087	0.237	0.356	0.418	0.451	0.470	0.483	0.498	0.505	0.519	0.572	0.661
	st.dev.	0.093	0.161	0.192	0.213	0.227	0.235	0.242	0.250	0.251	0.251	0.247	0.241
	max	0.738	0.990	0.979	0.989	0.997	0.999	1.000	0.997	0.999	0.999	1.000	0.991
F $\lvert\rho\rvert$	mean	0.067	0.165	0.264	0.323	0.356	0.388	0.429	0.475	0.537	0.616	0.720	0.816
	st.dev.	0.021	0.062	0.077	0.088	0.097	0.101	0.111	0.116	0.119	0.125	0.129	0.108
	max	0.100	0.302	0.405	0.458	0.505	0.545	0.591	0.647	0.710	0.795	0.885	0.912
Kibble $\lvert\rho\rvert$	mean	0.005	0.062	0.170	0.287	0.411	0.531	0.643	0.701	0.696	0.702	0.774	0.933
	st.dev.	0.004	0.035	0.059	0.086	0.119	0.164	0.201	0.208	0.211	0.202	0.149	0.074
	max	0.013	0.151	0.315	0.470	0.650	0.832	0.995	1.000	1.000	0.999	0.999	0.998
Cheriyan $\lvert\rho\rvert$	mean	0.223	0.306	0.321	0.370	0.396	0.438	0.469	0.497	0.556	0.597	0.632	0.695
	st.dev.	0.301	0.298	0.246	0.215	0.202	0.209	0.205	0.201	0.203	0.180	0.175	0.155
	max	0.906	0.950	0.974	0.953	0.977	0.975	0.978	0.985	0.991	0.974	0.985	0.988
F-I Beta $\lvert\delta_{1,2}\rvert$	mean	$4.4{\cdot}10^{-4}$	$7.0{\cdot}10^{-4}$	$1.2{\cdot}10^{-3}$	$1.7{\cdot}10^{-3}$	$2.3{\cdot}10^{-3}$	$3.2{\cdot}10^{-3}$	$4.4{\cdot}10^{-3}$	$6.4{\cdot}10^{-3}$	$9.4{\cdot}10^{-3}$	$1.6{\cdot}10^{-2}$	$3.4{\cdot}10^{-2}$	$1.1{\cdot}10^{-1}$
	st.dev.	$1.3{\cdot}10^{-3}$	$1.4{\cdot}10^{-3}$	$1.9{\cdot}10^{-3}$	$2.6{\cdot}10^{-3}$	$3.2{\cdot}10^{-3}$	$4.0{\cdot}10^{-3}$	$5.0{\cdot}10^{-3}$	$6.5{\cdot}10^{-3}$	$8.3{\cdot}10^{-3}$	$1.2{\cdot}10^{-2}$	$2.0{\cdot}10^{-2}$	$3.6{\cdot}10^{-2}$
	max	$8.0{\cdot}10^{-3}$	$2.5{\cdot}10^{-2}$	$1.1{\cdot}10^{-2}$	$2.8{\cdot}10^{-2}$	$1.7{\cdot}10^{-2}$	$2.0{\cdot}10^{-2}$	$2.4{\cdot}10^{-2}$	$2.8{\cdot}10^{-2}$	$3.6{\cdot}10^{-2}$	$4.8{\cdot}10^{-2}$	$9.6{\cdot}10^{-2}$	$2.1{\cdot}10^{-1}$
F $\lvert\delta_{1,2}\rvert$	mean	$4.8{\cdot}10^{-5}$	$1.6{\cdot}10^{-4}$	$2.7{\cdot}10^{-4}$	$3.2{\cdot}10^{-4}$	$3.5{\cdot}10^{-4}$	$3.9{\cdot}10^{-4}$	$4.7{\cdot}10^{-4}$	$7.0{\cdot}10^{-4}$	$9.8{\cdot}10^{-4}$	$2.0{\cdot}10^{-3}$	$6.9{\cdot}10^{-3}$	$1.8{\cdot}10^{-2}$
	st.dev.	$1.5{\cdot}10^{-4}$	$4.2{\cdot}10^{-4}$	$7.9{\cdot}10^{-4}$	$9.3{\cdot}10^{-4}$	$9.5{\cdot}10^{-4}$	$1.0{\cdot}10^{-3}$	$9.7{\cdot}10^{-4}$	$1.5{\cdot}10^{-3}$	$1.5{\cdot}10^{-3}$	$1.9{\cdot}10^{-3}$	$4.3{\cdot}10^{-3}$	$3.5{\cdot}10^{-3}$
	max	$9.0{\cdot}10^{-4}$	$4.7{\cdot}10^{-3}$	$7.0{\cdot}10^{-3}$	$8.4{\cdot}10^{-3}$	$9.1{\cdot}10^{-3}$	$1.0{\cdot}10^{-2}$	$8.1{\cdot}10^{-3}$	$1.4{\cdot}10^{-2}$	$1.7{\cdot}10^{-2}$	$1.9{\cdot}10^{-2}$	$2.9{\cdot}10^{-2}$	$2.5{\cdot}10^{-2}$
Kibble $\lvert\delta_{1,2}\rvert$	mean	$5.1{\cdot}10^{-5}$	$5.2{\cdot}10^{-4}$	$1.5{\cdot}10^{-3}$	$2.5{\cdot}10^{-3}$	$2.8{\cdot}10^{-3}$	$4.1{\cdot}10^{-3}$	$5.3{\cdot}10^{-3}$	$6.7{\cdot}10^{-3}$	$1.2{\cdot}10^{-2}$	$2.7{\cdot}10^{-2}$	$9.8{\cdot}10^{-2}$	$2.5{\cdot}10^{-1}$
	st.dev.	$7.3{\cdot}10^{-5}$	$7.6{\cdot}10^{-4}$	$2.5{\cdot}10^{-3}$	$5.1{\cdot}10^{-3}$	$2.8{\cdot}10^{-3}$	$5.3{\cdot}10^{-3}$	$1.1{\cdot}10^{-2}$	$1.0{\cdot}10^{-2}$	$1.6{\cdot}10^{-2}$	$3.0{\cdot}10^{-2}$	$9.3{\cdot}10^{-2}$	$1.4{\cdot}10^{-1}$
	max	$4.7{\cdot}10^{-4}$	$1.0{\cdot}10^{-2}$	$6.0{\cdot}10^{-2}$	$9.3{\cdot}10^{-2}$	$2.7{\cdot}10^{-2}$	$8.9{\cdot}10^{-2}$	$2.0{\cdot}10^{-1}$	$2.1{\cdot}10^{-1}$	$2.4{\cdot}10^{-1}$	$2.9{\cdot}10^{-1}$	$5.1{\cdot}10^{-1}$	$5.9{\cdot}10^{-1}$
Cheriyan $\lvert\delta_{1,2}\rvert$	mean	$1.4{\cdot}10^{-7}$	$2.9{\cdot}10^{-7}$	$2.4{\cdot}10^{-7}$	$1.7{\cdot}10^{-7}$	$3.1{\cdot}10^{-7}$	$3.5{\cdot}10^{-7}$	$5.7{\cdot}10^{-7}$	$9.9{\cdot}10^{-7}$	$1.4{\cdot}10^{-6}$	$1.9{\cdot}10^{-6}$	$4.9{\cdot}10^{-6}$	$1.7{\cdot}10^{-5}$
	st.dev.	$5.1{\cdot}10^{-7}$	$2.2{\cdot}10^{-6}$	$1.2{\cdot}10^{-6}$	$2.8{\cdot}10^{-7}$	$1.4{\cdot}10^{-6}$	$4.7{\cdot}10^{-7}$	$1.2{\cdot}10^{-6}$	$2.8{\cdot}10^{-6}$	$8.5{\cdot}10^{-6}$	$8.5{\cdot}10^{-6}$	$3.2{\cdot}10^{-5}$	$3.1{\cdot}10^{-5}$
	max	$4.1{\cdot}10^{-6}$	$5.4{\cdot}10^{-5}$	$3.0{\cdot}10^{-5}$	$4.9{\cdot}10^{-6}$	$3.7{\cdot}10^{-5}$	$4.9{\cdot}10^{-6}$	$2.2{\cdot}10^{-5}$	$4.7{\cdot}10^{-5}$	$2.6{\cdot}10^{-4}$	$2.6{\cdot}10^{-4}$	$9.2{\cdot}10^{-4}$	$2.6{\cdot}10^{-4}$

Where the adjusted variance is constant, i.e. does not depend on the value of observations, the conditional variances of the four distributions are not. The Cheriyan distribution however has a far more constant conditional variance than the other three distributions on the $5\%-95\%$ interquantile range of the condition: the minimum value of the conditional variance is on average 92% of the maximum conditional variance on the $5\%-95\%$ interquantile range of the condition. For the other distributions the ratios of the minimum and maximum value of the conditional variance are much lower with 39%, 52% and 42% for resp. the F-I Beta, F and Cheriyan distribution.

The adjusted variance is not a close approximation of the conditional variance for the F-I Beta, F and Kibble distribution. The average absolute error found over the $5\% - 95\%$ interquantile range of the condition, in a sample of $10,000$ cases of each of these distributions, is 75%, 19% and 27% respectively. When the condition is exactly as expected and equal to its mean, the errors are smaller with respectively 5%, 2% for the F-I Beta and F distribution and zero for the Kibble distribution. For F-I Beta distributions the adjusted variance overestimates the conditional variance on average with 59%. For F, Kibble and Cheriyan distributions the average biases are much smaller with resp. 1%, 0.07% and 0.04%.

For Cheriyan distributions the adjusted variance is a much better approximation of the conditional variance, as we might have expected since the conditional variance is relatively much more constant for these distributions. The average absolute error found in the sample of $10,000$ Cheriyan distributions is 2.2%, and the error for the condition equal to its expectation is on average 0.06%.

The skewness and kurtosis of both marginals form no indicator on their own for the size of the approximation error, i.e. we found no monotonous relationship with the size of the approximation error for the marginal skewness and kurtosis that is consistent for all four distributions. In the introduction of this chapter we mentioned that the adjusted and the conditional variance are equal for joint Normally distributed variables and in this chapter we showed that the adjusted variance is by far the best approximation to the conditional variance for the Cheriyan distributions. These Cheriyan distributions have a skewness and kurtosis that are for *both* marginals relatively close to the skewness and kurtosis of the Normal distribution. So having both marginals with a skewness close to 0 (approximately symmetric) and a kurtosis near 3 might indicate a small approximation error, but this needs further investigation.

For the F-I Beta and F distributions we noted that the approximation errors get smaller for (skewness, kurtosis)-values of the condition that are closer to (skewness, kurtosis)-values of

Pearson Type III distributions, as depicted in Figure 5.6.

In Section 5.4.5 we found that the correlation and the higher order correlation difference with the joint Normal distribution $\delta_{1,2}$ are good indicators of the approximation error, in the sense that higher absolute values of these on average indicate a larger approximation error within the same family of distributions. However, the absolute values of the correlation and $\delta_{1,2}$ cannot be used as an indicator of the size of the approximation error for these four distributions. So from an accuracy of the approximation perspective we would like the correlation and $\delta_{1,2}$ to be as small as possible. From a 'learning from observations' perspective however the opposite holds.

Practitioners we would therefore recommend not to use the adjusted variance as an approximation the conditional variance in general. Two exceptions however are in place to this recommendation: the adjusted variance might be a relative good approximation when the correlation is very small and for distributions for which the conditional variance is considered to be relatively constant, as is the case for the bivariate Normal and Cheriyan distribution.

Chapter 6

Bayes Linear Variance Adjustment using Higher Order Information

In the previous chapter it was shown that the Bayes linear adjusted variance might be a poor approximation of the conditional variance, especially when the conditional variance varies strongly with the value of the condition. An alternative way to approximate the conditional variance in the Bayes linear methodology is to calculate the variance from the adjusted first and second moment. In this chapter we evaluate the possible benefits of using higher order (product) moment information in Bayes linear belief adjustment. We show that this Bayes linear 'adjusted moment variance', the variance calculated from the adjusted first and second moment, can provide a much better approximation to the conditional variance. To calculate this adjusted moment variance, more (higher) moment assessments are needed than for the regular adjusted variance. In Section 6.3 a bivariate generalisation of the extended Pearson-Tukey method from Section 3.5.1.1 is evaluated for this purpose. Next, the approximation of the conditional variance by the adjusted moment variance using Pearson-Tukey approximated moment assessments is evaluated. In this chapter and the next, the term 'moments' will be used for both marginal and product moments together, unless specifically stated otherwise.

6.1 Bayes Linear Adjustment with Higher Order Information

The Bayes linear methodology has been introduced as a second order method in Chapter 4. A Bayes linear model is fully specified by first and second order moments of the quantities in the base of the model, and the belief adjustments are all operations using these moments only. Yet, knowledge of higher order moments can be used in Bayes linear adjustment by taking higher powers of the quantities in the base of a Bayes linear model. Goldstein, developer of the methodology, phrases this as follows: (Goldstein 1994, p.121): "Within a traditional Bayes formulation, [the specification of means and (co)variances] may be viewed as a 'second order' specification. However, by including as many functional forms as we feel are relevant to the problem, we may specify whatever product order of moments we choose, subject only to the constraint that we are able to make all of the necessary quantifications. We may even, in the limit, choose to specify all joint prior moments, which is equivalent to making a full probabilistic specification. Thus, the [...] specification [of means and (co)variances] may be viewed as reducing the full probabilistic approach to whatever level of detail we feel is both within our ability to specify and adequate to the problem at hand."

In this chapter the Bayes linear adjustment of moments is again viewed as approximation to their conditional counterparts in a full probabilistic approach. As the (conditional) variance can be derived as the (conditional) second moment of a variable minus the square of the (conditional) first, it will be evaluated whether the first two linearly adjusted moments can be used to form a better approximation to the conditional variance than the regular adjusted variance investigated in the previous chapter. The variance calculated from the adjusted moment will be referred to as 'adjusted moment variance' with notation ${}^{M}Var_Y(X)$, so:

$$^{M}Var_Y(X) = E_Y(X^2) - E_Y(X)^2,$$

where

$$\begin{aligned} E_Y(X) &= E(X) + Cov(X,Y)Var(Y)^{-1}(Y - E(Y)), \quad \text{and} \\ E_Y(X^2) &= E(X^2) + Cov(X^2,Y)Var(Y)^{-1}(Y - E(Y)). \end{aligned}$$

Note that by including X^2 in the base of a Bayes linear model, third order specifications are needed to be able to perform the linear adjustment of the expectation of X^2, since $Cov(X^2, Y) =$

$E(X^2Y) - E(X^2)E(Y)$.

But like higher powers of quantities can be included in the base of the model for which the expectations are to be adjusted, also higher powers of the quantities that are going the be observed can be included in the model. The notation $E_{\mathbf{Y}_n}(X^i)$ will be used here for the expectation of X^i that is linearly adjusted by the first n powers of observation Y, $\mathbf{Y}_n = [Y^1, \ldots, Y^n]$. From (4.1) in Chapter 4 it can be found that:

$$E_{\mathbf{Y}_n}(X^i) = E(X^i) + Cov(X^i, \mathbf{Y}_n)Var(\mathbf{Y}_n)^\dagger(\mathbf{Y}_n - E(\mathbf{Y}_n)).$$

where $Cov(X^i, \mathbf{Y}_n)$ is the (1 x n)-vector of covariances $Cov(X^i, Y^j)$ for $j = 1, \ldots, n$, $Var(\mathbf{Y}_n)^\dagger$ the Moore-Penrose inverse of the (n x n)-variance-covariance matrix of vector $\mathbf{Y}_n$ and $(\mathbf{Y}_n - E(\mathbf{Y}_n))$ the (n x 1)-vector of the differences $(Y^j - E(Y^j))$, $j = 1, \ldots, n$.

$E_{\mathbf{Y}_n}(X^i)$, which will be called the n-order adjusted moment of X^i by Y, is in fact an n-order polynomial in Y:

$$E_{\mathbf{Y}_n}(X^i) = k_0 + \sum_{i=1}^{n} k_i y^i,$$

with coefficients

$$\begin{array}{rcl} k_0 & = & E(X^i) - Cov(X^i, \mathbf{Y}_n)Var(\mathbf{Y}_n)^\dagger E(\mathbf{Y}_n), \quad \text{and} \\ \begin{pmatrix} k_1 & \ldots & k_n \end{pmatrix} & = & Cov(X^i, \mathbf{Y}_n)Var(\mathbf{Y}_n)^\dagger \end{array}$$

The adjusted moment variance derived from the n-order adjusted first and second moment will be referred to as the n-adjusted moment variance, ${}^M Var_{\mathbf{Y}_n}(X)$:

$${}^M Var_{\mathbf{Y}_n}(X) = E_{\mathbf{Y}_n}(X^2) - E_{\mathbf{Y}_n}(X)^2.$$

${}^M Var_{\mathbf{Y}_n}(X)$ is a polynomial of order $2n$ and requires a $2n$-th order moment specification. When the conditional mean is linear in the condition, the Bayes linear adjusted mean is equal to the conditional mean (see Section 5.1, p.61), thus also linear and will have coefficients of higher order equal to zero. A linear conditional mean will thus reduce the adjusted moment variance to be an n-order polynomial.

Recall the remark made in Section 4.2 that to be able to perform Bayes linear adjustment of moments, the moments used in the adjustment need to be finite. This might not be the case when dealing with distribution with heavy or long tails. Transformations of the variables in the base of the belief structure might offer a solution, but this has not been explored in this thesis.

Finally, note that when the required $2n$-th order moment specification is coherent, that also the n-order adjusted moments are. The n-adjusted moment variance will thus always be nonnegative when the moment specification is coherent. In the next section the n-adjusted moment variance will be evaluated as an approximation to the full conditional variance.

6.2 n-Adjusted Moment Variance with Exact Moments

Following the methodology of Section 5.4, the 1- and 2-adjusted moment variances are compared with the conditional variance in this section, for the same samples of $10,000$ cases of the bivariate F-I Beta, F, Kibble and Cheriyan distributions (see Section 5.2 for details about these distributions, and the sampling strategy applied). The difference measures used are defined in Section 5.4.1, and measure the average absolute relative difference (AARD), the maximum relative difference (MRD) and the average relative difference (ARD), all on the interval between the 5%- and the 95%-quantile of the condition Y, and finally the relative difference at $Y = E(Y)$ (RDE), when the condition Y is equal to its expectation.

The results are displayed in Table 6.1. For all four distributions and for all four of the difference measures the 1-adjusted moment variance has a larger mean and maximum difference with the conditional variance than the regular adjusted variance investigated in the previous chapter. For the F-I Beta distribution the AARD rises from 75% for the regular adjusted variance to 520% for the 1-adjusted moment variance. For the F distribution the AARD doubles from 19% to 39%. For both the bivariate Gamma distributions the AARD rises from 27% to 31% (Kibble), and from 2% to 51% (Cheriyan) respectively.

The 2-adjusted moment variance on the other hand constitutes an excellent approximation to the conditional variance for the four distributions considered, much better than the regular adjusted variance. The average AARD and MRD for the bivariate F-I Beta distribution found are $3 \cdot 10^{-6}$% and $4 \cdot 10^{-4}$% respectively. The maximum relative difference encountered in the 10,000 cases considered, for values of condition Y between its 5%- and 95%-quantile, is -3%. For the F, Kibble and Cheriyan distributions the maximum relative differences found in the samples are only $2 \cdot 10^{-6}$%, $4 \cdot 10^{-7}$% and $-8 \cdot 10^{-3}$% respectively.

However, the Bayes linear belief structure is designed to represent someone's beliefs: expectations, degree of confidence in these expectation judgements and judgements about the interrelations of the quantities in the base of the belief structure (Goldstein 1994, p.120), all expressed in moments. Goldstein warns us therefore, as mentioned in the introduction above, that

Table 6.1: Differences of the regular adjusted variance $Var_Y(X)$, the 1- and the 2-adjusted moment variances ${}^M Var_{\mathbf{Y}_1}(X)$ and ${}^M Var_{\mathbf{Y}_1}(X)$ with the conditional variance, when exact moments are used in the linear adjustments, for 10,000 F-I Beta, F, Kibble and Cheriyan distributions. Differences are in percentages.

		$Var_Y(X)$			${}^M Var_{\mathbf{Y}_1}(X)$			${}^M Var_{\mathbf{Y}_2}(X)$		
		mean*	st.dev.*	max	mean*	st.dev.*	max	mean*	st.dev.*	max
F-I Beta	AARD	75.408	1431.948	$1.342 \cdot 10^5$	519.879	$1.660 \cdot 10^4$	$1.175 \cdot 10^6$	$3.344 \cdot 10^{-6}$	$2.520 \cdot 10^{-4}$	$2.505 \cdot 10^{-2}$
	MRD	$1.161 \cdot 10^5$	$1.033 \cdot 10^7$	$1.031 \cdot 10^9$	$1.337 \cdot 10^5$	$7.840 \cdot 10^6$	$-6.847 \cdot 10^8$	$3.655 \cdot 10^{-4}$	$3.139 \cdot 10^{-2}$	-3.106
	ARD	58.556	1430.590	$1.341 \cdot 10^5$	-184.320	7854.760	$-7.176 \cdot 10^5$	$2.487 \cdot 10^{-6}$	$2.520 \cdot 10^{-4}$	$2.505 \cdot 10^{-2}$
	RDE	4.911	10.384	199.510	388.125	$1.534 \cdot 10^4$	$1.416 \cdot 10^6$	$2.499 \cdot 10^{-6}$	$2.376 \cdot 10^{-4}$	$2.374 \cdot 10^{-2}$
F	AARD	18.770	10.902	59.977	39.286	65.217	700.853	$4.512 \cdot 10^{-8}$	$8.443 \cdot 10^{-8}$	$1.157 \cdot 10^{-6}$
	MRD	51.187	42.150	256.026	179.026	353.929	-4139.489	$6.648 \cdot 10^{-8}$	$1.204 \cdot 10^{-7}$	$1.593 \cdot 10^{-6}$
	ARD	1.021	3.251	18.832	4.547	4.187	24.775	$4.344 \cdot 10^{-10}$	$9.573 \cdot 10^{-8}$	$1.157 \cdot 10^{-6}$
	RDE	1.873	2.359	13.951	42.060	57.374	527.846	$3.718 \cdot 10^{-10}$	$9.338 \cdot 10^{-8}$	$1.137 \cdot 10^{-6}$
Kibble	AARD	26.530	18.472	315.114	30.860	453.586	$3.558 \cdot 10^4$	$1.207 \cdot 10^{-10}$	$2.099 \cdot 10^{-9}$	$1.804 \cdot 10^{-7}$
	MRD	153.584	804.716	$4.811 \cdot 10^4$	5346.492	$1.639 \cdot 10^5$	$-1.388 \cdot 10^7$	$2.487 \cdot 10^{-10}$	$4.331 \cdot 10^{-9}$	$3.792 \cdot 10^{-7}$
	ARD	0.071	8.618	315.114	16.264	454.739	$3.558 \cdot 10^4$	$2.206 \cdot 10^{-11}$	$2.087 \cdot 10^{-9}$	$1.789 \cdot 10^{-7}$
	RDE	0	0	0	381.737	7114.959	$6.359 \cdot 10^5$	$2.102 \cdot 10^{-11}$	$1.831 \cdot 10^{-9}$	$1.538 \cdot 10^{-7}$
Cheriyan	AARD	2.202	1.681	15.155	51.082	124.796	3886.515	$5.176 \cdot 10^{-5}$	$2.130 \cdot 10^{-4}$	$7.521 \cdot 10^{-3}$
	MRD	4.584	3.659	37.725	133.707	330.977	$-1.015 \cdot 10^4$	$5.667 \cdot 10^{-5}$	$2.314 \cdot 10^{-4}$	$-8.291 \cdot 10^{-3}$
	ARD	0.044	0.109	2.135	7.022	17.150	531.335	$-4.279 \cdot 10^{-7}$	$2.192 \cdot 10^{-4}$	$-7.521 \cdot 10^{-3}$
	RDE	0.056	0.056	1.000	71.564	174.752	5447.620	$-4.273 \cdot 10^{-7}$	$2.191 \cdot 10^{-4}$	$-7.514 \cdot 10^{-3}$

* For the MRD the mean and the standard deviation are calculated from the absolute value of the MRD.

we are constrained by someone confidently being able to provide all of the necessary moment quantifications. In the next section a method is introduced to derive higher order moments for two variables from subjective assessments. In Section 6.4 this analysis will be repeated for moments derived using this method, to evaluate whether the 2-adjusted moment variance then still provides a better approximation to the conditional variance then the regular adjusted variance.

6.3 Bivariate Pearson-Tukey Moment Derivation

In Section 3.5 it was argued that the direct assessments of moments is generally advised against, since especially the higher moments are usually not observable quantities for the assessors. Pearson & Tukey (1965) have provided a procedure for deriving first and second moments (means and variances) from quantile assessments, which have been shown to provide accurate approximations for a wide selection of distributions (see Section 3.5 for details). Keefer & Bodily (1983) simplified the procedure proposed by Pearson and Tukey by proposing a three point distribution approximation from which marginal moments are estimated as Equation (3.1), which will be repeated here:

$$E(X^n) = 0.185(x_{0.05})^n + 0.63(x_{0.50})^n + 0.185(x_{0.95})^n,$$

where $x_{0.05}$, $x_{0.50}$ and $x_{0.95}$ are resp. the 5%−, 50%− and 95%−quantiles of variable X. The extended Pearson-Tukey method of Keefer and Bodily has the advantage that also higher marginal moments than the second can be estimated with it, while still using only the same three quantiles.

Yet, assessments of product-moments can not be derived with this method. The Pearson-Tukey method is based on the observation that the h%-distance:

$$h\text{\%-distance} = \frac{x_{(100-h)} - x_h}{\sqrt{Var(X)}},$$

is surprisingly constant for many well-known distributions (Pearson & Tukey 1965), where $h = 5$ in the extended Pearson-Tukey method. Since this observation trivially also holds for conditional marginal distributions, Keefer and Bodily note that a multivariate generalisation of method, by using conditional quantiles, is straightforward. We will work out the bivariate case here. A three point extended Pearson-Tukey approximation to the conditional distribution of X

given $Y = y$ is constructed by assigning probability mass 0.185, 0.63 and 0.185 to respectively the 5%−, 50%− and 95%−conditional quantiles of X given $Y = y$: $x_{0.05|y}$, $x_{0.50|y}$ and $x_{0.95|y}$. So we have that

$$\begin{aligned} P\left(X = x_{0.05|y} | Y = y\right) &= P\left(X = x_{0.95|y} | Y = y\right) = 0.185, \quad \text{and} \\ P\left(X = x_{0.50|y} | Y = y\right) &= 0.63. \end{aligned}$$

By conditioning on the 5%−, 50%− and 95%−quantiles of condition Y, and assigning the probability masses 0.185, 0.63 and 0.185 respectively to these unconditional quantiles of Y (the univariate 3-point extended Pearson-Tukey approximation), we can thus construct a 9-point bivariate distribution approximation in the following way:

$$\begin{aligned} P\begin{pmatrix} X = x_{0.05|y_{0.05}}, Y = y_{0.05} \\ X = x_{0.50|y_{0.05}}, Y = y_{0.05} \\ X = x_{0.95|y_{0.05}}, Y = y_{0.05} \\ X = x_{0.05|y_{0.50}}, Y = y_{0.50} \\ X = x_{0.50|y_{0.50}}, Y = y_{0.50} \\ X = x_{0.95|y_{0.50}}, Y = y_{0.50} \\ X = x_{0.05|y_{0.95}}, Y = y_{0.95} \\ X = x_{0.50|y_{0.95}}, Y = y_{0.95} \\ X = x_{0.95|y_{0.95}}, Y = y_{0.95} \end{pmatrix} &= \begin{pmatrix} P\left(X = x_{0.05|y_{0.05}} | Y = y_{0.05}\right) \cdot P\left(Y = y_{0.05}\right) \\ P\left(X = x_{0.50|y_{0.05}} | Y = y_{0.05}\right) \cdot P\left(Y = y_{0.05}\right) \\ P\left(X = x_{0.95|y_{0.05}} | Y = y_{0.05}\right) \cdot P\left(Y = y_{0.05}\right) \\ P\left(X = x_{0.05|y_{0.50}} | Y = y_{0.50}\right) \cdot P\left(Y = y_{0.50}\right) \\ P\left(X = x_{0.50|y_{0.50}} | Y = y_{0.50}\right) \cdot P\left(Y = y_{0.50}\right) \\ P\left(X = x_{0.95|y_{0.50}} | Y = y_{0.50}\right) \cdot P\left(Y = y_{0.50}\right) \\ P\left(X = x_{0.05|y_{0.95}} | Y = y_{0.95}\right) \cdot P\left(Y = y_{0.95}\right) \\ P\left(X = x_{0.50|y_{0.95}} | Y = y_{0.95}\right) \cdot P\left(Y = y_{0.95}\right) \\ P\left(X = x_{0.95|y_{0.95}} | Y = y_{0.95}\right) \cdot P\left(Y = y_{0.95}\right) \end{pmatrix} \\ &= \begin{pmatrix} 0.185^2 \\ 0.63 \cdot 0.185 \\ 0.185^2 \\ 0.185 \cdot 0.63 \\ 0.63^2 \\ 0.185 \cdot 0.63 \\ 0.185^2 \\ 0.63 \cdot 0.185 \\ 0.185^2 \end{pmatrix} = \begin{pmatrix} 0.034225 \\ 0.11655 \\ 0.034225 \\ 0.11655 \\ 0.3969 \\ 0.11655 \\ 0.034225 \\ 0.11655 \\ 0.034225 \end{pmatrix}. \end{aligned} \tag{6.1}$$

In Figure 6.1 the nine points of the bivariate distribution approximation are depicted for two positively dependent variables. Note that the roles of X and Y in the distribution approximation are not symmetric: there are only three distinct values for Y and (usually) nine for X.

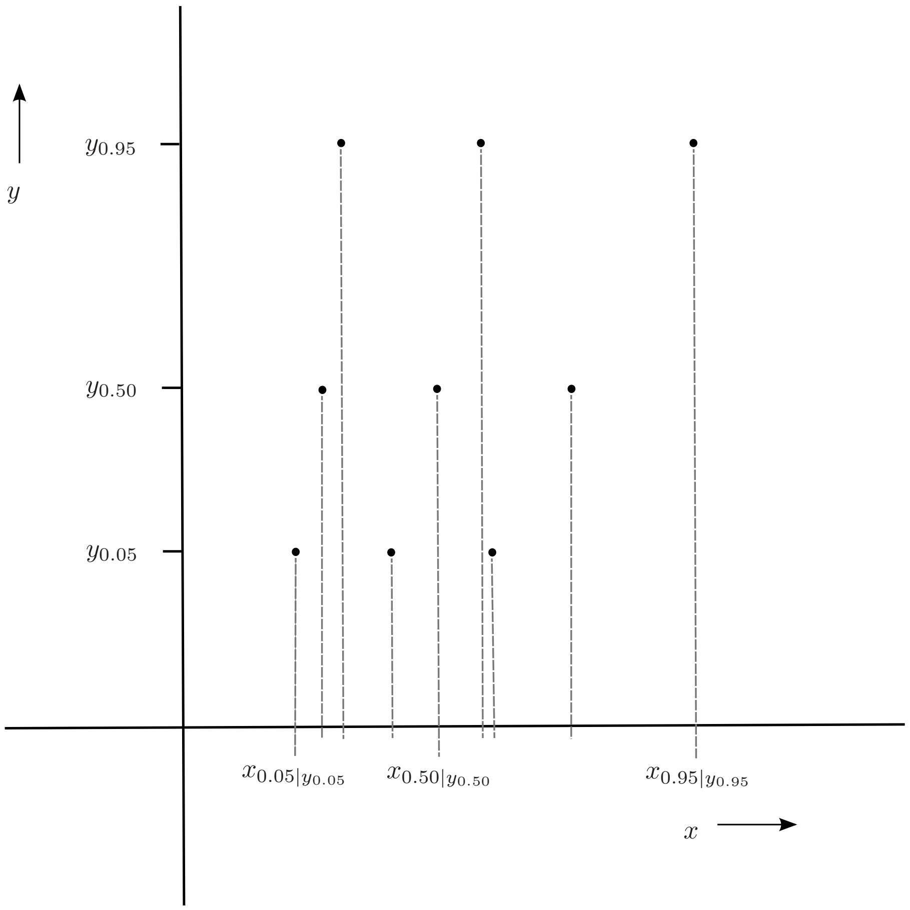

Figure 6.1: An example 9−point bivariate distribution approximation.

Since the 9-point discrete distribution described here is a fully specified probability distribution, both marginal and product-moments can be calculated for it:

$$E(X^iY^j) \quad = \quad \sum_{q_y=0.05,0.50,0.95} \sum_{q_x=0.05,0.50,0.95} p_{q_x,q_y} \cdot x^i_{q_x|y_{q_y}} \cdot y^j_{q_y}, \tag{6.2}$$

with

$$p_{q_x,q_y} \quad = \quad \begin{cases} 0.3969, & \text{when } q_x = q_y = 0.50 \\ 0.11655, & \text{when either } q_x = 0.50 \text{ or } q_y = 0.50\text{, but not both} \\ 0.034225, & \text{otherwise.} \end{cases}$$

The accuracy of the bivariate extended Pearson-Tukey derived moments from (6.2) has been tested for the same distributions for which the accuracy of the adjusted moment variance is evaluated in this chapter, with the exception of the Kibble distribution. For this distribution the conditional quantiles could not be determined with sufficient precision. As an extra reference point, the accuracy has also been tested on the moments of the bivariate Normal distribution.

The sample of $10,000$ Normal distributions used in this evaluation was obtained by sampling for each bivariate Normal distribution both means independently from a Normal distribution with zero mean and a standard deviation of 10. Both standard deviations were sampled, again independently, from the same Normal distribution. The correlation for each distribution, finally, was drawn uniformly between -1 and 1.

In Table 6.2 the mean, standard deviation and the maximum relative errors of the extended Pearson-Tukey approximation of moments are displayed, in percentages, for the first eight marginal moments and product moments of up to the sixth order. The results are based on samples of $10,000$ cases from each distribution. Since all four distributions considered are symmetrical in both variables, only the approximation of the marginal moments of one of the variables is analysed. Symmetrical product moments are left out of Table 6.2 for the same reason.

The relative errors in Table 6.2 look promising. We will discuss the results found for moments of up to the fourth order here, since these will be of interest to us in the next section. The errors found for marginal moments of up to the eighth order, and for product moments of up to the sixth order are summarised in Table 6.2. The average absolute error is for the first four marginal moments not much larger than 6% (for the fourth moment of F-I Beta) and is even below 1% in the Cheriyan sample. Even though the generated cases for the Cheriyan distribution are close to the Normal distribution on the Pearson diagram (see Figure 5.6), i.e. have a skewness and excess kurtosis close to 0, the average absolute relative errors for the marginal moments are quite different from those for the Normal distribution. For the FI-beta distribution the maximum relative errors of approximately -100% are striking. When we plot the relative errors of the marginal moments against the skewness of the variable, in Figure 6.2, we see that these errors occur for highly positive skewed Beta distributions. For the F distributions the moments are more and more underestimated as the skewness increases, and the size of the underestimation increases with the order of the moment (Figure 6.3). For Gamma distributions we find the opposite result. The higher the skewness, the more the moment is overestimated on average, where the magnitude of the overestimation increases with the order of the moment (Figure 6.4).

Keefer & Bodily (1983) evaluate the accuracy of their 3-point discrete Pearson-Tukey distribution approximation for 78 Beta distributions. They find an average absolute error of 0.02% for the first moment, and a maximum of 0.07%. We find much larger errors of resp. 0.58% for the average absolute error and -100% for the maximum. For the variance we find and average absolute error of 2.1% and a maximum of 100%, where Keefer and Bodily report 0.46% and

Table 6.2: The mean, standard deviation and the maximum of the relative errors of the Pearson-Tukey approximation of moments, based on a sample of $10,000$ cases from each of the bivariate Normal, F-I Beta, F and Cheriyan distributions, in percentages.

		Marginal moments							
		$E(X)$	$E(X^2)$	$E(X^3)$	$E(X^4)$	$E(X^5)$	$E(X^6)$	$E(X^7)$	$E(X^8)$
Normal	mean*	0.000	0.027	0.041	1.176	1.854	6.271	8.990	14.909
	st.dev.*	0.000	0.031	0.034	2.019	2.379	9.833	11.606	18.673
	max	0.000	0.105	0.105	-9.690	-9.706	-50.958	-51.056	-80.935
F-I Beta	mean*	0.581	1.469	3.290	6.088	9.621	13.564	17.705	21.925
	st.dev.*	5.182	8.453	12.479	16.971	21.438	25.530	29.079	32.012
	max	-100.000	-100.000	-100.000	-100.000	-100.000	-100.000	-100.000	-100.000
F	mean*	0.009	0.181	1.058	3.190	6.947	12.429	19.365	27.230
	st.dev.*	0.024	0.501	2.903	6.933	11.844	17.250	22.418	26.390
	max	-0.457	-8.789	-40.063	-72.954	-91.193	-97.846	-99.627	-99.970
Cheriyan	mean*	0.000	0.159	0.473	0.940	1.558	2.323	3.233	4.283
	st.dev.*	0.001	0.346	0.984	1.847	2.866	4.016	5.314	6.765
	max	-0.043	19.494	51.814	86.576	110.398	114.229	121.752	137.808

		Product moments								
		$E(XY)$	$E(XY^2)$	$E(XY^3)$	$E(X^2Y^2)$	$E(XY^4)$	$E(X^2Y^3)$	$E(XY^5)$	$E(X^2Y^4)$	$E(X^3Y^3)$
Normal	mean*	0.065	0.062	3.174	0.521	3.556	4.169	19.684	4.327	12.251
	st.dev.*	1.167	0.566	27.509	1.320	33.412	73.483	293.020	8.075	151.867
	max	-92.359	38.700	1730.895	-14.366	2524.298	-4420.878	-21800.063	-68.710	9409.254
F-I Beta	mean*	0.987	1.521	2.796	2.420	5.035	3.420	7.988	5.283	5.089
	st.dev.*	6.913	8.875	11.507	11.018	15.360	12.824	19.551	15.836	15.443
	max	-100.000	-100.000	-100.000	121.031	-100.000	114.703	-100.000	108.015	122.506
F	mean*	0.109	0.746	2.849	1.879	7.060	5.120	13.334	10.598	8.752
	st.dev.*	0.216	1.404	5.178	3.370	11.193	8.399	17.808	15.192	13.274
	max	-2.461	-21.021	-58.959	-32.033	-86.762	-69.980	-97.275	-92.074	-80.265
Cheriyan	mean*	0.000	0.001	0.002	0.159	0.005	0.159	0.011	0.161	0.473
	st.dev.*	0.001	0.008	0.065	0.345	0.213	0.351	0.434	0.404	0.985
	max	-0.043	-0.802	-6.417	19.453	-20.927	19.431	-42.004	-20.928	51.629

* The mean and the standard deviation are taken over the absolute values of the relative errors.

-1.6% respectively. When we only consider Beta distributions with a skewness smaller than 3, we find an average absolute error of 0.07% and still a maximum error as high as 19% for the mean. And for the variance 0.90% and 85%. For samples of size 78 instead of $10,000$ and distributions with a skewness smaller than 3, the average absolute errors for the mean and variance do correspond to the results reported by Keefer and Bodily, but the maximum errors are still an order of magnitude larger. Apparently increasing the sample size allows for the more extreme cases with larger relative errors to have a bigger impact on the results found.

The average absolute relative errors for the product moments are even better. For product moments up to the fourth order these errors are smaller than 3.2%. For the Normal, F-I Beta and the F distribution the average absolute errors for the product moments are comparable. The Normal distribution does have higher maximum errors, and often a higher standard deviation of the errors than the F-I Beta and the F distribution. For the Cheriyan distribution the errors are much smaller, with the largest average absolute error being only 0.5% and a maximum relative error of 52% for $E(X^3Y^3)$.

6.3.1 Limitations of the (bivariate) Pearson-Tukey Moment Derivation

We have evaluated the accuracy of the bivariate extended Pearson-Tukey moments derivation here for bivariate Normal, F-I Beta, F and Cheriyan distributions. Even though the average relative errors reported in Table 6.2 are relatively small, indicating good performance of the approximation, we also found that the errors increase fast with skewness for all distributions (with the exception of the Normal distributions of course, which are not skewed), see 6.2, 6.3 and 6.4. For the distributions investigated in this section the absolute value of the relative errors is different for similar skewness values, so we can not advise a bound for the skewness above which the Pearson-Tukey approximation could be considered inaccurate. It might be an interesting topic for further research to see whether it is possible to derive such a bound for the ratio of the interquantile ranges $(x_{0.50} - x_{0.05})$ and $(x_{0.95} - x_{0.50})$. For the distributions considered here the approximation can be considered accurate for a value of 1 for that ratio. Can a relationship be identified between the accuracy of the Person-Tukey approximation and this ratio? Finally we note that the results reported in this section do not necessarily generalise to other (bivariate) distributions.

6.4 Adjusted Moment Variance with Pearson-Tukey Derived Moments

In Section 6.2 it was shown that the 2-adjusted moment variance is an excellent approximation of the conditional variance for the F-I Beta, F, Kibble and Cheriyan distributions and in the previous section it was found that the bivariate Pearson-Tukey method in general provides a good approximation to the first four order moments. In this section the 2-adjusted moment variance will be evaluated for moments derived from (conditional) quantiles using the bivariate Pearson-Tukey method. But only for the F-I Beta, F and Cheriyan distribution, since the conditional quantiles could not be determined with sufficient precision the Kibble distribution. The results are given in Table 6.3.

For F-I Beta and F distributions the 2-adjusted moment variance with Pearson-Tukey approximated moments provides a much better approximation to the conditional variance than the regular adjusted variance. For the Cheriyan distribution we find the contrary, even though the Pearson-Tukey approximation of moments is by far the most accurate for these distributions. In the previous chapter (Table 5.1 and discussion) we noted that the conditional variance of the Cheriyan distribution is relatively constant in the sample we have taken. Yet, it is unlikely that the relatively large differences between the 2-adjusted moment and conditional variance is due to the applying a quadratic approximation to a relatively constant conditional variance. For in Table 6.3 we find that the bias ARD is almost identical to and the maximum error MRD is very close to the average absolute difference AARD, which implies that the approximation error made is very constant over the values of the condition considered, and must be in the constant term of the quadratic approximation.

For the F distribution the relative difference is completely constant over the values of the condition. The 2-adjusted moment variance underestimates the conditional variance on average with 0.8% in the sample of F distributions, with a maximum difference encountered of -12%. The average bias for F-I Beta distributions found is -0.9%. For the Cheriyan distribution the conditional variance is on average overestimated by 77%.

6.5 Summary and Conclusions

In this chapter we have shown that the 2-adjusted moment variance provides an excellent approximation to the conditional variance for all distribution evaluated: the F-I Beta, F, Kibble and Cheriyan distributions. The highest average absolute error found is of order $10^{-5}\%$ for the

Table 6.3: Differences of the regular adjusted variance $Var_Y(X)$, the 1- and the 2-adjusted moment variances ${}^{M}Var_{\mathbf{Y}_1}(X)$ and ${}^{M}Var_{\mathbf{Y}_2}(X)$ with the conditional variance, when Pearson-Tukey derived moments are used in the linear adjustments, for 10,000 F-I Beta, F and Cheriyan distributions. Differences are in percentages.

		$Var_Y(X)$			${}^{M}Var_{\mathbf{Y}_1}(X)$			${}^{M}Var_{\mathbf{Y}_2}(X)$		
		mean*	st.dev.*	max	mean*	st.dev.*	max	mean*	st.dev.*	max
F-I Beta	AARD	75.408	1431.948	$1.342 \cdot 10^5$	484.184	$1.609 \cdot 10^4$	$1.133 \cdot 10^6$	3.206	13.039	669.437
	MRD	$1.161 \cdot 10^5$	$1.033 \cdot 10^7$	$1.031 \cdot 10^9$	$1.099 \cdot 10^5$	$7.563 \cdot 10^6$	$-7.354 \cdot 10^8$	3.249	15.625	1091.817
	ARD	58.556	1430.590	$1.341 \cdot 10^5$	-151.922	7036.469	$-6.746 \cdot 10^5$	-0.912	13.137	615.718
	RDE	4.911	10.384	199.510	388.222	$1.540 \cdot 10^4$	$1.422 \cdot 10^6$	-0.903	13.578	704.953
F	AARD	18.881	11.087	62.565	31.538	44.440	414.655	0.775	0.861	12.322
	MRD	51.730	43.345	275.108	140.658	244.232	-2585.743	0.775	0.861	-12.322
	ARD	1.064	3.363	20.587	3.978	5.239	32.186	-0.775	0.861	-12.322
	RDE	1.907	2.439	15.171	40.024	54.423	485.479	-0.775	0.861	-12.322
Cheriyan	AARD	2.202	1.681	15.155	106.187	117.705	3905.935	76.742	21.797	100.143
	MRD	4.584	3.659	37.725	167.651	301.493	-9968.396	79.155	21.225	100.186
	ARD	0.044	0.109	2.135	83.630	29.133	627.554	76.532	22.522	100.143
	RDE	0.056	0.056	1.000	148.179	177.535	5542.975	76.546	22.560	100.184

* For the MRD the mean and the standard deviation are calculated from the absolute value of the MRD.

Cheriyan distributions, and the highest average absolute maximum error (MRD) found is in the order of $10^{-4}\%$, for the F-I Beta distributions.

Next we evaluated the approximation of marginal and joint moments from conditional quantiles using the bivariate generalisation of the extended Pearson-Tukey method. For the Normal distribution the average absolute errors in a sample of $10,000$ cases are below 1.2% for the marginal moments of up to the fourth order, and below 3.2% for the product moments of up to the fourth order, the highest error found for these product moments for all four distributions. The average absolute errors were smallest for the Cheriyan sample with errors below 1% and 0.2% for resp. the marginal and product moments of up to the fourth order. The largest errors for marginal moments to the fourth order were found for the F-I Beta sample, with an average absolute error of 6.1%. For this distribution the average absolute errors for the product moments to the fourth order are below 2.8%. For the F distributions we found the average absolute errors to be below 3.2% and 2.9% for resp. the marginal and product moments of up to the fourth order. For all distributions we found that the errors increase with the skewness of the distribution.

The results found for the approximation errors of the first two marginal moments for the Beta distribution are much larger than reported earlier by Keefer & Bodily (1983). Yet when the sample size of 78 is taken as Keefer and Bodily have done, instead of the sample size of $10,000$ used in this research, the average errors found do correspond in general with those found in the earlier research of Keefer and Bodily, but the maximum errors are still and order of magnitude larger. Apparently increasing the sample size allows for the more extreme cases with larger relative errors to have a bigger impact on the results found.

Even though the average relative errors reported here are relatively small, indicating good performance of the moment approximation, we also found that the errors increase fast with skewness for all distributions (with the exception of the Normal distributions of course, which are not skewed), see 6.2, 6.3 and 6.4. For the distributions investigated in this chapter the absolute value of the relative errors is different for similar skewness values, so we can not advise a bound for the skewness above which the Pearson-Tukey approximation could be considered inaccurate. It might be an interesting topic for further research to see whether it is possible to derive such a bound for the ratio of the interquantile ranges $(x_{0.50} - x_{0.05})$ and $(x_{0.95} - x_{0.50})$. For the distributions considered here the approximation can be considered accurate for a value of 1 for that ratio. Can a relationship be identified between the accuracy of the Person-Tukey approximation and this ratio? Finally we note that the results reported in this section do not necessarily generalise to other (bivariate) distributions.

Next we evaluated the 1- and 2-adjusted moment variance approximation to the conditional variance using moments derived with the extended Pearson-Tukey method. For the F-I Beta and F distributions the 2-adjusted moment variance provides a very good approximation to the conditional variance, much better than the regular adjusted variance evaluated in the previous chapter. The average AARD found for the F-I Beta distribution reduced from 75% for the regular adjusted variance to 3.2% for the 2-adjusted moment variance. The average MRD (maximum error) is $1.2 \cdot 10^5\%$ for the regular adjusted variance for this distribution, but only 3.3% for the 2-adjusted moment variance. The average AARD for the F distribution reduced from 19% for the regular adjusted variance to 0.8% for the 2-adjusted moment variance, and the average MRD from 52% to 0.8%.

For the Cheriyan distributions we find the opposite however. The regular adjusted variance provides a much better approximation than the 1- and 2-adjusted moment variances, even though the Pearson-Tukey approximation of moments is by far the most accurate for these distributions. The average AARD is 2.2% for the regular adjusted variance, and 77% for the 2-adjusted moment variance.

The Cheriyan distribution has by far the most constant conditional variance (see Table 5.1). In Section 5.5 we suggested that the regular adjusted variance might be a good approximation for distributions for which the conditional variance is considered to be relatively constant over the condition. It might a good suggestion for further research to see whether a decision statistic can be derived that can help to choose between the regular adjusted and the 2-adjusted moment variance to approximate the conditional variance. It might be possible to base this decision statistic on the conditional quantiles used in the bivariate Pearson-Tukey method, which are readily available. From these conditional quantiles, the $5\%-$, $50\%-$ and the $95\%-$quantiles, the conditional variance can be approximated using the univariate Pearson-Tukey method for three values of the condition: the $5\%-$, $50\%-$ and the $95\%-$quantiles of the condition. When these three derived conditional variances are relatively constant, the regular adjusted variance might be preferred as approximation. If not, the 2-adjusted moment variance might be the better option.

Striking in the approximation results of the 2-adjusted moment variance is that the bias in the approximation error is relatively large compared to the average absolute error. For the F-I Beta distribution the bias is -0.9%, and the average absolute error 3.2%. For the F and Cheriyan distributions the approximation error is practically constant over the evaluation interval of the condition, with biases of resp. -0.8% and 77% for average absolute errors of as well -0.8% and 77%. It would be an interesting subject for further research to see if this bias could be reduced.

Finally, in this chapter we have worked with exact (conditional) quantile assessments. In reality experts might not be willing or able to assess these precisely, and indifferent to small changes in their assessments. The impact of small changes in the quantile assessments has not been taken into consideration in this evaluation, and is an interesting topic for further research.

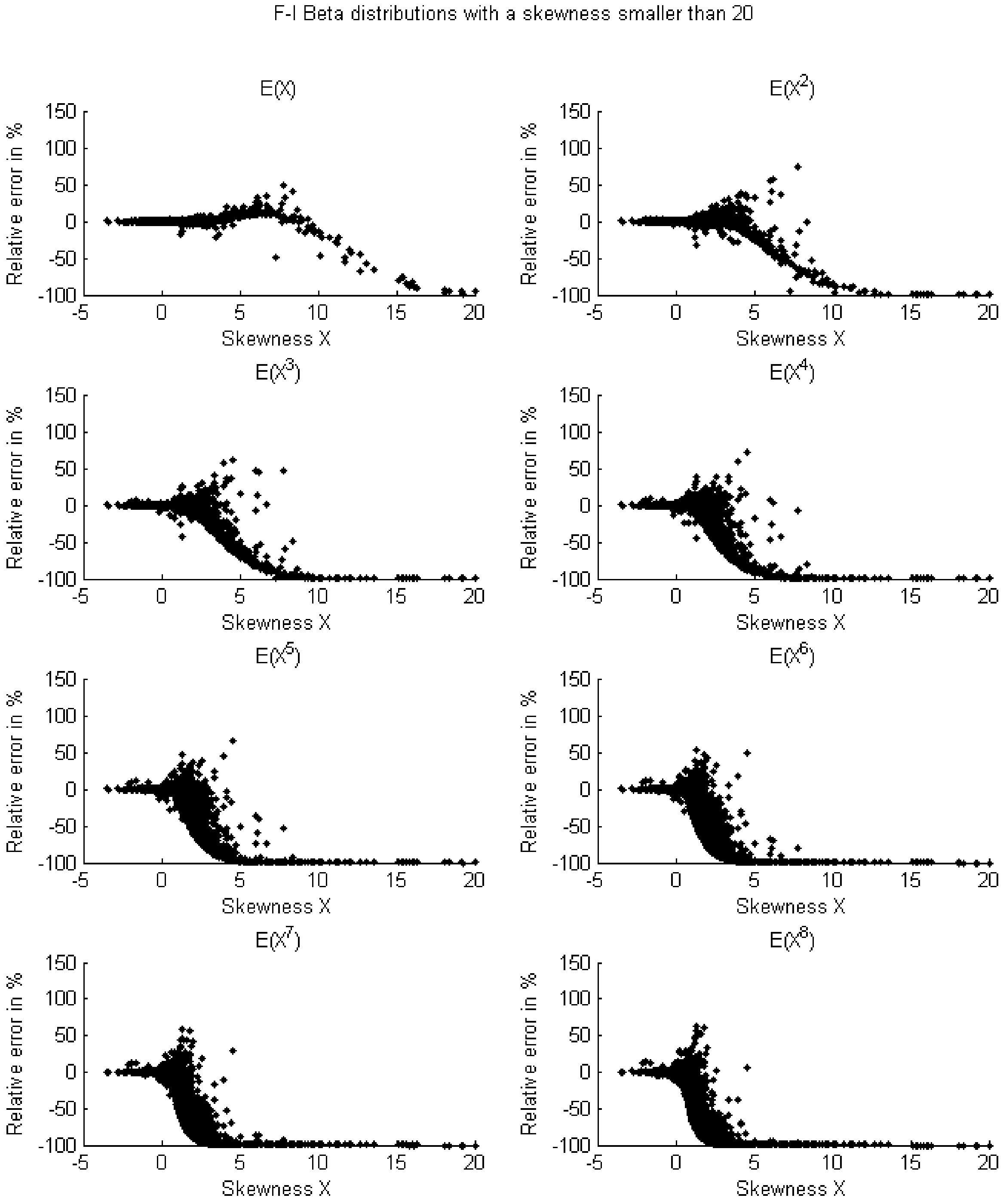

Figure 6.2: Relative errors in % of Pearson-Tukey approximated moments against the skewness for Beta distributions with a skewness smaller than 20.

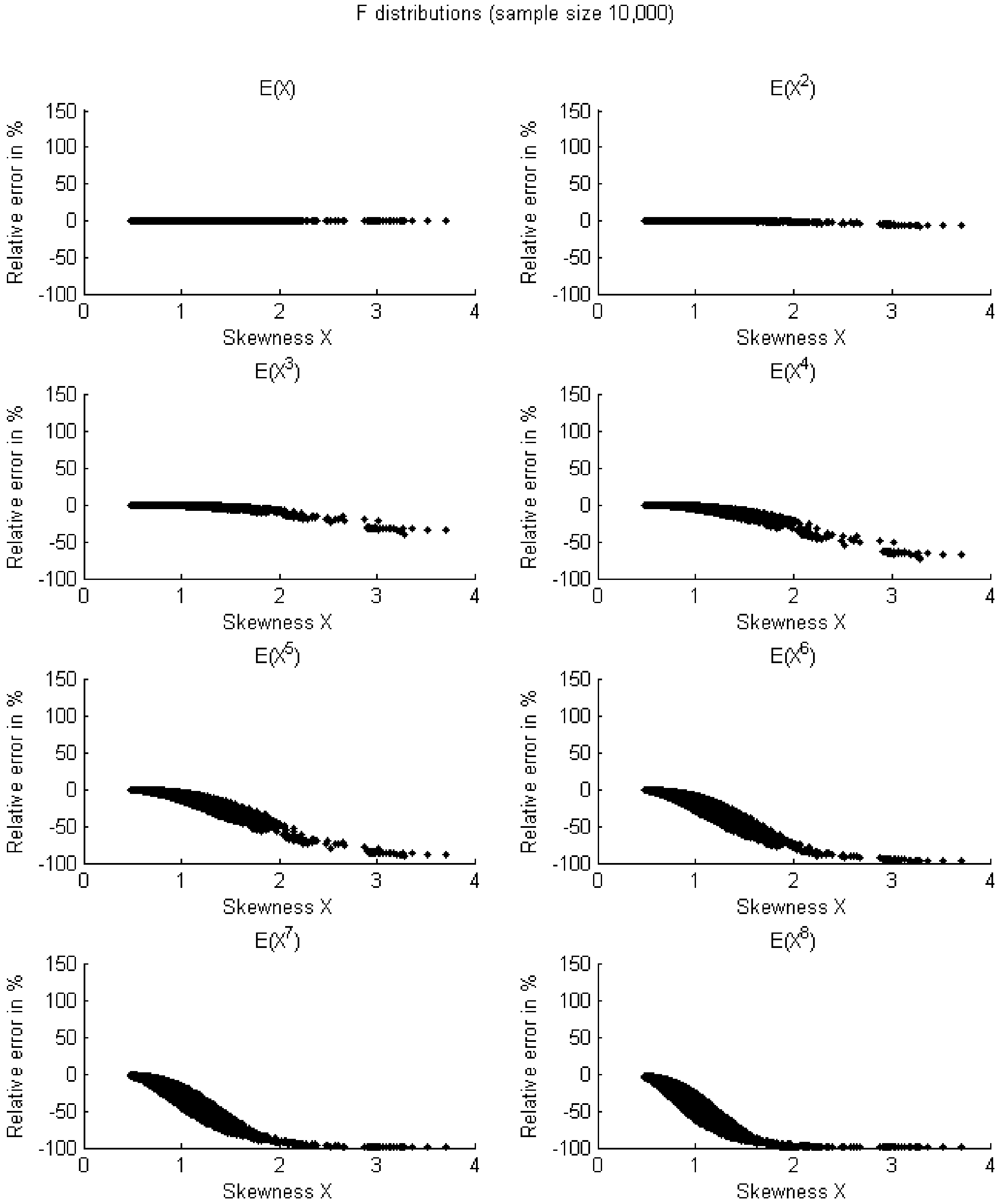

Figure 6.3: Relative errors in % of Pearson-Tukey approximated moments against the skewness for 10,000 F distributions.

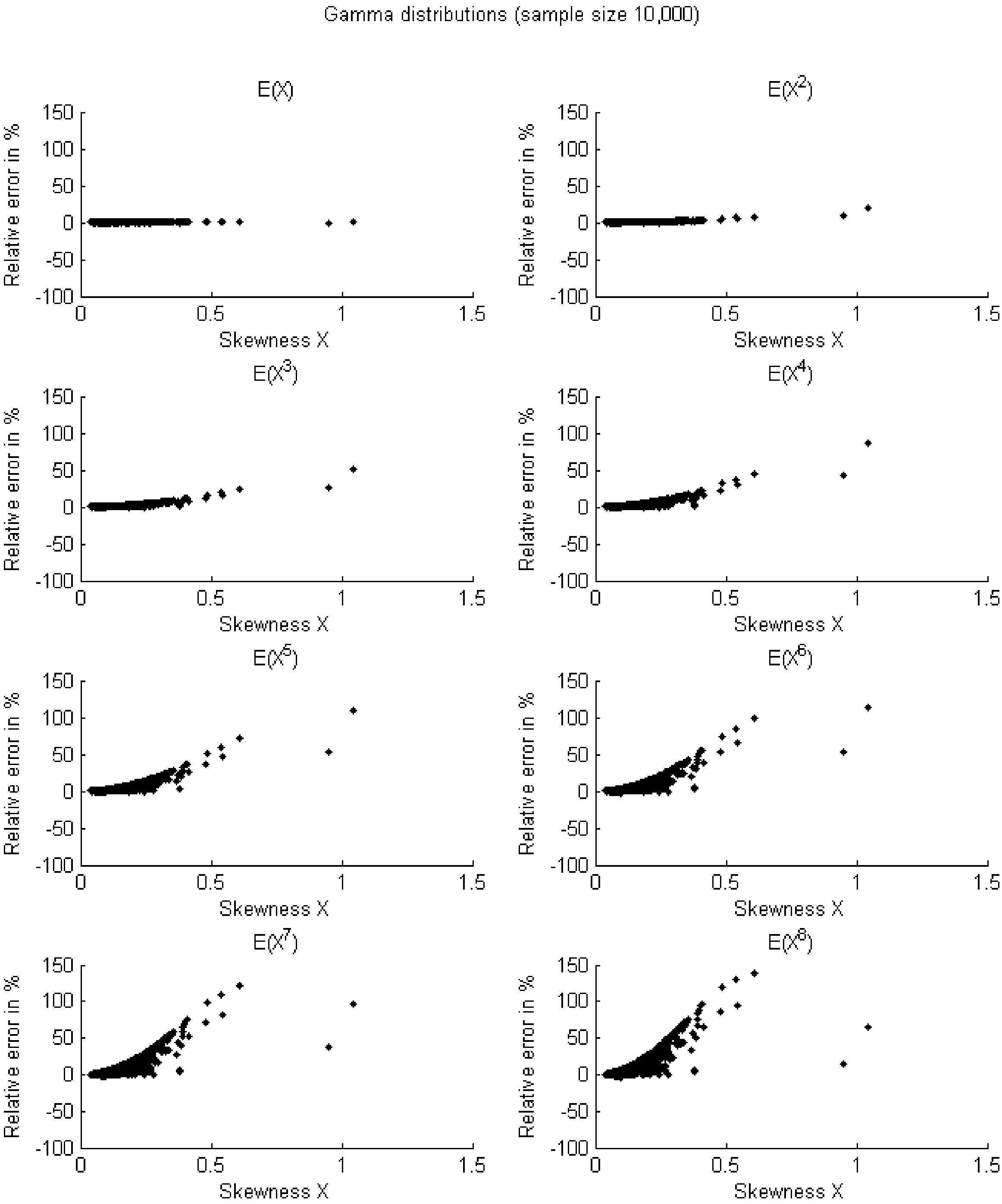

Figure 6.4: Relative errors in % of Pearson-Tukey approximated moments against the skewness for 10,000 Gamma distributions.

Chapter 7

Combining Moment Assessments of Multiple Experts

In this chapter we will discuss the mathematical aggregation of moment assessments of different experts, and develop a performance based combination method from first principles. We will take the situation in which a decision maker (DM) wishes to base his moment assessments solely on the experts' assessments, with as special case a DM who wishes to base his Bayes linear belief structure on those of individual experts. This problem is also referred to as the expert problem (French 1985, French 2011), and has been discussed for the aggregation of probabilistic assessments in Section 3.6. Since the behavioural methods for aggregating expert assessments described in that section do not depend on the type of assessments queried for, the focus in this chapter is on mathematical aggregation methods.

In Section 7.1 we translate the properties that mathematical combinations of probabilities can possess, discussed in Section 3.6.1, to a moment context. These properties will be more formally defined in Section 7.2, where we show that the requirement for an aggregation method for moments assessments to possess the marginalisation and zero preservation properties is equivalent to requiring the aggregation method to be a linear pool of the experts assessments. In Section 7.3 we develop a performance based weighting scheme for moment assessments. The performance of the proposed weighting scheme is compared with that of the classical model weighting scheme of Cooke in Section 7.4. The results are summarised and discussed in the final section.

7.1 Expert Combination Properties Revisited

In Section 3.6.1 an overview is given of properties of mathematically combined expert probabilities, with discussion. In this section we will repeat these properties loosely for expectations rather than probabilities. A formal definition of these properties for the combination of Bayes linear belief structures is given in the next section. Here we consider the case in which the experts' collections of expectations are combined to form the decision maker's collection of expectations.

Marginalisation Property. The same combined expectations are found whether (a) the assessors' expectations for uncertain quantities are first combined and then expectations for all finite linear combinations of these quantities are derived, or (b) the assessors' expectations for the finite linear combinations of the quantities are first derived individually, and the resulting expectations are then combined.

In principle a marginalisation property for any combination function of the quantities can be defined. We have chosen to base the property on linear combinations here though to correspond to the linearity of the expectation operator.

Zero Preservation Property. If all assessors judge a quantity to have expectation zero, then the combined expectation also equals zero.

Strong Setwise Function Property. The combined expectation for a quantity X depends only the expectations assessed for X by the individual assessors.

Independence Preservation Property. If all assessors regard two quantities X and Y as uncorrelated then in the combined assessment X and Y are also uncorrelated. Therefore 'Zero Correlation Preservation' is the appropriate description in this context.

External Bayesian Property. The result of first combining, and then processing the results of new observations via Bayes linear adjustment is the same as first letting the experts process the results of the new observations and then combining their Bayes linear adjusted expectations.

The arguments in favour and against the desirability of some of these properties are discussed in Section 3.6.1 for the combination of probabilities. The following example, based on the flashlight example in (Cooke 1991), shows that the marginalisation property is also compelling when combining expectations.

Example 7.1. *Suppose I move into a new house and I am interested in the number of days D it will take to get broadband internet at the new address. The process of getting broadband for people in my situation consists of two distinct, consecutive phases: getting the phone line in the new house connected (Phase A) and making the connected phone line ready for broadband usage (Phase B). Two experts, George and Anna, who I esteem equally, give independently of each other 25 days as their expectation of D. When I ask about their expectation of the duration of the phases A and B (D_A and D_B), George's expectations are respectively 5 and 20 days, but Anna's expectations are resp. 20 and 5 days. Suppose the combination function I use to combine their assessments results in an expectation of 10 days for phase A and an expectation of 10 days for phase B. Deriving the total duration in this way would result in an expectation of 20 days for me. But both experts agree on their estimation of the total duration as 25, and agree that at least one of the phases already takes 20 days to finish.*

Assuming that assessors stay consistent in their assessment of $E(D) = 25$, the marginalisation property demands that the combined expectation of D remains 25 whatever phases are identified for the process.

The marginalisation property is adopted as the guiding principle in the investigation into appropriate combination functions for the combination of collections of expectations in the next section.

7.2 The Linear Pool of Bayes Linear Belief Structures

In this section we shall loosely follow the line of reasoning adopted by McConway (McConway 1981) to show that also in an expectation context, the marginalisation property together with the zero preservation property is equivalent to the strong setwise function property. Then we show that the strong setwise function property is equivalent to specifying a linear pool as the expert combination rule. While the results are similar to those in (McConway 1981), the proofs are technically slightly different and a little simpler. Will will first introduce the terminology used in this section.

7.2.1 Terminology

In the following we assume that the decision maker is interested in assessing means, variances and covariances for a collection of unknown real-valued quantities $\{X_1, \ldots, X_n\}$. We will make use of the following definitions:

$e :=$ number of expert assessors.

$B :=$ some subset of $\{X_1, \ldots, X_n\}$.

$B^{1,2} :=$ the collection of quantities consisting of the elements of B, the squares of the elements of B and the cross products of the elements of B.

$\langle B \rangle :=$ collection of all linear combinations of elements of B.

$[B] :=$ coherent belief structure on B, fully defined by B and a coherent set of expectations for the elements of $B^{1,2}$ (for a definition see second paragraph of Section 4.2).

$[B](Z) :=$ expectation of linear combination $Z \in \langle B \rangle$, derived from the expectations for the elements of $B^{1,2}$ specified for belief structure $[B]$.

$[B]^i :=$ coherent belief structure of assessor i, $i = 1, \ldots, e$.

$[B]_T :=$ restriction of a belief structure on B to T, $T \subseteq B$.

$\mathcal{B}(B) :=$ collection of all coherent belief structures for B.

$C_B([B]^1, \ldots, [B]^e) :=$ combination function that combines e coherent belief structures on B into one coherent belief structure on B; $C_B : \mathcal{B}(B)^e \rightarrow \mathcal{B}(B)$.

$\mathcal{C} :=$ a class of combination functions defined on $T^{1,2}$, with $T \subseteq \langle B \rangle$.

The $B^{1,2}$ notation is introduced to make explicit the quantities for which expectations are defined in a BL belief structure, since the pooling is performed on these expectations. Note that the subscript B in C_B is just there to indicate the domain on which the combination function C is defined. We denote the restriction of the function C_B to the subdomain T with C_T. So we have by definition that if $C_B \in \mathcal{C}$ and $T \subseteq \langle B \rangle$, then C_T is also an element of $\mathcal{C}$: $C_T \in \mathcal{C}$. In the following we further assume the expert assessors to give consistent assessments across different belief structures in the sense that if the collections $B^{1,2}$ and $T^{1,2}$ of two belief structures $[B]$ and $[T]$ share the same elements, we assume the experts to assess the same expectations for these elements for both belief structures. This means that we assume $[B]^i_T = [T]^i$ for all $T \subseteq B$, $i = 1, \ldots, e$. Further we note that coherency implies that the expected value of a constant c must equal c.

7.2.2 Definitions and Support for the Linear Pool

The marginalisation property for Bayes linear belief structures is defined as:

Definition 7.1. *BL-Marginalisation Property (BL-MP).* We take $T \subseteq \langle B \rangle$ a non-empty subcollection of finite linear combinations of elements of B. A class $\mathcal{C}$ of combination functions has the marginalisation property if and only if, for all B, all $T \subseteq \langle B \rangle$ and all $C_B \in \mathcal{C}$, there exists a combination function $C_T \in \mathcal{C}$ such that for all $Z \in T^{1,2}$,

$$[C_B([B]^1, \ldots, [B]^e)](Z) = [C_T([T]^1, \ldots, [T]^e)](Z) \tag{7.1}$$

for all $[T] \in \mathcal{B}(T)$ and for all $[B] \in \mathcal{B}(B)$.

The BL-marginalisation property thus states that the same expectations are found whether (a) the more refined individual expectations of elements of $B^{1,2}$ are first combined and then the expectations for the elements of $T^{1,2}$ are determined from these expectations, or (b) first individual expectations for the elements of $T^{1,2}$ are determined and then these expectations are combined.

Definition 7.2. *BL-Weak Setwise Function Property (BL-WSFP).* A class $\mathcal{C}$ of combination functions has the weak setwise function property if and only if for each $Z \in \langle B \rangle$ there exists a function $F_Z : \mathbb{R}^e \to \mathbb{R}$ such that for all B and all $C_B \in \mathcal{C}$,

$$[C_B([B]^1, \ldots, [B]^e)](Z) = F_Z\left([B]^1(Z), \ldots, [B]^e(Z)\right) \tag{7.2}$$

for all $[B]^1, \ldots, [B]^e \in \mathcal{B}(B)$.

In words, the BL-WSFP says that the combined expectation for the linear combination Z depends only on the individual expectations for Z of each of the assessors and possibly on the specific linear combination Z itself, e.g. on which quantities of B have a nonzero coefficient in the linear combination Z.

Theorem 7.1. *A class $\mathcal{C}$ of combination functions has the BL-WSFP if and only if it has the BL-MP.*

Proof. (I) BL-WSFP ⇒ BL-MP. Take $T \subseteq \langle B \rangle$, $Z \in T$ and $C_B, C_T \in \mathcal{C}$. Using Definition 7.2 and assuming consistency of the experts in the sense that $[B]^i(Z) = [T]^i(Z)$, for $i = 1, \ldots, e$, we find

$$\begin{aligned}[C_T([T]^1, \ldots, [T]^e)](Z) &= F_Z\left([T]^1(Z), \ldots, [T]^e(Z)\right)\\ &= F_Z\left([B]^1(Z), \ldots, [B]^e(Z)\right) = [C_B([B]^1, \ldots, [B]^e)](Z).\end{aligned}$$

for all B, for all $T \subseteq \langle B \rangle$, for all $Z \in T$, for all $[B]^1, \ldots, [B]^e \in \mathcal{B}(B)$ and all $[T]^1, \ldots [T]^e \in \mathcal{B}(T)$. So by definition, the marginalisation property holds, and in particular $[C_T([T]^1, \ldots, [T]^e)]$ and $[C_B([B]^1, \ldots, [B]^e)]$ are coherent belief structures on resp. T and B.

(II) BL-MP ⇒ BL-WSFP. Take $Z \in \langle B \rangle$, let $[Z]$ be a coherent belief structure on Z. If $\mathcal{C}$ satisfies BL-MP, we get from (7.1) that for all $C_B \in \mathcal{C}$ there is a $C_Z \in \mathcal{C}$ such that

$$[C_B([B]^1, \ldots, [B]^e)](Z) = [C_Z([Z]^1, \ldots, [Z]^e)](Z) \tag{7.3}$$

for all B, for all $Z \in \langle B \rangle$, for all $[B]^1, \ldots, [B]^e \in \mathcal{B}(B)$ and all $[Z]^1, \ldots [Z]^e \in \mathcal{B}(Z)$. The right-hand side of (7.3) depends only on Z and $[Z]^i(Z)$, for $i = 1, \ldots, e$. Therefore $\mathcal{C}$ also satisfies BL-WSFP. □

But it would be even nicer if the combined expectation of Z would not depend on the specific linear combination Z itself, e.g. whether Z is a linear combination of 2 or 2000 elements of B and which elements are or are not in the linear combination Z, but only on the expectations of each of the assessors.

Definition 7.3. *BL-Strong Setwise Function Property (BL-SSFP).* A class $\mathcal{C}$ of combination functions has the strong setwise function property if and only if there exists a function $G : \mathbb{R}^e \to \mathbb{R}$ such that for all B, all $Z \in \langle B \rangle$ and all $C_B \in \mathcal{C}$,

$$[C_B([B]^1, \ldots, [B]^e)](Z) = G\left([B]^1(Z), \ldots, [B]^e(Z)\right) \tag{7.4}$$

for all $[B]^1, \ldots, [B]^e \in \mathcal{B}(B)$.

Definition 7.4. *BL-Zero Preservation Property (BL-ZPP).* A class $\mathcal{C}$ of combination functions has the zero preservation property if and only if, for all $C_B \in \mathcal{C}$, for all B, for all $[B] \in \mathcal{B}(B)$

and all $Z \in \langle B \rangle$,

$$[B]^1(Z) = \ldots = [B]^e(Z) = 0 \qquad \Rightarrow \qquad [C_B([B]^1, \ldots, [B]^e)](Z) = 0$$

The BL-ZPP thus states that if all assessors judge a quantity to have expectation zero, then the combination of their expectation is also zero.

Theorem 7.2. *For a class $\mathcal{C}$ of combination functions the following are equivalent: (a) $\mathcal{C}$ satisfies BL-MP and BL-ZPP, and (b) $\mathcal{C}$ satisfies BL-SSFP.*

Proof. (I) BL-SSFP ⇒ BL-MP and BL-ZPP. Suppose $C_B \in \mathcal{C}$ satisfies BL-SSFP, so that there exists a G as in (7.4); then $[B]^1(Z) = \ldots = [B]^e(Z) = 0$ implies

$$[C_B([B]^1, \ldots, [B]^e)](Z) = G\,(0, \ldots, 0) \tag{7.5}$$

from (7.4). Remember that coherency implies that the expected value of a constant c must equal c. By taking the particular case $Z \equiv 0$, we see that the left-hand side of (7.5) is 0, and hence that $G(0, \ldots, 0) = 0$. So C satisfies BL-ZPP. But BL-SSFP implies BL-WSFP by definition, and BL-WSFP implies BL-MP by Theorem 7.1. So BL-SSFP implies BL-MP and BL-ZPP.

(II) BL-MP and BL-ZPP ⇒ BL-SSFP.

Suppose $C_B \in \mathcal{C}$ satisfies BL-MP and BL-ZPP. By Theorem 7.1, C_B then satisfies BL-WSFP, so there exists $F_Z : \mathbb{R}^e \to \mathbb{R}$ such that (7.2) holds. Since C_B satisfies BL-ZPP, we also have

$$F_Z(0, \ldots, 0) = 0 \tag{7.6}$$

for each $Z \in \langle B \rangle$.

Now suppose $X, Y \in \langle B \rangle$. By linearity of coherent expectations (see (2.1) in Section 2.3.1) we have,

$$[C_B([B]^1, \ldots, [B]^e)](X + Y) = [C_B([B]^1, \ldots, [B]^e)](X) + [C_B([B]^1, \ldots, [B]^e)](Y) \tag{7.7}$$

It follows that,

$$F_X(x^1, \ldots, x^e) + F_Y(y^1, \ldots, y^e) = F_{X+Y}(x^1 + y^1, \ldots, x^e + y^e)$$

Similarly, taking $X + (Y - X) = Y$, we have

$$F_X(x^1, \ldots, x^e) + F_{Y-X}(y^1 - x^1, \ldots, y^e - x^e) = F_Y(y^1, \ldots, y^e) \tag{7.8}$$

Hence, using (7.6) and (7.8) we see that,

$$F_X(x^1, \ldots, x^e) = F_X(x^1, \ldots, x^e) + F_{Y-X}(0, \ldots, 0) = F_Y(x^1, \ldots, x^e).$$

We find $F_X \equiv F_Y$, so F_X does not depend on X and therefore BL-SSFP holds. □

Theorem 7.3. *For a class $\mathcal{C}$ of combination functions the following are equivalent:*

I *$\mathcal{C}$ satisfies BL-SSFP*

II *For each $C_B \in \mathcal{C}$ there exist real numbers $\alpha_1, \ldots, \alpha_e$, nonnegative and summing to 1, such that for all B, all $Z \in \langle B \rangle$ and all $[B] \in \mathcal{B}(B)$,* $[C_B([B]^1, \ldots, [B]^e)](Z) = \sum_{i=1}^{e} \alpha_i [B]^i(Z).$

Proof. (II) ⇒ (I). The combined expectation for $Z \in \langle B \rangle$ (left-hand side) depends only on the expectations assessed by each of the individual assessors (right-hand side), so BL-SSFP is satisfied. It remains to show that the linear pool gives a coherent belief structure. This is shown separately in Theorem 7.4 below.

(I) ⇒ (II). Suppose $\mathcal{C}$ satisfies BL-SSFP. Then there exists a function G as in (7.4). From (7.7) it follows that,

$$G(x^1, \ldots, x^e) + G(y^1, \ldots, y^e) = G(x^1 + y^1, \ldots, x^e + y^e) \tag{7.9}$$

Repeatedly using (7.9), we have for any $z^1, \ldots, z^e \in \mathbb{R}$,

$$\begin{aligned} G(z^1, \ldots, z^e) &= G(z^1, 0, \ldots, 0) + G(0, z^2, \ldots, z^e) \\ &= G(z^1, 0, \ldots, 0) + \ldots + G(0, \ldots, 0, z^{e-1}, 0) + G(0, \ldots, 0, z^e) \end{aligned}$$

So if we define $G_i(z) = G(\overbrace{0, \ldots, 0}^{i-1}, z_i, \overbrace{0, \ldots, 0}^{e-i})$ for $i = 1, \ldots, e$, we can write $G(z^1, \ldots, z^e) = \sum_{i=1}^{e} G_i(z)$. Also (7.9) implies that $G_i(x+y) = G_i(x) + G_i(y)$ for $x, y \in \mathbb{R}$, so each G_i satisfies Cauchy's functional equation (Aczél 1966, Sec. 2.1.1). Further we see that

$$\lim_{\triangle x \to 0} G_i(x + \triangle x) = \lim_{\triangle x \to 0} G_i(x) + G_i(\triangle x)$$

Note that $\triangle x$ is a constant and recall again that coherency implies that the expected value of a constant C must equal C. We find that

$$\lim_{\triangle x \to 0} G_i(x + \triangle x) = \lim_{\triangle x \to 0} G_i(x) + G_i(\triangle x) = \lim_{\triangle x \to 0} G_i(x) + \triangle x = G_i(x).$$

Hence $G_i(x)$ is continuous and we find that $G_i(z) = \alpha_i z$, where α_i is a constant.

We complete the proof by examining the case in which z is a constant and all experts agree that $z \equiv C \neq 0$. We find by $G(C, \ldots, C) = C = \sum_{i=1}^{e} G_i(C) = \sum_{i=1}^{e} \alpha_i C$ that $\sum_{i=1}^{e} \alpha_i = 1$. □

As mentioned in the introduction of this chapter we call the set of expectations obtained by taking the linear pool of expectations of experts, the *decision maker (DM) opinion.* We now give the remaining promised result, that if the experts have all given a coherent set of expectations then the DM using a linear pool will automatically do so. In particular, if each of the experts have provided a vector of expectations and a non-negative definite covariance matrix, then that derived for the DM should be so too.

Theorem 7.4. *If a set of experts each provides a coherent set of expectations then the decision maker based on a linear pool of these sets of expectations is also coherent.*

Proof. This follows immediately from the convex hull interpretation of coherency (see last paragraph of Section 2.3.1). If each expert's vector of expectations is coherent, then each of these vectors is in the closed convex hull of the realm of the assessed quantities. Since an affine linear combination of these vectors remains in this convex hull, the linear pool must also be coherent. □

Finally, we note that the linear pool does not have the zero correlation preservation and externally Bayesian properties. This can be easily shown by the following example.

Example 7.2. Consider the case where two experts are consulted for their expectations for the uncertain quantities X, Y and XY. Suppose Expert 1 assesses these as $E_1(X) = E_1(Y) = E_1(XY) = 0$ and Expert 2 as $E_2(X) = E_2(Y) = E_2(XY) = 1$. So both experts assess X and Y to be uncorrelated, since for both the covariance $Cov(X, Y) = E(XY) - E(X)E(Y) = 0$. If we combine these assessments by using a linear pool with a weight of $\frac{1}{2}$ for both experts, we find the DM expectations to be $E_{DM}(X) = E_{DM}(Y) = 0.5 = E_{DM}(XY)$, resulting in a covariance of the DM of 0.25. Hence we find that the linear pool does not posses the zero correlation preservation property.

Staying with this example, suppose now the realisation of Y is observed. By looking at the adjustment rule for the expectation for X given Y, $E_Y(X)$,

$$E(X) + Cov(X,Y)[Var(Y)]^{-1}(Y - E(Y))$$

we see that the adjusted expectation for each of the experts does not change, where it does change for the DM. Therefore the linear pool is not externally Bayesian.

We now proceed to discuss how the weights for experts might be chosen.

7.3 Performance Based Weighting

We speak of performance based weighting when weights are derived from the performance on seed variables. The performance on these seed variables is then applied to derive a combined assessment for the other variables. In order to gain confidence in the assessments on the other variables, the seed variables should closely match them. The classical model of Cooke (Cooke 1991), described in Section 3.6.2, is an implementation of the linear pool that uses performance based weighting to combine probability assessments. In this section we develop a performance based weighting scheme for moment assessments analogous to that of the classical model. The concepts of calibration and information used in the classical model weighting scheme do not translate directly into a Bayes linear context, but we shall show that it is possible to build a scoring rule that possesses some similar properties.

The basis for a scoring rule can be obtained from the observation that my expected value of a quantity X is defined in a moment context as the value $E(X)$ I would choose if I were to take part in a lottery with a small monetary penalty of $c(x - E(X))^2$ when the realisation of X becomes known and equals x. Hence when a number of expected values are being assessed for seed variables $X_1, \ldots, X_n$, we can define a penalty function for an expert giving assessed expectations $a_1, \ldots, a_n$ for $X_1, \ldots, X_n$, by

$$\phi(a_1, \ldots, a_n) = \sum_{j=1}^{n} c_j (x_j - a_j)^2, \tag{7.10}$$

where x_j is the realisation of variable X_j (unknown to the expert at the time of assessment), and the quantities c_j are positive scaling variables bringing each quantity to a common monetary scale. Using ϕ as the basis for weighting is plausible as the expert will minimize future expected loss ϕ by assessing his expected values - remember from the definition of expectation, De Finetti's second criterion of prevision given in Section 2.3.1, that the expected value minimizes the expected quadratic loss - as is appropriate. We shall return to a discussion of the choice of scaling quantities c_j later. For the moment we note that ϕ is a loss function, that is, we should have more confidence in an expert with a smaller ϕ.

7.3.1 The Moment Model Weighting Scheme

A weighting scheme that uses the ϕ function can be defined as follows. First we compute the ϕ value for each expert, giving values $\phi_1, \ldots, \phi_e$. To be able to exclude bad performing experts from the linear pool we introduce a cut-off value $\alpha > 0$ (at the moment this is arbitrary, but a specific choice will be made shortly). Any expert with loss $\phi \geqslant \alpha$ will be given weight 0. The unnormalised weight w'_i for expert i is the difference between the cut-off and the loss,

$$w'_i = (\alpha - \phi_i) \cdot 1_{\alpha > \phi}, \tag{7.11}$$

where the indicator function $1_{\alpha > \phi}$ is there to assign weight 0 to experts that have a loss ϕ greater than or equal to α. The performance based expert weights w_i are subsequently derived by normalising the weights w'_i:

$$w_i = \frac{w'_i}{\sum_i w'_i} \tag{7.12}$$

The cut-off value α is chosen within the interval $(\min(\phi_1, \ldots, \phi_e), \infty)$ by optimisation so that the loss of the combined expert is minimized. Note that the two extreme cases for α in this interval give:

$\alpha \searrow \min(\phi_1, \ldots, \phi_e)$: All the weight is given to the expert with the smallest loss, or equally to those experts with smallest loss in case of a tie.

$\alpha \nearrow \infty$: Equal weight is given to each expert.

7.3.2 Weighting Scheme Properties

Clearly there is some degree of arbitrariness in the scheme proposed. However, two important properties in its favour should be mentioned. The first is already clear from the discussion above: if an expert wishes to choose a set of expected values so as to maximize his unnormalised score, then he should simply use the same assessment procedure he would have used if he were to minimize quadratic loss (thereby providing his genuine expectation) for each quantity individually.

The second property relates to the choice of scaling variables. Clearly the choice of these variables is important in determining the overall loss ϕ. However, only relative values of the scaling variables are important:

Theorem 7.5. *If ϕ is a loss function defined as above, then for any constant $c > 0$, the function $c\phi$ is also a loss function, and the normalized weights obtained from the two loss functions are identical.*

This follows easily from the observation that if α is an optimal cut-off value for ϕ then $c\alpha$ is an optimal cut-off value for $c\phi$, so that upon normalization the constant c will disappear.

It is worth pausing to consider in what way calibration and information are reflected in this weighting scheme. Cooke requires the experts to use the same quantiles for all seed variables in order that there is a common "scale" defined by the probability bins (usually 0-5%, 5% - 50% , 50% - 95% , 95% -100%). Thus calibration can be defined in terms of quality of assessing likelihood of quantile bins over all the seed variables. In the scheme we are proposing, the common scale is defined by a choice of scaling variables, and the calibration is performed directly in terms of quadratic loss. Assuming that variances of quantities are being assessed, the overall loss ϕ contains also terms such as $(x^2 - E(X^2))^2$ which take account of the spread of the variables. Hence uncertainty about the spread is taken into account, though in a different way to Cooke's information function (which rewards low spread per se).

7.3.3 Expert Weighting Properties and Geometric Interpretation

We begin by discussing some theoretical properties of the weighting scheme and then show some examples and give the geometric interpretation.

Theorem 7.6. *The weighting scheme has the following properties:*

1. *The unnormalised weight of an expert is a proper scoring rule.*

2. *The expert with the smallest loss always remains in the pool.*

3. *The DM loss is always smaller than, or equal to the loss of any individual expert.*

4. *The DM loss is always smaller than, or equal to the loss obtained when using equal weights for all experts.*

5. *The weighting scheme defines a continuous mapping from a vector of expert losses to a vector of expert weights.*

Proof. (1) It is well known that the expected value minimizes quadratic loss. Hence for the expert to minimize his overall loss he should give his own mean value of each quantity. The score remains proper with the use of cut-off α. (2) This follows directly from the definition of the weighting. (3) and (4). These follow from the fact that they correspond to the two cases $\alpha = \min(\phi_1, \ldots, \phi_e)$ and $\alpha = \infty$ as noted before. (5) Continuity is obvious when all the expert losses differ from the cut-off value. For an expert i who does not have the smallest loss, as $\phi_i \nearrow \alpha$ we have $w_i \searrow 0$, and $w_i = 0$ for $\phi_i \geq \alpha$. For an expert with the smallest loss, continuity at $\phi_i = \alpha$ is not an issue as this is outside the allowed range of α. Hence the weighting scheme defines a continuous mapping from a vector of expert scores to a vector of expert weights. □

To understand the geometry of the weighting scheme we first observe that the quadratic scoring rule is related to a Euclidean norm (just the usual norm, but scaled by our scaling constants). Hence the loss ϕ is in fact the squared distance (using this norm) from the realisation vector to the expert's vector of expectations. We now look at some simple examples to gain insight in the behaviour of the scoring rule.

Example 7.3. Suppose we have four expert assessors, who are asked to state their expectations for two uncertain quantities X and Y. Let $\underline{e_i} = (E_i(X), E_i(Y))$ be the vector of expectations of X and Y of Expert i. Suppose the experts would give the following coherent assessments:

Expert assessments			
$\underline{e}_1$	$(4, 1)$	$\underline{e}_3$	$(4, 4)$
$\underline{e}_2$	$(2, 3)$	$\underline{e}_4$	$(7, 3)$

In Figure 7.1 these assessments are plotted on the (X, Y)-plane. The line drawn around the assessments in this figure is in fact the convex hull of the experts' expectations. Because each of the experts has given coherent expectations, all the points (x, y) inside this convex hull also represent coherent expectations for (X, Y), and hence a decision maker that uses a linear pool of the experts' assessments will also be coherent (Theorem 7.4).

Suppose we choose to score the assessments with the penalty function (**??**) with scales $c_X = c_Y = 1$ and that after the experts have given their expectations, we observe realisation $\underline{r}_1 = (4.75, 2)$. Then we can compute the normalised weights w_1, w_2, w_3 and w_4 for the experts and the DM expectations $(E_{DM}(X), E_{DM}(Y))$ based on the linear pool using these weights, for all possible cut-off values α. We choose α to be greater than the minimum penalty score $\phi_{min} = min(\phi_1, \ldots \phi_4)$ of the experts, thereby leaving at least one (the best) expert in the pool. The dashed line in Figure 7.1 represents the range of values of the DM expectation vector for $\alpha \in (\phi_{min}, \infty)$. So the graphical interpretation of choosing the optimal cut-off α, i.e. the

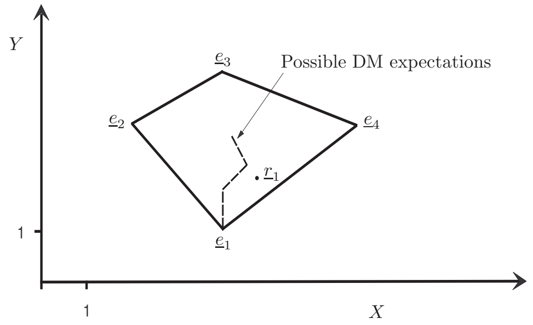

Figure 7.1: Realisation in convex hull.

α that minimizes the DM penalty, is simply choosing the point on the dashed line that is the closest to the vector of realisations.

Example 7.4. We take the same assessments and scales as in the previous example, yet now we take a look at what happens when we observe a realisation outside of the convex hull of the experts' assessments, $\underline{r}_2 = (10, 2)$.

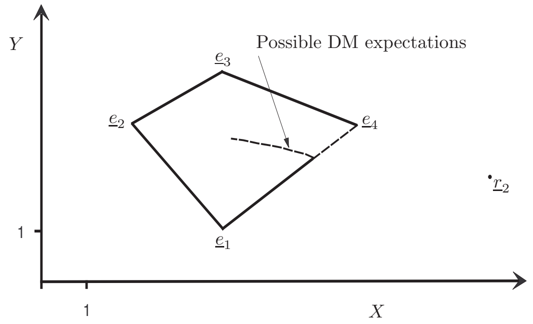

Figure 7.2: One expert selected.

The dashed line of possible DM values (Figure 7.2) starts when an infinitely great cut-off value α is evaluated, leading to equal normalised weights $w_i = \frac{1}{4}$ for all experts and the corresponding vector $(4.25, 2.75)$ for the DM. With α decreasing the scores of the experts become relevant and the DM vector starts to move in the direction of the better experts, all the way until it meets the best expert, Expert 4. Expert 4's assessment is in this case also the point on the line of possible DM assessments that is the closest to the realisation.

Example 7.5. In this example we evaluate the case where we have three experts, of which two have equal but both relatively poor performance. Suppose we have the following coherent expectations:

Expert assessments	
$\underline{e}_4$	$(1, 1)$
$\underline{e}_5$	$(11, 3)$
$\underline{e}_6$	$(21, 1)$

Now, after receiving these expectations we observe $\underline{r}_3 = (11, 0.9)$. The DM score is now optimised by taking cut-off α as large as possible, leading to equal weights $w_i = \frac{1}{3}$ for all experts (Figure 7.3). In fact, it does not matter how poor experts 4 and 6 judgements of $E(X)$ are, as long as these have about the same norm distance to the realisation x_{r_3} and are in opposite direction of x_{r_3}, we will get the same expectations for the decision maker, after optimisation over α.

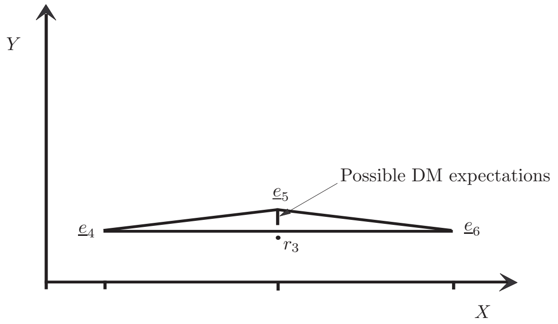

Figure 7.3: Experts with relatively big losses selected.

We will now turn to Cooke's principles for the use of expert judgement, discussed in Section 3.4.1, and show that the linear pool with the performance based weighting scheme as described in Section 7.3.1 can comply with these principles. All quotes in the remaining part of this section are taken from (Cooke 1991).

The principle of reproducibility requires the possibility "for scientific peers to review and if necessary reproduce all calculations". This property can be fulfilled by specifying the weighting scheme explicitly and making the experts' assessments accessible. When also the source of the assessments is specified explicitly, accountability is achieved. By using performance based weights, the principle of empirical control is satisfied. The neutrality principle requires that "the method for combining/evaluating expert opinion should encourage experts to state their true opinions", which is exactly what is pursued by using an unnormalised score with the strictly proper scoring rule property. As all experts are treated equally, also the principle of fairness is fulfilled.

Finally, we will apply the moment model introduced in this paper on expert judgement data gathered in an application of Cooke's classical model as an illustrative example and compare the results with those obtained when using the classical model.

7.4 Moment Model vs. Classical Model

The moment and the classical model have been compared on data from five actual applications of the classical model: prime rent assessments, dikering safety assessments, thermal comfort in buildings assessments, radionuclide transport in soil assessments and atmospheric deposition assessments. Detailed results of the comparison on these five applications can be found in Appendix C. Before discussing the results, we will first describe the method of comparison and discuss one of the cases in more depth as an illustrative example.

7.4.1 Method of Comparison

To compare the weighting schemes five applications with a relatively high amount of seed variables were selected from applications of the classical model. For each case the first half of the seed variables was used for deriving performance based weights in both the moment and classical model. The second half of the variables was used to test the performance of the linear pools using both these sets of weights. These linear pools are in the following referred to as M_{DM} for the moment weighting scheme and C_{DM} for the classical scheme. The performance of both these linear pools for the second half of the seed variables was evaluated through computing for each pool the scores under both weighting schemes: the penalty ϕ (moment model) and the product of calibration and information (classical model). The calculation of the calibration and information scores is described in Section 3.6.2.

The original data were expert assessments for the 5%-, 50%- and 95%-quantiles for the seed variables. To be able to use the moment scheme, these assessments needed to be translated into assessments of moments of the variables. We are interested in the experts' abilities to assess means and variances, therefore assessments for first and second moments are sufficient. As translation method the extended Pearson-Tukey (EP-T) method described in Section 3.5.1.1 is used.

7.4.2 Illustrative Case: Prime Rent Assessments

The prime rent case involved 5 investments managers who each gave assessments on the rent indices of office space for the major cities in the Netherlands for the future (Qing 2002). Apart from the 16 original seed variables, realisations for 15 more variables were observed post hoc. Assessments for the 16 original seed variables have been used to derive the performance based linear pools for both weighting schemes: M_{DM} and C_{DM}. Secondly, the performance of these linear pools was evaluated on the remaining 15 variables.

To be able to calculate the moment model score for the assessments of each of the experts we need to determine appropriate values for the coefficients c_j of the score (7.10) first. Since all variables assessed by the experts are of the same scale (prime rents in Dutch Guilders per m^2), the errors made by the assessors are as well. We therefore use the same coefficients for all seed variables. Yet, since we score both the assessment of means and variances via the first and second moments, we need to choose a coefficient c_1 that brings first moment errors to a monetary scale and a coefficient c_2 that brings assessment errors of second moments to a monetary scale. The penalty function ϕ_1 used thus becomes:

$$\phi_1(a_1,\ldots,a_n,b_1,\ldots,b_n) = \sum_{j=1}^{n} c_1(x_j - a_j)^2 + \sum_{j=1}^{n} c_2((x_j)^2 - b_j)^2, \tag{7.13}$$

where a_j and b_j are the derived assessments for an expert of resp. the first and second moment of variable X_j, x_j the observed realisation of X_j and n the number of variables that are evaluated in the score. The first summation in (7.13) penalises deviations of first moment assessments from the realisations, the second summation penalises deviations of second moment assessments from the square of the realisations. When determining an appropriate value for c_1 to bring assessment errors on first moments to a monetary scale, we need to make sure that the penalty is large enough for the assessors to matter, but small enough to avoid risk averseness of the assessors against large money losses. For this purpose a value of 1 Euro cent, entailing total penalty scores in the order of tens to hundreds of Euros, was deemed appropriate. In the default case the value for c_2 was chosen such that both summations in (7.13) are on average equal for all experts, on average resulting in an equal total penalty for first moment assessments and second moment assessments. Or, more formally, choosing c_2 such that $r_2 = 0.5$ in

$$r_2 = \frac{\sum_{j=1}^{n} c_2((x_j)^2 - b_j)^2}{\phi_1}. \tag{7.14}$$

The results of the comparison for $r_2 = 0.5$ are shown Table 7.1. The optimised performance based weights are very different for both methods. The classical model assigns in this case weight 1 to its best performing expert (i.e. the expert with the highest product of calibration and information). The moment model on the other hand finds a very diverse pool of experts to

Table 7.1: Comparison of MM and CM for Prime Rent data ($\boldsymbol{r_2 = 0.5}$)

	M_{DM}	C_{DM}	Eq. weights	Exp. 1	Exp. 2	Exp. 3	Exp. 4	Exp. 5
Weights								
Expert 1	0.143	1	0.2	1	0	0	0	0
Expert 2	0.172	0	0.2	0	1	0	0	0
Expert 3	0	0	0.2	0	0	1	0	0
Expert 4	0.537	0	0.2	0	0	0	1	0
Expert 5	0.148	0	0.2	0	0	0	0	1
Moment model score								
16 original seed variables								
penalty ϕ_1	*5257*	*18001*	*12518*	*18001*	*17262*	*116911*	*8157*	*17863*
15 performance variables								
penalty ϕ_1	**2756**	**2887**	**4680**	**2887**	**15261**	**34616**	**7006**	**9448**
Classical model score								
16 original seed variables								
calibration	0.3053	0.3305	0,0561	0.3305	0.1472	0.0201	0.0001	0.0042
information	0.5099	0.8572	0.1790	0.8572	0.9554	0.1556	1.5357	0.6126
score	*0.1557*	*0.2833*	*0.0100*	*0.2833*	*0.1407*	*0.0031*	*0.0001*	*0.0026*
15 performance variables								
calibration	0.2880	0.3579	0.1824	0.3579	0.0000	0.0006	0.0390	0.0390
information	0.5026	0.6724	0.1674	0.6724	0.7465	0.1641	1.3837	0.9623
score	**0.1448**	**0.2406**	**0.0305**	**0.2406**	**0.0000**	**0.0001**	**0.054**0	**0.0375**

M_{DM}, C_{DM}: performance based linear pools using resp. moment and classical model.

be the optimal choice, based on the experts' performance on the 16 seed variables. If we look at the performance of the linear pools C_{DM} and M_{DM} on the additional 15 variables for which the realisations are known, we find a mixed picture. The M_{DM} slightly outperforms C_{DM} when the penalty ϕ_1 is considered (2756 (Euro cents) versus 2887 (Euro cents)), but the C_{DM} has a better score in the classical model (0.1448 versus 0.2406). Note here that the lower the penalty score, the better the performance, whilst the opposite holds for the score in the classical model. Both performance based linear pools perform better than the linear pool using equal weights.

The results in Table 7.1 are based on $r_2 = 0.5$. However, when r_2 is varied the weights of M_{DM} change and also the results change. Figure 7.4 gives an impression of how the weights change for different values of r_2. The higher the value of r_2, i.e. the more important the assessments of the second moments become, the lower the weight assigned to expert 1, eventually resulting in weight zero for high values of r_2. Indeed the penalty score for the derived second moments assessments for expert 1 is 26% worse than the worst penalty score of the experts 2, 4 and 5.

In Table 7.2 the results for M_{DM} and C_{DM} are summarized for $r_2 = 0.1$, $r_2 = 0.5$ and $r_2 = 0.9$. The results for $r_2 = 0.1$ are consistent with those for $r_2 = 0.5$, only the differences are greater. The weights are even more spread over the experts for M_{DM} (see Figure 7.4). The M_{DM} pool now more distinctly outperforms the C_{DM} pool on the penalty score ϕ_1, while the opposite holds for the unnormalised weight of the classical model. When r_2 is set to 0.9, a different picture is obtained. Only the experts 2, 4 and 5 have a non-zero weight in M_{DM}. Expert 1, who is assigned weight 1 in the classical model, is now assigned a zero weight in the moment model. C_{DM} now has a slightly smaller penalty score, whilst M_{DM} has a slightly better score in the classical model.

The prime rent case explored in this section thus shows that the moment weighting scheme can lead to very different weights than Cooke's classical model. Neither method however had a strictly better performance. Both performance based weighting schemes outperformed the equal weights linear pool.

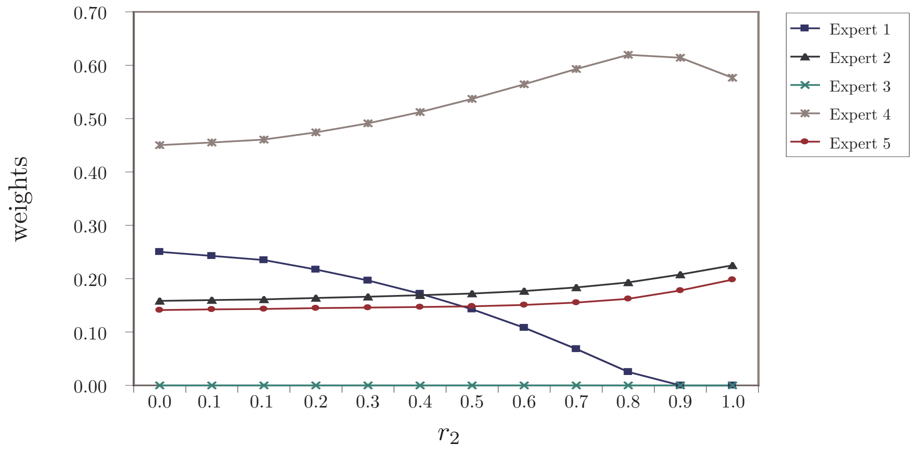

Figure 7.4: M_{DM} weights for $\boldsymbol{r_2}$, prime rent data

Table 7.2: Comparison of MM and CM for Prime Rent data ($\boldsymbol{r_2 = 0.1, 0.5}$ and $\mathbf{0.9}$)

	$r_2 = 0.1$		$r_2 = 0.5$		$r_2 = 0.9$	
	M_{DM}	C_{DM}	M_{DM}	C_{DM}	M_{DM}	C_{DM}
Moment model score						
16 original seed variables						
penalty ϕ_1	*2931*	*8887*	*5257*	*18001*	*24705*	*100667*
15 performance variables						
penalty ϕ_1	**1749**	**2009**	**2756**	**2887**	**11686**	**10851**
Classical model score						
16 original seed variables						
calibration	0.3053	0.3305	0.3053	0.3305	0.3378	0.3305
information	0.4972	0.8572	0.5099	0.8572	0.5688	0.8572
score	*0.1518*	*0.2833*	*0.1557*	*0.2833*	*0.1921*	*0.2833*
15 performance variables						
calibration	0.1824	0.3579	0.2880	0.3579	0.4314	0.3579
information	0.4801	0.6724	0.5026	0.6724	0.5747	0.6724
score	**0.0875**	**0.2406**	**0.1448**	**0.2406**	**0.2479**	**0.2406**

7.4.3 Comparison on Five Applications

A similar analysis has been applied to the other four applications. The results are summarized in Table 7.3. Only for two out of the five cases different choices of r_2 lead to different performance based weights in the moment model. In none of the cases both models select the same experts

Table 7.3: Comparison of the classical model and moment model on 5 applications

	$r_2 = 0.1$	$r_2 = 0.5$	$r_2 = 0.9$
M_{DM} scores better than C_{DM} on penalty ϕ	3/5	3/5	2/5
M_{DM} scores better than C_{DM} on cal × info	2/5	2/5	3/5
Both models select same experts	0/5	0/5	0/5
Both models agree on best expert	0/5	0/5	0/5
Both models agree on ranking experts	0/5	0/5	0/5
M_{DM} at least as good as equal weights on ϕ	3/5	4/5	4/5
C_{DM} at least as good as equal weights on ϕ	3/5	3/5	3/5
M_{DM} at least as good as equal weights on cal × info	4/5	5/5	4/5
C_{DM} at least as good as equal weights on cal × info	4/5	4/5	4/5
M_{DM} at least as good as best expert on ϕ	4/5	4/5	4/5
C_{DM} at least as good as best expert on ϕ	3/5	3/5	3/5
M_{DM} at least as good as best expert on cal × info	3/5	3/5	4/5
C_{DM} at least as good as best expert on cal × info	4/5	4/5	4/5

M_{DM}, C_{DM}: performance based linear pools using resp. moment and classical model.

for their linear opinion pool (i.e. assign the same set of experts a non-zero weight). Even, in none of the cases the moment and the classical model agree on which expert has the best performance on the seed variables.

The remainder of the results will be discussed for $r_2 = 0.5$. The results are only slightly different for different values of r_2, as can be seen in Table 7.3.

When loss function ϕ_1 is used to evaluate the performance of the linear pools M_{DM} and C_{DM}, we find that the moment model DM, M_{DM}, has a better score on the performance variables than the classical model DM, C_{DM}, in 3 out of the 5 cases. In 4 out of 5 cases M_{DM} performs at least as good as the equal weight linear pool and as when choosing the best expert as DM. In 3 out of 5 cases C_{DM} performs at least as good as the equal weight linear pool and as when choosing the best expert as DM.

When the classical model score is used as performance measure, C_{DM} has a better score than M_{DM} in 3 out of 5 cases. C_{DM} performs at least as good as the equal weights linear pool in 4 out of 5 cases, against 5 out of 5 cases for M_{DM}. Finally, in 4 out of 5 cases C_{DM} scores at least as good on the performance variables as when choosing the best expert as DM, against 3 out of 5 cases for the moment model DM.

7.5 Summary and Conclusions

In this chapter we have developed a performance based aggregation method for moment assessments from first principles. We have translated the most predominantly discussed desirable properties mathematical aggregation methods of probabilistic assessments can have into the context of moment methods. We have shown that also when expectation is taken as the primitive, that the requirement for the aggregation method to possess the marginalisation and zero preservation properties is equivalent to requiring the aggregation method to be a linear pool of the experts' assessments. Note that this result has been derived here in an expectation framework. The result is analogous to McConway's argumentation for the support of using a linear pool of expert probability distributions (McConway 1981), is consistent with McConway's result since a linear combination of probability distributions entails a linear combination of expectations as well, but does not follow from his result.

The linear pool of experts' expectations has the nice property that when all experts have provided a coherent set of assessments, any linear pool of these assessments also constructs a coherent set of assessments. The meaning of the pooled assessment through a combination function is questioned by Garthwaite et al. (2005): "Another criticism of all these pooling methods is that it is not clear whose opinion (if anyones) the resulting probability distribution represents". Yet we think it is clear that in many cases the decision maker is genuinely unable to specify a prior distribution and is genuinely unable to choose between a number of technical experts with excellent qualifications and different opinions. Rather than choosing uninformative priors as a solution, or forcing the decision maker into choices that he or she feels uncomfortable with, we believe that the combined expert represents a synthetic but plausible prior that the decision maker could more fairly adopt than other alternatives that are available.

In Section 7.3 we have developed a weighting scheme that is based on the performance of the experts. This weighting scheme is similar to Cooke's classical model weighting scheme discussed in Section 3.6.2, but has certain theoretical advantages over it. Firstly, the moment model weighting scheme is much simpler and derived from first principles: the basis of the performance assessment is a quadratic scoring rule which also constitutes the foundation of De Finetti's definition of an expectation (see Section 2.3.1). In the classical model, the 'weighing' of the calibration and information scores in the overall score, the unnormalised weight ((3.4) in Section 3.6.2), is fixed but arbitrary. In the moment model the concepts of calibration and information are both accounted for in one score, the loss: a smaller loss corresponds to both better calibration and more informativeness. It must be noted though that the moment model does not prescribe how the losses for the different quantities and for different moments of the same quantity should be 'weighed' in the overall loss ϕ. Secondly, the unnormalised weight of the moment model is a strictly proper scoring rule, where the unnormalised weight of the classical model is only weakly asymptotic strictly proper.

A third advantage of the moment model over the classical model is that, given true values of the seed quantities, the moment model weighting scheme forms a continuous mapping from the experts' assessments to a vector of expert weights. In the classical model there can be a huge difference in weights for experts that have given very similar assessments, depending on whether probability bins are just hit or just missed by the realisations of the seed variables. Fourthly, the moment model also does not require an arbitrary choice of bounds for the seed variables, as the classical model does. Finally, but maybe most important, the moment model enables us to also evaluate and score the performance of the experts in assessing dependencies between quantities as well.

Like the classical model the moment model always keeps the best performing expert in the pool, but the moment model linear pool is certain to have at least as good a performance on the seed quantities as the best expert and equal weights linear pool. The results in Section 7.4 do not give evidence that either of the weighting schemes has a better performance than the other. Where classical model scheme regularly has only the best expert with nonzero weight, the weights seem more spread over experts for the moment model.

For all the reasons summarised in this section, we therefore recommend the moment model to aggregate sets of expert moments.

Chapter 8

Conclusions

In this final chapter we will summarise the results reported in this thesis. We will discuss the answers formulated to each research question, describe the limitations of these answers and give suggestions for further research. Finally we will discuss the implications of the results of this research for Defence.

8.1 Derivation of Moments from Expert Assessments

We propose the use of the bivariate Pearson-Tukey method to derive the assessments of the moments needed to quantify Bayes linear models, for two reasons. Firstly, this method has a good performance. Previous research has shown that with the univariate Pearson-Tukey method means and variances can be approximated very well for a wide variety of distributions. In fact, the method performed best in all publications we have found on this topic. We refer to Section 3.5.1.1 for details and the references to this research.

In Section 6.3 we have evaluated the bivariate extension of this method, for bivariate Normal, F-I Beta, F and Cheriyan distributions. The maximum average absolute error found for these distributions is 6.1% for marginal moments of up to the fourth order, and 3.2% for product moments to this order. We found that the errors for the marginal moments correlate strongly with the skewness of the variable, where larger errors are found for stronger skewed distributions. It would be a valuable topic for further research to investigate whether adjustments to the bivariate Pearson-Tukey method can be proposed that would reduce the errors found for more strongly skewed distributions.

The results found for univariate Beta distributions correspond to the results reported in a previous study when only Beta distributions with a skewness smaller than 3 are considered, although the maximum errors we encountered are an order of magnitude larger.

The second important reason for our preference for the bivariate Pearson-Tukey method is due to the military context wherein this research is placed, and wherein simplicity and robustness are required. This method only requires marginal and conditional quantiles as input from experts, which form in our opinion much less complicated assessments to make than some of the alternative approaches require that have been investigated in the literature (discussed in Section 3.5.2).

The results reported in this thesis are based on large samples drawn from bivariate Normal, F-I Beta, F and Cheriyan distributions. Although with these samples a wide variety of bivariate distributions is covered, the result reported in this research do not necessarily generalise to other bivariate distributions.

8.2 The Accuracy of the Bayes Linear Adjustment Rules

The Bayes linear adjustment of the mean and variance is exact for multivariate Normal distributions, i.e. the adjusted mean and variance are equal to the conditional mean and variance for this distribution. We have evaluated the accuracy of the adjustment rules for bivariate F-I Beta, F, Kibble and Cheriyan distributions. These distributions all have a conditional mean that is linear in the condition, which entails that the adjusted and conditional mean are equal for these distributions (see Section 5.1). We evaluated the accuracy of the adjusted variance analytically and by calculating the difference with the conditional variance for large samples from the four distributions. These samples cover a wide variety of bivariate distributions common in practice. It should be noted however that the results reported in this thesis do not necessarily generalise to other bivariate distributions.

The adjusted variance is not a close approximation of the conditional variance for the F-I Beta, F and Kibble distribution. The average absolute error found over the $5\% - 95\%$ interquantile range of the condition, in a sample of $10,000$ cases of each of these distributions, is 75%, 19% and 27% respectively. When the condition is exactly as expected and equal to its mean, the errors are smaller with respectively 5%, 2% for the F-I Beta and F distribution and zero for the Kibble distribution. For F-I Beta distributions the adjusted variance overestimates the conditional variance on average with 59%. For F, Kibble and Cheriyan distributions the average biases are much smaller with resp. 1%, 0.07% and 0.04%.

For Cheriyan distributions the adjusted variance is a much better approximation of the conditional variance. The average absolute error found in the sample of $10,000$ Cheriyan distributions is 2.2%, and the error for the condition equal to its expectation is on average 0.06%. Where the adjusted variance is constant, i.e. does not depend on the value of observations, the conditional variances of the four distributions investigated are not. The Cheriyan distribution however has a far more constant conditional variance than the other three distributions on the $5\% - 95\%$ interquantile range of the condition: the minimum value of the conditional variance is on average 92% of the maximum conditional variance on the $5\% - 95\%$ interquantile range of the condition. For the other distributions the ratios of the minimum and maximum value of the conditional variance are much lower with 39%, 52% and 42% for resp. the F-I Beta, F and Cheriyan distribution.

Practitioners we would therefore recommend not to use the adjusted variance as an approximation the conditional variance in general. Two exceptions however are in place to this recommendation: the adjusted variance might be a relative good approximation when the correlation is very small and for distributions for which the conditional variance is considered to be relatively constant, as is the case for the bivariate Normal and Cheriyan distribution.

8.3 The Benefits of Using Higher Order Information

In Chapter 6 we have shown that the conditional variance can be approximated extremely well by calculating the variance from the adjusted first and second moment and using fourth order (product) moments in the adjustment. We refer to this approximation as the (Bayes linear) 2-adjusted moment variance. The good results are found for all four distributions evaluated: the F-I Beta, F, Kibble and Cheriyan distributions. The highest average absolute error found is of order $10^{-5}\%$ for the Cheriyan distributions, and the highest average absolute maximum error found is in the order of $10^{-4}\%$, for the F-I Beta distributions.

We also evaluated the 2-adjusted moment variance for fourth order moments derived from (conditional) quantiles with the bivariate Pearson-Tukey method. For the F-I Beta and F distributions the 2-adjusted moment variance provides a very good approximation to the conditional variance, much better han the regular Bayes linear adjusted variance. The average absolute error found for the F-I Beta distribution reduced from 75% for the regular adjusted variance to 3.2% for the 2-adjusted moment variance. The average maximum error is $1.2 \cdot 10^5\%$ for the regular adjusted variance for this distribution, but only 3.3% for the 2-adjusted moment variance. The average absolute error for the F distribution reduced from 19% for the regular adjusted variance to 0.8% for the 2-adjusted moment variance, and the average maximum error from 52% to 0.8%.

For the Cheriyan distributions we found the opposite however. The regular adjusted variance provides a much better approximation than the 2-adjusted moment variance, even though the Pearson-Tukey approximation of moments is by far the most accurate for these distributions. The average absolute error we found is 2.2% for the regular adjusted variance, and 77% for the 2-adjusted moment variance.

The Cheriyan distribution has by far the most constant conditional variance (see Table 5.1). In the previous section we suggested that the regular adjusted variance might be a good approximation for distributions for which the conditional variance is considered to be relatively constant over the condition. It might a good suggestion for further research to see whether a decision statistic can be derived that can help to choose between the regular adjusted and the 2-adjusted moment variance to approximate the conditional variance. It might be possible to base this decision statistic on the conditional quantiles used in the bivariate Pearson-Tukey method, which are readily available. From these conditional quantiles, the 5%−, 50%− and the 95%−quantiles, the conditional variance can be approximated using the univariate Pearson-Tukey method for three values of the condition: the 5%−, 50%− and the 95%−quantiles of the condition. When these three derived conditional variances are relatively constant, the regular adjusted variance might be preferred as approximation. If not, the 2-adjusted moment variance might be the better option.

Striking in the approximation results of the 2-adjusted moment variance is that the bias in the approximation error is relatively large compared to the average absolute error. For the F-I Beta distribution the bias is −0.9%, and the average absolute error 3.2%. For the F and Cheriyan distributions the approximation error is practically constant over the evaluation interval of the condition, with biases of resp. −0.8% and 77% for average absolute errors of as well −0.8% and 77%. It would be an interesting subject for further research to see if this bias could be reduced, thereby reducing the error in general even further.

In this thesis we have worked with exact (conditional) quantile assessments. In reality experts might not be willing or able to assess these precisely, and indifferent to small changes in their assessments. The impact of small changes in the quantile assessments has not been taken into consideration in this evaluation, and is an interesting topic for further research.

8.4 Performance Based Aggregation of Moment Assessments

In Chapter 7 we have developed a method to aggregate sets of (product) moment assessments from different experts based on their performance on test questions. By using the extended Pearson-Tukey method to translate quantile assessments to moment assessments, we have been able to compare the performance of our proposed moment model with that of a performance based aggregation method for quantile assessments, the classical model of Cooke, on five actual applications. Both models have shown to be of comparable performance in these applications. We have shown that the moment model possesses nice and desirable theoretical properties, and has many theoretical advantages over the classical model from the literature. The moment model e.g. enables us to test the performance of the experts in assessing dependencies as well. We recommend the moment model to aggregate sets of expert moments.

8.5 Implications of Research for Defence: Complementing the Defence Methodology Toolbox

Based on the results presented in this thesis we strongly advise to complement the Defence methodology toolbox with the Bayes linear methodology. We consider the Bayes linear methodology to be the best option available in general for situations in which subject matter experts are asked for quantitative assessments about interrelated quantities, and when assessments of magnitude and the uncertainty about these assessments are desired. The Bayes linear methodology reflects the discrete character of quantitative expert assessments and is flexible in the amount of detail that can both be specified by the experts and is needed for the decision problem at hand. This entails that the methodology can be applied within a relatively short time frame, leading to a short response time.

Furthermore, the methodology is assumption free as in that it does not require the quantities to have a probability distribution from a certain family of distributions. The assessments needed from subject matter experts to quantify a Bayes linear model are not highly involved. We belief these assessments do not require the subject matter experts to be expert in probabilistic methodology as well, and that they can be provided after only some basic introduction.

As mentioned in the introduction of this thesis, we hold the viewpoint that when sufficient data is available, this data should be preferred over subject matter expert judgements. The

Bayes linear methodology provides the vehicle to gradually switch from expert assessments to actually observed data when this becomes available. With the methodology the expert assessments are revised by data in a rational and coherent way. The Bayes linear methodology has the nice property that the more data is available, the less the revised beliefs will rely on the initial subject matter expert assessments.

Two limitations for the application of the methodology need to be mentioned. Firstly, most results reported in this thesis about the accuracy of the methodology are found by evaluation a set of bivariate distribution families, covering a wide spectrum of distributions common in practice. The results however do not necessarily generalise to other distribution families. The second limitation is related to the case of a high probability of observing extreme values. A Bayes linear model quantified using the bivariate Pearson-Tukey method, as proposed in this thesis, might provide less accurate results for cases in which variables are heavily skewed. Furthermore, for distributions with heavy tails not all moments needed for Bayes linear belief adjustment are necessarily finite. Transformations of the variables in the base of the belief structure might offer a solution, but this has not been explored in this thesis. We therefore identify this as an important area for further research.

Finally we note that although the conclusions in this thesis are formulated with respect to Defence applications, these conclusion are by no means restricted to this area of application.

Bibliography

Aczél, J. (1966), *Lectures on functional equations and their applications*, Academic Press.

Balakrishnan, N. & Lai, C. D. (2009), *Continuous bivariate distributions*, Springer Verlag.

Barnett, V. (1982), *Comparative statistical inference*, Wiley series in probability and mathematical statistics, 2nd ed. edn, John Wiley & Sons Ltd., New York; Chichester.

Barros, A. I. (2009), Dutch insights in irregular warfare, *in* 'SAS-071 NATO Proceedings 2009'.

Bedford, T. J. & Cooke, R. M. (2001), *Probabilistic Risk Analysis. Foundations and methods*, Cambridge University Press, Cambridge; New York.

Bolger, F. & Wright, G. (1993), 'Coherence and calibration in expert probability judgment', *Omega-International Journal Of Management Science* **21**(6), 629–644.

Carnap, R. (1950), *Logical foundations of probability*, University of Chicago Press.

Cleaves, D. A. (1986), Cognitive biases and corrective techniques: Proposals for improving elicitation procedures for knowledge-based systems, *in* 'Proceedings from the AAAI sponsored 2nd Annual Knowledge Acquisition for Knowledge-Based Systems Workshop', Banff, Canada, pp. 9–0 to 9–11.

Clemen, R. T., Fischer, G. W. & Winkler, R. L. (2000), 'Assessing dependence: Some experimental results', *Management Science* **46**(8), 1100–1115.

Cooke, R. M. (1986), 'Conceptual fallacies in subjective probability', *Topoi* **5**(1), 21–27.

Cooke, R. M. (1991), *Experts in uncertainty: opinion and subjective probability in science*, Environmental Ethics and Science Policy Series, Oxford University Press, New York.

Cooke, R. M. (2004), 'The anatomy of the squizzel: The role of operational definitions in representing uncertainty', *Reliability Engineering & System Safety* **85**(1-3), 313.

Cooke, R. M., Mendel, M. & Thijs, W. (1988), 'Calibration and information in expert resolution - a classical approach', *Automatica* **24**(1), 87–93.

Cowan, N. (2005), *Working memory capacity*, Psychology Press.

De Finetti, B. (1974), *Theory of probability*, Vol. vol I, John Wiley & Sons Ltd.

De Wit, M. S. (2001), Uncertainty in predictions of thermal comfort in buildings, PhD thesis, Delft University of Technology.

Ericson, W. A. (1969), 'A note on the posterior mean of a population mean', *Journal of the Royal Statistical Society. Series B (Methodological)* **31**(2), 332–334.

Ferrell, W. R. (1994), Discrete subjective probabilities and decision analysis: Elicitation, calibration and combination, *in* G. Wright & P. Ayton, eds, 'Subjective Probability', John Wiley & Sons, pp. 411–451.

Fishburn, P. C. (1986), 'The axioms of subjective probability', pp. 335–345.

Fraassen, B. v. (1984), 'Belief and the will', *Journal of Philosophy* **81**(5), 235–256.

Fraassen, B. v. (1995), 'Belief and the problem of ulysses and the sirens', *Philosophical Studies* **77**(1), 7–37.

French, S. (1985), Group consensus probability distributions: a critical survey, *in* J. Bernardo, M. De Groot, D. V. Lindley & A. Smith, eds, 'Bayesian Statistics 2', North-Holland, pp. 183–201.

French, S. (1986), *Decision theory: an introduction to the mathematics of rationality*, Halsted Press, Chichester.

French, S. (1987), Conflict of belief: when advisers disagree, *in* P. Bennett, ed., 'Analysing conflict and its resolution: some mathematical contributions', Oxford University Press, Oxford, pp. 93–111.

French, S. (2011), 'Aggregating expert judgement', *Revista de la Real Academia de Ciencias Exactas, Fisicas y Naturales. Serie A. Matematicas* **105**(1), 181–206. 1578-7303.

Garthwaite, P. H., Kadane, J. B. & O'Hagan, A. (2005), 'Statistical methods for eliciting probability distributions', *Journal of the American Statistical Association* **100**, 680–700.

Genest, C. & Zidek, J. V. (1986), 'Combining probability distributions: a critique and annotated bibliography', *Statistical Science* **1**, 114–148.

Gillies, D. (2000), *Philosophical theories of probability*, Routledge.

Gokhale, D. V. & Press, S. J. (1982), 'Assessment of a prior distribution for the correlation coefficient in a bivariate normal distribution', *Journal of the Royal Statistical Society* **145**(2), 237–249.

Goldstein, M. (1981), 'Revising previsions: a geometrical interpretation (with discussion)', *Journal of the Royal Statistical Society. Series B (Methodological)* **43**, 105–130.

Goldstein, M. (1986), 'Exchangeable belief structures', *Journal of the American Statistical Association* **81**(396), 971–976.

Goldstein, M. (1988*a*), 'Adjusting belief structures', *Journal of the Royal Statistical Society. Series B (Methodological)* **50**(1), 133–154.

Goldstein, M. (1988*b*), The data trajectory, *in* J. Bernardo, M. De Groot, D. Lindley & A. Smith, eds, 'Bayesian Statistics 3', Vol. 3, Oxford University Press, pp. 189–209.

Goldstein, M. (1991), 'Belief transforms and the comparison of hypotheses', *The Annals of Statistics* **19**(4), 2067–2089.

Goldstein, M. (1994), Belief revision: subjectivist principles and practice, *in* D. Prawitz & D. Westersthl, eds, 'Logic and Philosophy of Science in Uppsala', Kluwer Academic Publishers Group, pp. 117–130.

Goldstein, M. (1998), Bayes linear analysis, *in* S. Kotz, ed., 'Encyclopaedia of Statistical Sciences, update volume 3', Wiley, pp. 29–34.

Goldstein, M. & Wooff, D. (2007), *Bayes linear statistics: theory and methods*, Wiley.

Goossens, L. H., Cooke, R. M. & Kraan, B. C. P. (1998), Evaluation of weighting schemes for expert judgement studies, *in* A. Mosley & R. A. Bari, eds, 'PSAM4 Proceedings', Springer, New York, pp. 1937–1942.

Hacking, I. (2001), *An introduction to probability and inductive logic*, Cambridge University Press.

Hájek, A. (2010), Interpretations of probability, *in* E. N. Zalta, ed., 'The Stanford Encyclopedia of Philosophy', spring 2010 edn.

Harper, F., Goossens, L., Cooke, R., Hora, S., Young, M., Psler-Sauer, J., Miller, L., Kraan, B., Lui, C., McKay, M., Helton, J. & Jones, J. (1995), Joint usnrc/cec consequence uncertainty study: Summary of objectives, approach, application, and results for the dispersion and deposition uncertainty assessment. prepared for u.s. nuclear regulatory commission and commission of european communities, Technical Report NUREG/CR-6244, EUR 15855, Volumes I, II, III.

Hartigan, J. A. (1969), 'Linear bayesian methods', *Journal of the Royal Statistical Society. Series B (Methodological)* **31**(3), 446–454.

Hogarth, R. M. (1975), 'Cognitive processes and the assessment of subjective probability distributions; (with discussion)', *Journal of the American Statistical Association* **70**, 271–294.

Hogarth, R. M. (1987), *Judgement and Choice: The psychology of decision*, 2nd edition edn, John Wiley & Sons, New York.

Huygens, C. (1657), *De ratiociniis in ludo aleae.*

Jaynes, E. T. (2003), *Probability theory: the logic of science*, Cambridge Univ Pr.

Jeffreys, H. (1939), *Theory of probability*, reprinted in Oxford Classics in the Physical Sciences.

Johnson, D. (2002), 'Triangular approximations for continuous random variables in risk analysis', *Journal Of The Operational Research Society* **53**, 457–467.

Johnson, W. E. (1921), *Logic*, Cambridge University Press, Cambridge.

Kadane, J. B. & Wolfson, L. J. (1998), 'Experiences in elicitation', *The Statistician* **47**(Part 1), 3–19.

Kahneman, D., Slovic, P. & Tversky, A. (1982), *Judgement under uncertainty: Heuristics and biases*, Cambridge University Press, Cambridge; Massachusetts.

Keefer, D. L. & Bodily, S. E. (1983), 'Three-point approximations for continuous random variables', *Management Science* **29**(No. 5), 595–609.

Keefer, D. L. & Verdini, W. (1990), Comparison of simple approximations in estimating means and variances of beta distributions: Detailed results, Technical report, Department of Decision and Information Systems, College of Business, Arizona State University.

Keynes, J. M. (1921), *A Treatise on Probability*, Macmillan and Co.

Klugman, S. A., Panjer, H. H. & Willmot, G. E. (1998), *Loss models: from data to decisions*, Wiley New York.

Kolmogorov, A. N. (1956 (1933)), *Foundations of the theory of probability*, second english edition edn, Chelsea Publishing Company.

Koopman, B. O. (1940), 'The bases of probability', *Bulletin of the American Mathematical Society* **46**, Reprinted in Kyburg and Smokler (1964, 1980).

Kyburg, H. E. J. (1980), 'Conditionalization', *Journal of Philosophy* **77**, 98–114.

Kyburg, H. E. J. & Smokler, H., eds (1964), *Studies in Subjective Probability*, Wiley, New York.

Lad, F. (1996), *Operational subjective statistical methods: A mathematical, philosophical, and historical introduction*, John Wiley & Sons, New York.

Langer, E. (1975), 'The illusion of control', *The Journal of Personality and Social Psychology* **32**, 311–328.

Laplace, P. S. (1820), *A philosophical essay on probabilities*, Dover.

Lindley, D. V. (1985), Reconciliation of discrete probability distributions, *in* J. Bernardo, M. De Groot, D. Lindley & A. Smith, eds, 'Bayesian Statistics 2', Elsevier Science Publishers B.V., pp. 375–390.

Lindley, D. V. (2000), 'The philosophy of statistics', *Journal Of The Royal Statistical Society Series D-The Statistician* **49**, 293–319. Part 3.

Lindley, D. V., Tversky, A. & Brown, R. V. (1979), 'On the reconciliation of probability assessments; (with discussion)', *Journal of the Royal Statistical Society, Ser. A* **142**, 146–180.

Mardia, K. V. (1970), *Families of bivariate distributions*, Griffin's Statistical Monographs & Courses, Charles Griffin & Company Limited, London.

McConway, K. J. (1981), 'Marginalization and linear opinion pools', *Journal of the American Statistical Association* **76**, 410–414.

Meyer, M. A. & Booker, J. M. (2001), *Eliciting and analysing expert judgement: A practical guide*, ASA-SIAM Series on Statistics and Applied Probability, SIAM, Philadelphia.

Miller, G. A. (1956), 'The magical number seven, plus or minus two: some limits on our capacity for processing information', *Psychological Review* **63**, 81–97.

Morgan, M. G. & Henrion, M. (1990), *Uncertainty*, Cambridge University Press, Cambridge.

O'Hagan, A. (1988), *Probability: Methods and measurement*, Chapman & Hall, London.

O'Hagan, A., Buck, C. E., Daneshkhah, A., Eiser, J. R., Gathwaite, P. H., Jenkinson, D. J., Oakley, J. E. & Rakow, T. (2006), *Uncertain judgements. Eliciting experts' probabilities*, Statistics in Practice, John Wiley & Sons, Chichester.

O'Hagan, A. & Oakley, J. E. (2004), 'Probability is perfect, but we can't elicit it perfectly', *Reliability Engineering and System Safety* **85**, 239–248.

Pearson, E. S. & Tukey, J. W. (1965), 'Approximate means and standard deviations based on distances between percentage points of frequency curves', *Biometrika* **52**(3 and 4), 533–546.

Peterson, C. & Miller, A. (1964), 'Mode, median, and mean as optimal strategies', *Journal of Experimental Psychology* **68**(4), 363–367.

Popper, K. R. (1957), The propensity interpretation of the calculus of probability and the quantum theory, *in* S. Körner, ed., 'The Colston Papers', Vol. 9, pp. 65–70.

Popper, K. R. (1959), 'The propensity interpretation of probability', *British Journal of the Philosophy of Science* **10**, 25–42.

Qing, X. (2002), Risk analysis for real estate investment, Phd thesis, TU Delft.

Ramsey, F. (1926), *Truth and Probability*, reprinted in Ramsey (1931) and Kyburg and Smokler (1964, 1980).

Ramsey, F. (1931), *The Foundations of Mathematics and Other Logical Essays*, Paterson, NJ: Littlefield-Adams.

Reichenbach, H. (1949), *The Theory of Probability*, University of California Press, Berkeley.

Reilly, T. (2002), 'Estimating moments of subjectively assessed distributions', *Decision Sciences* **33**(1), 133. 00117315.

Revie, M., Bedford, T. J. & Walls, L. (2010), 'Evaluation of elicitation methods to quantify bayes linear models', *Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability* **224**(4), 322–332.

Rosenfield, I. (1988), *The invention of memory: A new view of the brain*, Basic Books, New York.

Savage, L. J. (1972 (1954)), *The foundations of statistics*, second revised edition edn, Dover.

Schafer, G. (2002), 'Review of operational subjective statistical methods by Frank Lad', *available at: http://www.glennshafer.com/assets/downloads/review14.pdf* .

Seidenfeld, T. (1979), 'Why i am not an objective bayesian; some reflections prompted by rosenkrantz', *Theory and Decision* **11**(4), 413–440.

Smith, J. E. (1993), 'Moment methods for decision analysis', *Management Science* **39**(3), 340. 00251909.

Stone, M. (1963), 'Robustness of non-ideal decision procedures', *Journal of the American Statistical Association* **58**(302), 480–486.

Van Elst, N. P. (1997), Betrouwbaarheid beweegbare waterkeringen [Reliability of movable water barriers], Technical Report WBBM report Series 35, Delft University Press.

Van Uven, M. J. (1947*a*), 'Extension of pearson's probability distribution to two variables I', *Ned. Akad. Wet. Proc.* **50**, 1063–1070.

Van Uven, M. J. (1947*b*), 'Extension of pearson's probability distribution to two variables II', *Ned. Akad. Wet. Proc.* **50**, 1252–1264.

Van Uven, M. J. (1948*a*), 'Extension of pearson's probability distribution to two variables III', *Ned. Akad. Wet. Proc.* **51**, 41–52.

Van Uven, M. J. (1948*b*), 'Extension of pearson's probability distribution to two variables IV', *Ned. Akad. Wet. Proc.* **51**, 191–196.

Venn, J. (1876), *The Logic of chance*, Macmillan.

Von Mises, R. V. (1957), *Probability, Statistics, and Truth*, revised english edition edn, Macmillan, New York.

Whittle, P. (1992), *Probability via expectation*, Springer Verlag.

Williams, B. S. (2010), 'Heuristics and biases in military decision making', *Military Review* (September-October 2010), 40–52.

Winkler, R. L. (1967), 'The assessment of prior distributions in bayesian analysis', *Journal of the American Statistical Association* **62**, 776–800.

Winkler, R. L. (1996), 'Uncertainty in probabilistic risk assessment', *Reliability Engineering and System Safety* **54**, 127–132.

Winkler, R. L. & Murphy, A. H. (1968), ''good' probability assessors', *Journal of applied meteorology* **7**, 751–758.

Wisse, B., Bedford, T. & Quigley, J. (2008*a*), 'Combining expert judgement using moment methods', *Reliability Engineering and System Safety, Special issue on the use of expert judgement* **93**, 675–686.

Wisse, B., Bedford, T. & Quigley, J. (2008*b*), 'Response to tony ohagans and simon french comments on expert judgement combination using moment methods', *Reliability Engineering and System Safety, Special issue on the use of expert judgement* **93**, 769–770.

Appendix A

Derivations of Equations for Chapter 5

The (condidional) means, variances and covariances in this appendix are taken from (Mardia 1970) and (Balakrishnan & Lai 2009).

A.1 Equality of Adjusted and Conditional Mean

A.1.1 F-I Beta

Conditional mean:

$$E(X|Y = y) = \frac{p_1}{p_1 + p_3}(1 - y)$$

Means, variance and covariance:

$$\begin{aligned} E(X) &= \tfrac{p_1}{p_1+p_2+p_3} \\ E(Y) &= \tfrac{p_2}{p_1+p_2+p_3} \\ Var(Y) &= \tfrac{p_2(p_1+p_3)}{(p_1+p_2+p_3)^2(p_1+p_2+p_3+1)} \\ Cov(X,Y) &= -\tfrac{p_1 p_2}{(p_1+p_2+p_3)^2(p_1+p_2+p_3+1)} \end{aligned}$$

Bayes linear adjusted mean:

$$\begin{aligned}
E_Y(X) &= E(X) + \frac{Cov(X,Y)}{Var(Y)}(y - E(Y)) \\
&= \frac{p_1}{p_1+p_2+p_3} \\
&\quad + \frac{-p_1p_2(p_1+p_2+p_3)^2(p_1+p_2+p_3+1)}{p_2(p_1+p_3)(p_1+p_2+p_3)^2(p_1+p_2+p_3+1)}\left(y - \frac{p_2}{p_1+p_2+p_3}\right) \\
&= \frac{p_1}{p_1+p_2+p_3} - \frac{p_1}{p_1+p_3}\left(y - \frac{p_2}{p_1+p_2+p_3}\right) \\
&= \frac{p_1(p_1+p_3)+p_1p_2}{(p_1+p_2+p_3)(p_1+p_3)} - \frac{p_1}{p_1+p_3}y \\
&= \frac{p_1}{p_1+p_3} - \frac{p_1}{p_1+p_3}y \\
&= \frac{p_1}{p_1+p_3}(1-y)
\end{aligned}$$

A.1.2 F

Conditional mean:

$$E(X|Y=y) = \frac{(\nu_0+\nu_2 y)}{\nu_0+\nu_2-2}$$

Means, variance and covariance:

$$\begin{aligned}
E(X) &= \tfrac{\nu_0}{\nu_0-2} \\
E(Y) &= \tfrac{\nu_0}{\nu_0-2} \\
Var(Y) &= \tfrac{2\nu_0^2(\nu_0+\nu_1-2)}{\nu_2(\nu_0-2)^2(\nu_0-4)} \\
Cov(X,Y) &= \tfrac{2\nu_0^2}{(\nu_0-2)^2(\nu_0-4)}
\end{aligned}$$

Bayes linear adjusted mean:

$$\begin{aligned}
E_Y(X) &= E(X) + \frac{Cov(X,Y)}{Var(Y)}(y - E(Y)) \\
&= \frac{\nu_0}{\nu_0-2} + \frac{2\nu_0^2\nu_2(\nu_0-2)^2(\nu_0-4)}{(\nu_0-2)^2(\nu_0-4)2\nu_0^2(\nu_0+\nu_1-2)}\left(y - \frac{\nu_0}{\nu_0-2}\right) \\
&= \frac{\nu_0}{\nu_0-2} + \frac{\nu_2}{(\nu_0+\nu_1-2)}\left(y - \frac{\nu_0}{\nu_0-2}\right) \\
&= \frac{\nu_0}{\nu_0-2} - \frac{\nu_0\nu_2}{(\nu_0+\nu_2-2)(\nu_0-2)} + \frac{\nu_2}{(\nu_0+\nu_2-2)}y \\
&= \frac{\nu_0(\nu_0+\nu_2-2)-\nu_0\nu_2}{(\nu_0+\nu_2-2)(\nu_0-2)} + \frac{\nu_2}{(\nu_0+\nu_2-2)}y \\
&= \frac{\nu_0(\nu_0-2)}{(\nu_0+\nu_2-2)(\nu_0-2)} + \frac{\nu_2}{(\nu_0+\nu_2-2)}y \\
&= \frac{(\nu_0+\nu_2 y)}{\nu_0+\nu_2-2}
\end{aligned}$$

A.1.3 Kibble

Conditional mean:

$$E(X|Y = y) = \rho(y - \alpha) + \alpha$$

Means, variance and covariance:

$$\begin{aligned} E(X) &= \alpha \\ E(Y) &= \alpha \\ Var(Y) &= \alpha \\ Cov(X, Y) &= \rho\alpha \end{aligned}$$

Bayes linear adjusted mean:

$$\begin{aligned} E_Y(X) &= E(X) + \frac{Cov(X, Y)}{Var(Y)}(y - E(Y)) \\ &= \alpha + \frac{\rho\alpha}{\alpha}(y - \alpha) \\ &= \rho(y - \alpha) + \alpha \end{aligned}$$

A.1.4 Cheriyan

Conditional mean:

$$E(X|Y = y) = \theta_1 + \frac{\theta_3}{\theta_2 + \theta_3} y$$

Means, variance and covariance:

$$\begin{aligned} E(X) &= \theta_1 + \theta_3 \\ E(Y) &= \theta_2 + \theta_3 \\ Var(Y) &= \theta_2 + \theta_3 \\ Cov(X, Y) &= \theta_3 \end{aligned}$$

Bayes linear adjusted mean:

$$\begin{aligned} E_Y(X) &= E(X) + \frac{Cov(X, Y)}{Var(Y)}(y - E(Y)) \\ &= \theta_1 + \theta_3 + \frac{\theta_3}{\theta_2 + \theta_3}(y - \theta_2 + \theta_3) \\ &= \theta_1 + \frac{\theta_3}{\theta_2 + \theta_3} y \end{aligned}$$

A.2 Derivation of d_{var}

A.2.1 F-I Beta

Conditional variance:

$$Var(X|Y=y) = \frac{p_1 p_3}{(p_1+p_3)^2(1+p_1+p_3)}(1-y)^2$$

Variances and covariance:

$$\begin{aligned}
Var(X) &= \tfrac{p_1(p_2+p_3)}{(p_1+p_2+p_3)^2(p_1+p_2+p_3+1)} \\
Var(Y) &= \tfrac{p_2(p_1+p_3)}{(p_1+p_2+p_3)^2(p_1+p_2+p_3+1)} \\
Cov(X,Y) &= -\tfrac{p_1 p_2}{(p_1+p_2+p_3)^2(p_1+p_2+p_3+1)}
\end{aligned}$$

Bayes linear adjusted variance:

$$\begin{aligned}
Var_Y(X) &= Var(X) - \frac{Cov(X,Y)^2}{Var(Y)} \\
&= \frac{p_1(p_2+p_3)}{(p_1+p_2+p_3)^2(p_1+p_2+p_3+1)} \\
&\quad - \frac{p_1^2 p_2^2(p_1+p_2+p_3)^2(p_1+p_2+p_3+1)}{(p_1+p_2+p_3)^4(p_1+p_2+p_3+1)^2 p_2(p_1+p_3)} \\
&= \frac{p_1(p_2+p_3)}{(p_1+p_2+p_3)^2(p_1+p_2+p_3+1)} \\
&\quad - \frac{p_1^2 p_2}{(p_1+p_2+p_3)^2(p_1+p_2+p_3+1)(p_1+p_3)} \\
&= \frac{p_1(p_2+p_3)(p_1+p_3)}{(p_1+p_2+p_3)^2(p_1+p_2+p_3+1)(p_1+p_3)} \\
&\quad - \frac{p_1^2 p_2}{(p_1+p_2+p_3)^2(p_1+p_2+p_3+1)(p_1+p_3)} \\
&= \frac{p_1(p_2+p_3)(p_1+p_3) - p_1^2 p_2}{(p_1+p_2+p_3)^2(p_1+p_2+p_3+1)(p_1+p_3)}
\end{aligned}$$

So $d_{var,Filon-Isserlis}$ is:

$$\begin{aligned}
d_{var,Filon-Isserlis}(y) &= Var_Y(X) - Var(X|Y=y) \\
&= \frac{p_1(p_2+p_3)(p_1+p_3)-p_1^2p_2}{(p_1+p_2+p_3)^2(p_1+p_2+p_3+1)(p_1+p_3)} \\
&\quad -\frac{p_1p_3}{(p_1+p_3)^2(1+p_1+p_3)}(1-y)^2 \\
&= \frac{p_1(p_2+p_3)(p_1+p_3)-p_1^2p_2}{(p_1+p_2+p_3)^2(p_1+p_2+p_3+1)(p_1+p_3)} \\
&\quad -\frac{p_1p_3}{(p_1+p_3)^2(1+p_1+p_3)}+\frac{p_1p_3}{(p_1+p_3)^2(1+p_1+p_3)}(2y-y^2) \\
&= \frac{p_1(p_2+p_3)(p_1+p_3)^2(1+p_1+p_3)-p_1^2p_2(p_1+p_3)(1+p_1+p_3)}{(p_1+p_3)^2(1+p_1+p_3)(p_1+p_2+p_3)^2(p_1+p_2+p_3+1)} \\
&\quad -\frac{p_1p_3(p_1+p_2+p_3)^2(p_1+p_2+p_3+1)}{(p_1+p_3)^2(1+p_1+p_3)(p_1+p_2+p_3)^2(p_1+p_2+p_3+1)} \\
&\quad +\frac{p_1p_3}{(p_1+p_3)^2(1+p_1+p_3)}(2y-y^2) \\
&= \frac{p_1p_2p_3(1+2p_1+p_2+2p_3)}{(p_1+p_3)^2(1+p_1+p_3)(p_1+p_2+p_3)(1+p_1+p_2+p_3)} \\
&\quad +\frac{p_1p_3}{(p_1+p_3)^2(1+p_1+p_3)}(2y-y^2)
\end{aligned}$$

A.2.2 F

Conditional variance:

$$Var(X|Y=y) = \frac{2(\nu_0+\nu_1+\nu_2-2)}{\nu_1(\nu_0+\nu_2-2)^2(\nu_0+\nu_2-4)}(\nu_0+\nu_2 y)^2$$

Variances and covariance:

$$\begin{aligned}
Var(X) &= \tfrac{2\nu_0^2(\nu_0+\nu_2-2)}{\nu_1(\nu_0-2)^2(\nu_0-4)} \\
Var(Y) &= \tfrac{2\nu_0^2(\nu_0+\nu_1-2)}{\nu_2(\nu_0-2)^2(\nu_0-4)} \\
Cov(X,Y) &= \tfrac{2\nu_0^2}{(\nu_0-2)^2(\nu_0-4)}
\end{aligned}$$

Bayes linear adjusted variance:

$$\begin{aligned}
Var_Y(X) &= Var(X) - \frac{Cov(X,Y)^2}{Var(Y)} \\
&= \frac{2\nu_0^2(\nu_0+\nu_2-2)}{\nu_1(\nu_0-2)^2(\nu_0-4)} - \frac{4\nu_0^4\nu_2(\nu_0-2)^2(\nu_0-4)}{(\nu_0-2)^4(\nu_0-4)^2 2\nu_0^2(\nu_0+\nu_1-2)} \\
&= \frac{2\nu_0^2(\nu_0+\nu_2-2)}{\nu_1(\nu_0-2)^2(\nu_0-4)} - \frac{2\nu_0^2\nu_2}{(\nu_0-2)^2(\nu_0-4)(\nu_0+\nu_1-2)} \\
&= \frac{2\nu_0^2(\nu_0+\nu_2-2)(\nu_0+\nu_1-2)}{\nu_1(\nu_0-2)^2(\nu_0-4)(\nu_0+\nu_1-2)} - \frac{2\nu_0^2\nu_1\nu_2}{\nu_1(\nu_0-2)^2(\nu_0-4)(\nu_0+\nu_1-2)} \\
&= \frac{2\nu_0^2(\nu_0+\nu_2-2)(\nu_0+\nu_1-2) - 2\nu_0^2\nu_1\nu_2}{\nu_1(\nu_0-2)^2(\nu_0-4)(\nu_0+\nu_1-2)} \\
&= \frac{2\nu_0^2[(\nu_0+\nu_2-2)(\nu_0+\nu_1-2) - \nu_1\nu_2]}{\nu_1(\nu_0-2)^2(\nu_0-4)(\nu_0+\nu_1-2)}
\end{aligned}$$

So $d_{var,F}$ is:

$$\begin{aligned}
d_{var,F}(y) &= Var_Y(X) - Var(X|Y=y) \\
&= \frac{2\nu_0^2[(\nu_0+\nu_2-2)(\nu_0+\nu_1-2) - \nu_1\nu_2]}{\nu_1(\nu_0-2)^2(\nu_0-4)(\nu_0+\nu_1-2)} \\
&\quad + \frac{2\nu_0^2(\nu_0+\nu_1+\nu_2-2)}{\nu_1(\nu_0+\nu_2-2)^2(\nu_0+\nu_2-4)} \\
&\quad + \frac{2(\nu_0+\nu_1+\nu_2-2)}{\nu_1(\nu_0+\nu_2-2)^2(\nu_0+\nu_2-4)}(-2\nu_0\nu_2 y - \nu_2^2 y^2)
\end{aligned}$$

A.2.3 Kibble

Conditional variance:

$$Var(X|Y=y) = (1-\rho)^2\alpha + 2\rho(1-\rho)y$$

Variances and covariance:

$$\begin{aligned}
Var(X) &= \alpha \\
Var(Y) &= \alpha \\
Cov(X,Y) &= \rho\alpha
\end{aligned}$$

Bayes linear adjusted variance:

$$\begin{aligned} Var_Y(X) &= Var(X) - \frac{Cov(X,Y)^2}{Var(Y)} \\ &= \alpha - \frac{\rho^2\alpha^2}{\alpha} \\ &= \alpha(1-\rho^2) \end{aligned}$$

So $d_{var,Kibble}$ is:

$$\begin{aligned} d_{var,Kibble}(y) &= Var_Y(X) - Var(X|Y=y) \\ &= \alpha(1-\rho^2) - (1-\rho)^2\alpha - 2\rho(1-\rho)y \\ &= \alpha(1-\rho^2) - \alpha(1-\rho^2) - \alpha(-2\rho+2\rho^2) - 2\rho(1-\rho)y \\ &= -\alpha(-2\rho+2\rho^2) - 2\rho(1-\rho)y \\ &= 2\alpha\rho(1-\rho) - 2\rho(1-\rho)y \\ &= -2\rho(1-\rho)(y-\alpha) \\ &= -2\rho(1-\rho)(y-E(Y)) \end{aligned}$$

A.2.4 Cheriyan

Conditional variance:

$$Var(X|Y=y) = \theta_1 + \frac{\theta_2\theta_3}{(\theta_2+\theta_3)^2(1+\theta_2+\theta_3)}y^2$$

Variances and covariance:

$$\begin{aligned} Var(X) &= \theta_1 + \theta_3 \\ Var(Y) &= \theta_2 + \theta_3 \\ Cov(X,Y) &= \theta_3 \end{aligned}$$

Bayes linear adjusted variance:

$$\begin{aligned} Var_Y(X) &= Var(X) - \frac{Cov(X,Y)^2}{Var(Y)} \\ &= \theta_1 + \theta_3 - \frac{\theta_3^2}{\theta_2+\theta_3} \end{aligned}$$

So $d_{var,Cheriyan}$ is:

$$\begin{aligned}
d_{var,Cheriyan}(y) &= Var_Y(X) - Var(X|Y=y) \\
&= \theta_1 + \theta_3 - \frac{\theta_3^2}{\theta_2+\theta_3} - \theta_1 - \frac{\theta_2\theta_3}{(\theta_2+\theta_3)^2(1+\theta_2+\theta_3)}y^2 \\
&= \theta_3 - \frac{\theta_3^2}{\theta_2+\theta_3} - \frac{\theta_2\theta_3}{(\theta_2+\theta_3)^2(1+\theta_2+\theta_3)}y^2 \\
&= \frac{\theta_3(\theta_2+\theta_3)}{\theta_2+\theta_3} - \frac{\theta_3^2}{\theta_2+\theta_3} - \frac{\theta_2\theta_3}{(\theta_2+\theta_3)^2(1+\theta_2+\theta_3)}y^2 \\
&= \frac{\theta_2\theta_3}{\theta_2+\theta_3} - \frac{\theta_2\theta_3}{(\theta_2+\theta_3)^2(1+\theta_2+\theta_3)}y^2
\end{aligned}$$

Appendix B

Additional Tables to Chapter 5

B.1 AARD against skewness and kurtosis: F, Kibble and Cheriyan

Table B.1: Mean, standard deviation and maximum value of skewness and excess kurtosis for different percentile ranges of AARD of 10,000 F distributions. AARD in %.

F		Percentiles of AARD											
		0-1%	1-10%	10-20%	20-30%	30-40%	40-50%	50-60%	60-70%	70-80%	80-90%	90-99%	99-100%
AARD	mean	2.320	5.979	9.610	11.648	13.081	14.676	16.667	19.439	23.357	29.803	42.384	56.576
	st.dev.	0.387	1.494	0.764	0.460	0.408	0.510	0.675	0.939	1.382	2.497	5.261	1.857
	max	2.862	8.111	10.803	12.398	13.809	15.575	17.882	21.159	25.891	34.823	53.628	59.977
skewness of $X < 0$	mean												
	st.dev.												
	min												
skewness of $X > 0$	mean	0.645	0.696	0.692	0.682	0.690	0.726	0.762	0.833	0.913	1.077	1.435	1.847
	st.dev.	0.244	0.288	0.309	0.302	0.273	0.280	0.236	0.284	0.247	0.262	0.282	0.207
	max	2.100	3.020	3.291	3.291	2.990	2.967	2.191	3.071	3.232	3.260	3.708	3.021
skewness of $Y < 0$	mean												
	st.dev.												
	min												
skewness of $Y > 0$	mean	2.293	1.145	0.717	0.631	0.617	0.665	0.691	0.762	0.856	1.017	1.380	1.799
	st.dev.	0.518	0.487	0.257	0.229	0.217	0.236	0.163	0.192	0.169	0.175	0.205	0.069
	max	2.993	3.370	3.651	2.605	2.673	2.811	2.308	2.649	2.020	2.119	2.009	1.868
kurtosis of X	mean	0.822	0.984	0.989	0.960	0.958	1.071	1.135	1.402	1.661	2.347	4.350	7.510
	st.dev.	0.857	1.193	1.380	1.337	1.091	1.242	0.845	1.374	1.233	1.459	2.022	1.819
	max	6.831	14.292	18.025	18.025	13.925	13.640	7.638	14.949	17.165	17.562	25.060	18.064
kurtosis of Y	mean	8.524	2.498	1.004	0.803	0.775	0.912	0.936	1.163	1.455	2.089	4.025	7.127
	st.dev.	3.788	2.412	1.278	0.952	0.953	1.112	0.617	0.878	0.735	0.900	1.335	0.622
	max	13.951	19.227	23.991	12.081	12.948	14.829	9.853	14.202	8.143	9.563	8.731	7.732

Table B.2: Mean, standard deviation and maximum value of skewness and excess kurtosis for different percentile ranges of AARD of 10, 000 Kibble distributions. AARD in %.

Kibble		Percentiles of AARD											
		0-1%	1-10%	10-20%	20-30%	30-40%	40-50%	50-60%	60-70%	70-80%	80-90%	90-99%	99-100%
AARD	mean	0.344	3.825	9.522	14.382	18.482	22.276	25.384	28.678	33.573	42.137	62.827	108.485
	st.dev.	0.228	1.788	1.567	1.243	1.193	0.997	0.848	1.117	1.800	3.422	11.018	26.175
	max	0.698	6.822	12.187	16.428	20.482	23.896	26.889	30.756	36.979	48.841	90.073	315.114
skewness of $X < 0$	mean												
	st.dev.												
	min												
skewness of $X > 0$	mean	1.079	1.010	0.968	0.983	0.894	0.956	0.978	1.036	1.264	1.696	2.847	5.095
	st.dev.	0.809	1.240	0.768	1.059	0.396	0.721	0.958	1.147	1.091	1.359	2.306	4.219
	max	5.928	23.609	15.436	20.833	5.556	11.366	14.917	28.035	16.156	13.546	25.507	21.107
skewness of $Y < 0$	mean												
	st.dev.												
	min												
skewness of $Y > 0$	mean	1.079	1.010	0.968	0.983	0.894	0.956	0.978	1.036	1.264	1.696	2.847	5.095
	st.dev.	0.809	1.240	0.768	1.059	0.396	0.721	0.958	1.147	1.091	1.359	2.306	4.219
	max	5.928	23.609	15.436	20.833	5.556	11.366	14.917	28.035	16.156	13.546	25.507	21.107
kurtosis of X	mean	2.720	3.836	2.290	3.131	1.433	2.149	2.810	3.579	4.180	7.080	20.120	65.372
	st.dev.	7.308	37.118	13.857	25.610	2.376	8.900	17.920	39.345	18.449	19.792	55.284	126.035
	max	52.708	836.075	357.402	650.992	46.311	193.779	333.763	1178.908	391.501	275.244	975.887	668.283
kurtosis of Y	mean	2.720	3.836	2.290	3.131	1.433	2.149	2.810	3.579	4.180	7.080	20.120	65.372
	st.dev.	7.308	37.118	13.857	25.610	2.376	8.900	17.920	39.345	18.449	19.792	55.284	126.035
	max	52.708	836.075	357.402	650.992	46.311	193.779	333.763	1178.908	391.501	275.244	975.887	668.283

Table B.3: Mean, standard deviation and maximum value of skewness and excess kurtosis for different percentile ranges of AARD of $10,000$ Cheriyan distributions. AARD in %.

Cheriyan		Percentiles of AARD											
		0-1%	1-10%	10-20%	20-30%	30-40%	40-50%	50-60%	60-70%	70-80%	80-90%	90-99%	99-100%
AARD	mean	0.030	0.291	0.653	0.944	1.230	1.540	1.941	2.450	3.159	4.061	5.443	8.761
	st.dev.	0.017	0.126	0.088	0.085	0.086	0.098	0.129	0.178	0.240	0.285	0.666	1.526
	max	0.062	0.492	0.800	1.089	1.378	1.721	2.175	2.771	3.603	4.611	7.213	15.155
skewness of $X < 0$	mean												
	st.dev.												
	min												
skewness of $X > 0$	mean	0.082	0.076	0.075	0.072	0.078	0.082	0.090	0.099	0.103	0.109	0.129	0.180
	st.dev.	0.031	0.043	0.038	0.024	0.033	0.028	0.033	0.041	0.038	0.038	0.059	0.062
	max	0.190	0.949	0.415	0.314	0.546	0.241	0.335	0.486	0.392	0.379	1.046	0.482
skewness of $Y < 0$	mean												
	st.dev.												
	min												
skewness of $Y > 0$	mean	0.135	0.105	0.093	0.086	0.082	0.080	0.079	0.080	0.078	0.077	0.086	0.132
	st.dev.	0.160	0.059	0.044	0.031	0.027	0.025	0.025	0.029	0.025	0.023	0.023	0.038
	max	1.513	0.668	0.695	0.498	0.224	0.288	0.314	0.426	0.422	0.427	0.264	0.301
kurtosis of X	mean	0.012	0.012	0.011	0.009	0.011	0.011	0.014	0.017	0.018	0.020	0.030	0.054
	st.dev.	0.010	0.046	0.019	0.008	0.019	0.010	0.015	0.023	0.019	0.018	0.063	0.046
	max	0.054	1.351	0.259	0.147	0.446	0.087	0.168	0.355	0.231	0.216	1.641	0.349
kurtosis of Y	mean	0.065	0.022	0.016	0.013	0.011	0.010	0.010	0.011	0.010	0.010	0.012	0.028
	st.dev.	0.346	0.041	0.030	0.016	0.010	0.009	0.009	0.013	0.011	0.010	0.008	0.020
	max	3.433	0.669	0.724	0.372	0.076	0.125	0.148	0.272	0.267	0.273	0.105	0.136

B.2 AARD against $|\delta 2, 1|$ and $|\delta 2, 2|$: F, Kibble and Cheriyan

Table B.4: Mean, standard deviation and maximum value of the absolute value of higher order correlation differences $|\delta_{2,2}|$ and $|\delta_{2,2}|$ for different percentile ranges of AARD of 10,000 F-I Beta, F, Kibble and Cheriyan distributions. AARD in %.

		Percentiles of AARD											
		0-1%	1-10%	10-20%	20-30%	30-40%	40-50%	50-60%	60-70%	70-80%	80-90%	90-99%	99-100%
F-I Beta $\lvert\delta_{2,1}\rvert$	mean	$4.0\cdot10^{-2}$	$2.0\cdot10^{-2}$	$9.3\cdot10^{-3}$	$5.5\cdot10^{-3}$	$4.1\cdot10^{-3}$	$4.2\cdot10^{-3}$	$3.1\cdot10^{-3}$	$3.1\cdot10^{-3}$	$3.0\cdot10^{-3}$	$2.8\cdot10^{-3}$	$2.8\cdot10^{-3}$	$2.3\cdot10^{-3}$
	st.dev.	$4.2\cdot10^{-2}$	$3.2\cdot10^{-2}$	$1.5\cdot10^{-2}$	$1.0\cdot10^{-2}$	$8.3\cdot10^{-3}$	$1.2\cdot10^{-2}$	$6.7\cdot10^{-3}$	$7.4\cdot10^{-3}$	$7.4\cdot10^{-3}$	$1.0\cdot10^{-2}$	$1.1\cdot10^{-2}$	$7.2\cdot10^{-3}$
	max	$2.6\cdot10^{-1}$	$6.0\cdot10^{-1}$	$1.7\cdot10^{-1}$	$1.2\cdot10^{-1}$	$9.4\cdot10^{-2}$	$2.4\cdot10^{-1}$	$9.5\cdot10^{-2}$	$1.1\cdot10^{-1}$	$7.8\cdot10^{-2}$	$1.7\cdot10^{-1}$	$2.5\cdot10^{-1}$	$4.9\cdot10^{-2}$
F $\lvert\delta_{2,1}\rvert$	mean	$6.0\cdot10^{-3}$	$2.0\cdot10^{-3}$	$3.5\cdot10^{-4}$	$2.2\cdot10^{-4}$	$2.0\cdot10^{-4}$	$2.4\cdot10^{-4}$	$1.9\cdot10^{-4}$	$3.3\cdot10^{-4}$	$6.4\cdot10^{-4}$	$1.7\cdot10^{-3}$	$6.4\cdot10^{-3}$	$1.7\cdot10^{-2}$
	st.dev.	$2.8\cdot10^{-3}$	$2.8\cdot10^{-3}$	$1.4\cdot10^{-3}$	$1.2\cdot10^{-3}$	$1.1\cdot10^{-3}$	$1.2\cdot10^{-3}$	$3.7\cdot10^{-4}$	$3.6\cdot10^{-4}$	$2.9\cdot10^{-4}$	$1.0\cdot10^{-3}$	$3.6\cdot10^{-3}$	$2.9\cdot10^{-3}$
	max	$1.4\cdot10^{-2}$	$2.2\cdot10^{-2}$	$2.8\cdot10^{-2}$	$1.6\cdot10^{-2}$	$1.6\cdot10^{-2}$	$1.6\cdot10^{-2}$	$9.0\cdot10^{-3}$	$5.8\cdot10^{-3}$	$2.4\cdot10^{-3}$	$1.2\cdot10^{-2}$	$1.9\cdot10^{-2}$	$2.0\cdot10^{-2}$
Kibble $\lvert\delta_{2,1}\rvert$	mean	$5.1\cdot10^{-5}$	$5.2\cdot10^{-4}$	$1.5\cdot10^{-3}$	$2.5\cdot10^{-3}$	$2.8\cdot10^{-3}$	$4.1\cdot10^{-3}$	$5.3\cdot10^{-3}$	$6.7\cdot10^{-3}$	$1.2\cdot10^{-2}$	$2.7\cdot10^{-2}$	$9.8\cdot10^{-2}$	$2.5\cdot10^{-1}$
	st.dev.	$7.3\cdot10^{-5}$	$7.6\cdot10^{-4}$	$2.5\cdot10^{-3}$	$5.1\cdot10^{-3}$	$2.8\cdot10^{-3}$	$5.3\cdot10^{-3}$	$1.1\cdot10^{-2}$	$1.0\cdot10^{-2}$	$1.6\cdot10^{-2}$	$3.0\cdot10^{-2}$	$9.3\cdot10^{-2}$	$1.4\cdot10^{-1}$
	max	$4.7\cdot10^{-4}$	$1.0\cdot10^{-2}$	$6.0\cdot10^{-2}$	$9.3\cdot10^{-2}$	$2.7\cdot10^{-2}$	$8.9\cdot10^{-2}$	$2.0\cdot10^{-1}$	$2.1\cdot10^{-1}$	$2.4\cdot10^{-1}$	$2.9\cdot10^{-1}$	$5.1\cdot10^{-1}$	$5.9\cdot10^{-1}$
Cheriyan $\lvert\delta_{2,1}\rvert$	mean	$3.7\cdot10^{-6}$	$2.9\cdot10^{-6}$	$2.1\cdot10^{-6}$	$7.5\cdot10^{-7}$	$5.1\cdot10^{-7}$	$4.8\cdot10^{-7}$	$5.6\cdot10^{-7}$	$8.7\cdot10^{-7}$	$9.0\cdot10^{-7}$	$9.0\cdot10^{-7}$	$8.3\cdot10^{-7}$	$5.2\cdot10^{-6}$
	st.dev.	$1.4\cdot10^{-5}$	$3.2\cdot10^{-5}$	$3.2\cdot10^{-5}$	$5.4\cdot10^{-6}$	$1.4\cdot10^{-6}$	$1.4\cdot10^{-6}$	$3.1\cdot10^{-6}$	$7.4\cdot10^{-6}$	$1.3\cdot10^{-5}$	$1.3\cdot10^{-5}$	$1.8\cdot10^{-6}$	$1.1\cdot10^{-5}$
	max	$1.1\cdot10^{-4}$	$8.9\cdot10^{-4}$	$9.7\cdot10^{-4}$	$1.4\cdot10^{-4}$	$1.9\cdot10^{-5}$	$2.8\cdot10^{-5}$	$9.2\cdot10^{-5}$	$2.0\cdot10^{-4}$	$4.0\cdot10^{-4}$	$4.1\cdot10^{-4}$	$3.1\cdot10^{-5}$	$9.2\cdot10^{-5}$
F-I Beta $\lvert\delta_{2,2}\rvert$	mean	$3.9\cdot10^{-2}$	$2.0\cdot10^{-2}$	$1.1\cdot10^{-2}$	$7.5\cdot10^{-3}$	$7.1\cdot10^{-3}$	$8.2\cdot10^{-3}$	$8.5\cdot10^{-3}$	$1.1\cdot10^{-2}$	$1.4\cdot10^{-2}$	$2.0\cdot10^{-2}$	$4.1\cdot10^{-2}$	$1.1\cdot10^{-1}$
	st.dev.	$4.0\cdot10^{-2}$	$3.1\cdot10^{-2}$	$1.5\cdot10^{-2}$	$1.0\cdot10^{-2}$	$8.7\cdot10^{-3}$	$1.3\cdot10^{-2}$	$8.1\cdot10^{-3}$	$9.3\cdot10^{-3}$	$1.1\cdot10^{-2}$	$1.4\cdot10^{-2}$	$2.3\cdot10^{-2}$	$3.6\cdot10^{-2}$
	max	$2.5\cdot10^{-1}$	$5.7\cdot10^{-1}$	$1.7\cdot10^{-1}$	$1.2\cdot10^{-1}$	$9.8\cdot10^{-2}$	$2.4\cdot10^{-1}$	$9.8\cdot10^{-2}$	$1.2\cdot10^{-1}$	$8.8\cdot10^{-2}$	$1.9\cdot10^{-1}$	$3.2\cdot10^{-1}$	$2.1\cdot10^{-1}$
F $\lvert\delta_{2,2}\rvert$	mean	$6.0\cdot10^{-3}$	$2.2\cdot10^{-3}$	$6.5\cdot10^{-4}$	$5.3\cdot10^{-4}$	$5.0\cdot10^{-4}$	$5.4\cdot10^{-4}$	$4.3\cdot10^{-4}$	$5.4\cdot10^{-4}$	$4.8\cdot10^{-4}$	$6.4\cdot10^{-4}$	$1.1\cdot10^{-3}$	$1.8\cdot10^{-3}$
	st.dev.	$2.8\cdot10^{-3}$	$2.8\cdot10^{-3}$	$1.6\cdot10^{-3}$	$1.6\cdot10^{-3}$	$1.5\cdot10^{-3}$	$1.7\cdot10^{-3}$	$1.1\cdot10^{-3}$	$1.6\cdot10^{-3}$	$1.5\cdot10^{-3}$	$1.9\cdot10^{-3}$	$2.6\cdot10^{-3}$	$3.0\cdot10^{-3}$
	max	$1.4\cdot10^{-2}$	$2.3\cdot10^{-2}$	$3.0\cdot10^{-2}$	$1.8\cdot10^{-2}$	$1.9\cdot10^{-2}$	$1.9\cdot10^{-2}$	$1.2\cdot10^{-2}$	$1.4\cdot10^{-2}$	$1.6\cdot10^{-2}$	$1.9\cdot10^{-2}$	$3.2\cdot10^{-2}$	$2.1\cdot10^{-2}$
Kibble $\lvert\delta_{2,2}\rvert$	mean	$8.9\cdot10^{-5}$	$8.6\cdot10^{-4}$	$2.2\cdot10^{-3}$	$3.3\cdot10^{-3}$	$3.5\cdot10^{-3}$	$4.3\cdot10^{-3}$	$4.8\cdot10^{-3}$	$5.4\cdot10^{-3}$	$9.4\cdot10^{-3}$	$1.9\cdot10^{-2}$	$3.7\cdot10^{-2}$	$2.3\cdot10^{-2}$
	st.dev.	$1.1\cdot10^{-4}$	$1.0\cdot10^{-3}$	$2.9\cdot10^{-3}$	$5.1\cdot10^{-3}$	$4.1\cdot10^{-3}$	$6.6\cdot10^{-3}$	$9.9\cdot10^{-3}$	$1.0\cdot10^{-2}$	$1.6\cdot10^{-2}$	$2.7\cdot10^{-2}$	$3.8\cdot10^{-2}$	$3.1\cdot10^{-2}$
	max	$6.7\cdot10^{-4}$	$9.8\cdot10^{-3}$	$5.4\cdot10^{-2}$	$8.0\cdot10^{-2}$	$3.2\cdot10^{-2}$	$8.0\cdot10^{-2}$	$1.4\cdot10^{-1}$	$1.4\cdot10^{-1}$	$1.5\cdot10^{-1}$	$1.6\cdot10^{-1}$	$1.5\cdot10^{-1}$	$1.6\cdot10^{-1}$
Cheriyan $\lvert\delta_{2,2}\rvert$	mean	$3.6\cdot10^{-6}$	$2.9\cdot10^{-6}$	$2.3\cdot10^{-6}$	$9.7\cdot10^{-7}$	$8.6\cdot10^{-7}$	$8.4\cdot10^{-7}$	$1.1\cdot10^{-6}$	$1.9\cdot10^{-6}$	$1.8\cdot10^{-6}$	$2.1\cdot10^{-6}$	$5.1\cdot10^{-6}$	$1.6\cdot10^{-5}$
	st.dev.	$1.4\cdot10^{-5}$	$3.2\cdot10^{-5}$	$3.2\cdot10^{-5}$	$5.8\cdot10^{-6}$	$2.0\cdot10^{-6}$	$1.7\cdot10^{-6}$	$3.2\cdot10^{-6}$	$8.7\cdot10^{-6}$	$6.3\cdot10^{-6}$	$7.0\cdot10^{-6}$	$3.3\cdot10^{-5}$	$3.6\cdot10^{-5}$
	max	$1.2\cdot10^{-4}$	$9.0\cdot10^{-4}$	$9.8\cdot10^{-4}$	$1.6\cdot10^{-4}$	$3.9\cdot10^{-5}$	$2.8\cdot10^{-5}$	$8.3\cdot10^{-5}$	$2.3\cdot10^{-4}$	$1.8\cdot10^{-4}$	$2.0\cdot10^{-4}$	$9.3\cdot10^{-4}$	$2.7\cdot10^{-4}$

Appendix C

Detailed Results of the Comparison of the Classical Model and the Moment Model

The moment model scores and weights reported in this thesis were calculated with Microsoft Excel. The classical model calibration and information scores, and the classical model weights were calculated in the Excalibur v1.0 Light software, provided by Roger Cooke from TU Delft.

C.1 Case 1: Prime Rent Indices

The prime rent case involved 5 investment managers who each gave assessments on the rent indices of office space for the major cities in the Netherlands for the future, see (Qing 2002). The 16 original seed variables were used to derive the decision maker weights for both the classical model, C_{DM}, and the moment model, M_{DM}. The 15 additional variables for which the realisations were observed post hoc were used to test the performance of C_{DM} and M_{DM}. Coefficient $c_1 = 1$ has been used for the moment model. In Tables C.1, C.2 and C.3 the results are given for resp. $r_2 = 0.1$, 0.5 and 0.9. The results for both the moment and the classical model linear pool are summarised in Table C.4.

Table C.1: Comparison of MM and CM for prime rent data ($\boldsymbol{r_2 = 0.1}$)

	M_{DM}	C_{DM}	Eq. weights	Exp. 1	Exp. 2	Exp. 3	Exp. 4	Exp. 5
Weights								
Expert 1	0.235	1	0.2	1	0	0	0	0
Expert 2	0.161	0	0.2	0	1	0	0	0
Expert 3	0.000	0	0.2	0	0	1	0	0
Expert 4	0.461	0	0.2	0	0	0	1	0
Expert 5	0.143	0	0.2	0	0	0	0	1
Moment model score								
16 original seed variables								
penalty ϕ_1	*2931*	*8887*	*6736*	*8887*	*10301*	*61738*	*4557*	*10646*
15 performance variables								
penalty ϕ_1	**1749**	**2009**	**3180**	**2009**	**10073**	**22972**	**4933**	**5937**
Classical model score								
16 original seed variables								
calibration	0.3053	0.3305	0.0561	0.3305	0.1472	0.0201	0.0001	0.0042
information	0.4972	0.8572	0.1790	0.8572	0.9554	0.1556	1.5357	0.6126
score	*0.1518*	*0.2833*	*0.0100*	*0.2833*	*0.1407*	*0.0031*	*0.0001*	*0.0026*
15 performance variables								
calibration	0.1824	0.3579	0.1824	0.3579	0.0000	0.0006	0.0390	0.0390
information	0.4801	0.6724	0.1674	0.6724	0.7465	0.1641	1.3837	0.9623
score	**0.0875**	**0.2406**	**0.0305**	**0.2406**	**0.0000**	**0.0001**	**0.0540**	**0.0375**

M_{DM}, C_{DM}: performance based linear pools using resp. moment and classical model.

Table C.2: Comparison of MM and CM for prime rent data ($\boldsymbol{r_2 = 0.5}$)

	M_{DM}	C_{DM}	Eq. weights	Exp. 1	Exp. 2	Exp. 3	Exp. 4	Exp. 5
Weights								
Expert 1	0.143	1	0.2	1	0	0	0	0
Expert 2	0.172	0	0.2	0	1	0	0	0
Expert 3	0	0	0.2	0	0	1	0	0
Expert 4	0.537	0	0.2	0	0	0	1	0
Expert 5	0.148	0	0.2	0	0	0	0	1
Moment model score								
16 original seed variables								
penalty ϕ_1	*5257*	*18001*	*12518*	*18001*	*17262*	*116911*	*8157*	*17863*
15 performance variables								
penalty ϕ_1	**2756**	**2887**	**4680**	**2887**	**15261**	**34616**	**7006**	**9448**
Classical model score								
16 original seed variables								
calibration	0.3053	0.3305	0,0561	0.3305	0.1472	0.0201	0.0001	0.0042
information	0.5099	0.8572	0.1790	0.8572	0.9554	0.1556	1.5357	0.6126
score	*0.1557*	*0.2833*	*0.0100*	*0.2833*	*0.1407*	*0.0031*	*0.0001*	*0.0026*
15 performance variables								
calibration	0.2880	0.3579	0.1824	0.3579	0.0000	0.0006	0.0390	0.0390
information	0.5026	0.6724	0.1674	0.6724	0.7465	0.1641	1.3837	0.9623
score	**0.1448**	**0.2406**	**0.0305**	**0.2406**	**0.0000**	**0.0001**	**0.054**0	**0.0375**

M_{DM}, C_{DM}: performance based linear pools using resp. moment and classical model.

Table C.3: Comparison of MM and CM for prime rent data ($\boldsymbol{r_2 = 0.9}$)

	M_{DM}	C_{DM}	Eq. weights	Exp. 1	Exp. 2	Exp. 3	Exp. 4	Exp. 5
Weights								
Expert 1	0.000	1	0.2	1	0	0	0	0
Expert 2	0.208	0	0.2	0	1	0	0	0
Expert 3	0.000	0	0.2	0	0	1	0	0
Expert 4	0.614	0	0.2	0	0	0	1	0
Expert 5	0.178	0	0.2	0	0	0	0	1
Moment model score								
16 original seed variables								
penalty ϕ_1	*24705*	*100667*	*64966*	*100667*	*80403*	*617347*	*40810*	*83328*
15 performance variables								
penalty ϕ_1	**11686**	**10851**	**18289**	**10851**	**62315**	**140222**	**25807**	**41296**
Classical model score								
16 original seed variables								
calibration	0.3378	0.3305	0.0561	0.3305	0.1472	0.0201	0.0001	0.0042
information	0.5688	0.8572	0.1790	0.8572	0.9554	0.1556	1.5357	0.6126
score	*0.1921*	*0.2833*	*0.0100*	*0.2833*	*0.1407*	*0.0031*	*0.0001*	*0.0026*
15 performance variables								
calibration	0.4314	0.3579	0.1824	0.3579	0.0000	0.0006	0.0390	0.0390
information	0.5747	0.6724	0.1674	0.6724	0.7465	0.1641	1.3837	0.9623
score	**0.2479**	**0.2406**	**0.0305**	**0.2406**	**0.0000**	**0.0001**	**0.0540**	**0.0375**

M_{DM}, C_{DM}: performance based linear pools using resp. moment and classical model.

Table C.4: Summary of Comparison of MM and CM for prime rent data

	$r_2 = 0.1$		$r_2 = 0.5$		$r_2 = 0.9$	
	M_{DM}	C_{DM}	M_{DM}	C_{DM}	M_{DM}	C_{DM}
Moment model score						
16 original seed variables						
penalty ϕ_1	*2931*	*8887*	*5257*	*18001*	*24705*	*100667*
15 performance variables						
penalty ϕ_1	**1749**	**2009**	**2756**	**2887**	**11686**	**10851**
Classical model score						
16 original seed variables						
calibration	0.3053	0.3305	0.3053	0.3305	0.3378	0.3305
information	0.4972	0.8572	0.5099	0.8572	0.5688	0.8572
score	*0.1518*	*0.2833*	*0.1557*	*0.2833*	*0.1921*	*0.2833*
15 performance variables						
calibration	0.1824	0.3579	0.2880	0.3579	0.4314	0.3579
information	0.4801	0.6724	0.5026	0.6724	0.5747	0.6724
score	**0.0875**	**0.2406**	**0.1448**	**0.2406**	**0.2479**	**0.2406**

C.2 Case 2: Dikering Safety

The dikering safety case (Van Elst 1997) involves 17 experts assessing 47 seed variables. Seven seed variables were judged to be of a different scale and left out of the current evaluation. These are the variables with the identifiers: 'mod6', 'mod7', 'mod8', 'mod9', 'mod10', 'mod11' and 'mod12'. The first 20 seed variables of the 40 left we used to derive the performance based weights for both the classical model, C_{DM}, and the moment model, M_{DM}. The remaining 20 were used to test the performance of the classical and moment model linear pools. Coefficient $c_1 = 1000$ has been used for the moment model. The moment model results in the same DM weights for $r_2 = 0.1$, 0.5 and 0.9. The results for $r_2 = 0.5$ are displayed in Tables C.5 and C.6.

Table C.5: Comparison of MM and CM for dikering safety data, for exp. 1-7. ($\boldsymbol{r_2 = 0.5}$)

	M_{DM}	C_{DM}	Eq. weights	Exp. 1	Exp. 2	Exp. 3	Exp. 4	Exp. 5	Exp. 6	Exp.7
Weights										
Expert 1	1	0	1/17	1	0	0	0	0	0	0
Expert 2	0	0.5	1/17	0	1	0	0	0	0	0
Expert 3	0	0	1/17	0	0	1	0	0	0	0
Expert 4	0	0	1/17	0	0	0	1	0	0	0
Expert 5	0	0	1/17	0	0	0	0	1	0	0
Expert 6	0	0	1/17	0	0	0	0	0	1	0
Expert 7	0	0	1/17	0	0	0	0	0	0	1
Expert 8	0	0	1/17	0	0	0	0	0	0	0
Expert 9	0	0	1/17	0	0	0	0	0	0	0
Expert 10	0	0.5	1/17	0	0	0	0	0	0	0
Expert 11	0	0	1/17	0	0	0	0	0	0	0
Expert 12	0	0	1/17	0	0	0	0	0	0	0
Expert 13	0	0	1/17	0	0	0	0	0	0	0
Expert 14	0	0	1/17	0	0	0	0	0	0	0
Expert 15	0	0	1/17	0	0	0	0	0	0	0
Expert 16	0	0	1/17	0	0	0	0	0	0	0
Expert 17	0	0	1/17	0	0	0	0	0	0	0
Moment model score										
20 seed variables										
penalty ϕ_1	*3356*	*4906*	*4868*	*3356*	*4722*	*4859*	*5015*	*5115*	*5120*	*5103*
20 performance variables										
penalty ϕ_1	**4506**	**2642**	**33718**	**4506**	**3455**	**4583**	**3376**	**16866**	**11447**	**142282**
Classical model score										
20 seed variables										
calibration	1.19E-08	2.16E-02	6.67E-04	1.19E-08	2.16E-02	6.67E-04	8.81E-09	1.07E-17	7.65E-03	3.57E-15
information	1.2031	0.3356	0.4519	1.2031	0.3356	0.5167	0.9160	2.3397	0.0182	1.7215
score	*1.43E-08*	*7.26E-03*	*3.02E-04*	*1.43E-08*	*7.26E-03*	*3.45E-04*	*8.07E-09*	*2.51E-17*	*1.39E-04*	*6.15E-15*
20 performance variables										
calibration	0.2200	0.2200	0.1704	0.2200	0.0003	0.5505	0.0106	0.4280	0.0106	0.0106
information	0.7289	0.3356	0.4519	0.7289	1.5683	0.7351	1.2538	0.6011	0.3663	0.3187
score	**0.1604**	**0.0738**	**0.0770**	**0.1604**	**0.0005**	**0.4046**	**0.0132**	**0.2573**	**0.0039**	**0.0034**

M_{DM}, C_{DM}: performance based linear pools using resp. moment and classical model.

Table C.6: Comparison of MM and CM for dikering safety data, for exp. 8-17. ($\boldsymbol{r_2 = 0.5}$)

	Exp. 8	Exp. 9	Exp. 10	Exp. 11	Exp. 12	Exp. 13	Exp. 14	Exp. 15	Exp. 16	Exp. 17
Weights										
Expert 1	0	0	0	0	0	0	0	0	0	0
Expert 2	0	0	0	0	0	0	0	0	0	0
Expert 3	0	0	0	0	0	0	0	0	0	0
Expert 4	0	0	0	0	0	0	0	0	0	0
Expert 5	0	0	0	0	0	0	0	0	0	0
Expert 6	0	0	0	0	0	0	0	0	0	0
Expert 7	0	0	0	0	0	0	0	0	0	0
Expert 8	1	0	0	0	0	0	0	0	0	0
Expert 9	0	1	0	0	0	0	0	0	0	0
Expert 10	0	0	1	0	0	0	0	0	0	0
Expert 11	0	0	0	1	0	0	0	0	0	0
Expert 12	0	0	0	0	1	0	0	0	0	0
Expert 13	0	0	0	0	0	1	0	0	0	0
Expert 14	0	0	0	0	0	0	1	0	0	0
Expert 15	0	0	0	0	0	0	0	1	0	0
Expert 16	0	0	0	0	0	0	0	0	1	0
Expert 17	0	0	0	0	0	0	0	0	0	1
Moment model score										
20 seed variables										
penalty ϕ_1	*5052*	*5103*	*4722*	*5115*	*4464*	*5109*	*5103*	*5103*	*4970*	*5103*
20 performance variables										
penalty ϕ_1	**4940**	**2106**	**54800**	**3139**	**3347**	**16995**	**8302**	**332281**	**332445**	**332234**
Classical model score										
20 seed variables										
calibration	8.81E-09	3.57E-15	2.16E-02	1.07E-17	4.92E-17	3.76E-15	3.57E-15	3.57E-15	1.41E-07	3.57E-15
information	1.1100	1.7215	0.3356	2.3397	2.1895	1.9200	1.7215	1.7215	0.7592	1.7215
score	*9.78E-09*	*6.15E-15*	*7.26E-03*	*2.51E-17*	*1.08E-16*	*7.23E-15*	*6.15E-15*	*6.15E-15*	*1.07E-07*	*6.15E-15*
20 performance variables										
calibration	0.0106	0.2392	0.2200	0.0000	0.0000	0.4280	0.0000	0.2200	0.2200	0.2200
information	0.7963	0.7744	0.6332	1.8366	2.3068	0.9305	2.0388	0.6364	0.5691	0.6550
score	**0.0084**	**0.1853**	**0.1393**	**0.0000**	**0.0000**	**0.3983**	**0.0000**	**0.1400**	**0.1252**	**0.1441**

M_{DM}, C_{DM}: performance based linear pools using resp. moment and classical model.

C.3 Case 3: Thermal Comfort in Buildings

The thermal comfort in buildings case (De Wit 2001) involves 6 experts assessing 48 seed variables. The first 24 seed variables we used to derive the performance based weights for both the classical model, C_{DM}, and the moment model, M_{DM}. The last 24 were used to test the performance of the classical and moment model linear pools. Coefficient $c_1 = 1000$ has been used for the moment model. The moment model results in the same DM weights for $r_2 = 0.1$, 0.5 and 0.9. The results for $r_2 = 0.5$ are displayed in Table C.7.

Table C.7: Comparison of MM and CM for thermal comfort in buildings data ($\boldsymbol{r_2 = 0.5}$)

	M_{DM}	C_{DM}	Eq. weights	Exp. 1	Exp. 2	Exp. 3	Exp. 4	Exp. 5	Exp. 6
Weights									
Expert 1	0	0	1/6	1	0	0	0	0	0
Expert 2	0	0	1/6	0	1	0	0	0	0
Expert 3	0	0	1/6	0	0	1	0	0	0
Expert 4	1	0	1/6	0	0	0	1	0	0
Expert 5	0	0	1/6	0	0	0	0	1	0
Expert 6	0	1	1/6	0	0	0	0	0	1
Moment model score									
13 seed variables									
penalty ϕ_1	*1773*	*6994*	*2673*	*3872*	*3222*	*5534*	*1773*	*3430*	*6994*
13 performance variables									
penalty ϕ_1	**2284**	**6263**	**2752**	**4344**	**3076**	**5743**	**2284**	**3593**	**6263**
Classical model score									
13 seed variables									
calibration	0.0000	0.1816	0.0008	0.0003	0.1203	0.0003	0.0000	0.0000	0.181633
information	0.1650	0.7305	0.1421	0.6449	0.5527	0.4630	0.1650	0.8092	0.730458
score	*0.0000*	*0.1327*	*0.0001*	*0.0002*	*0.0665*	*0.0002*	*0.0000*	*0.0000*	*0.132676*
13 performance variables									
calibration	2.41E-08	1.04E-03	2.41E-08	9.85E-06	1.60E-02	3.27E-08	2.41E-08	1.39E-11	1.04E-03
information	0.1650	0.7305	0.1421	0.6449	0.5527	0.4630	0.1650	0.8092	0.730458
score	**3.97E-09**	**7.61E-04**	**3.42E-09**	**6.35E-06**	**8.84E-03**	**1.51E-08**	**3.97E-09**	**1.13E-11**	**7.61E-04**

M_{DM}, C_{DM}: performance based linear pools using resp. moment and classical model.

C.4 Case 4: Radionuclide Transport in Soils

The radionuclide transport in soil case is a study from the Joint EU-USNRC Project on Uncertainty Analysis of Probabilistic Accident Consequence Codes (Harper, Goossens, Cooke, Hora, Young, Psler-Sauer, Miller, Kraan, Lui, McKay, Helton & Jones 1995). These codes estimate the risks and other endpoints associated with accidents from hypothesised nuclear installations. The case consists of 31 seed variables, assessed by 4 experts. Five seed variables were judged to be of a different scale and left out of the current evaluation. These are the variables with the identifiers: 'S2_RU_CS_CA', 'S2_RU_CS_PO', 'S2_RU_CS_BA', 'S2B_CR_CS_SS' and 'S2B_CR_CS_LS'. The first 13 seed variables of the 26 left we used to derive the performance based weights for both the classical model, C_{DM}, and the moment model, M_{DM}. The remaining 13 were used to test the performance of the classical and moment model linear pools. Coefficient $c_1 = 10000$ has been used for the moment model. In Tables C.8, C.9 and C.10 the results are given for resp. $r_2 = 0.1$, 0.5 and 0.9. The results for both the moment and the classical model linear pool are summarised in Table C.11.

Table C.8: Comparison of MM and CM for radionuclide transport in soil data ($\boldsymbol{r_2 = 0.1}$)

	M_{DM}	C_{DM}	Eq. weights	Exp. 1	Exp. 2	Exp. 3	Exp. 4
Weights							
Expert 1	0.1649	0	0.25	1	0	0	0
Expert 2	0.1957	0	0.25	0	1	0	0
Expert 3	0.3438	0.4614	0.25	0	0	1	0
Expert 4	0.2956	0.5386	0.25	0	0	0	1
Moment model score							
13 seed variables							
penalty ϕ_1	*1790*	*2564*	*1904*	*4767*	*4487*	*3139*	*3577*
13 performance variables							
penalty ϕ_1	**15670**	**27590**	**14255**	**12542**	**9172**	**9034**	**66129**
Classical model score							
13 seed variables							
calibration	0.0195	0.1431	0.0195	0.0000	0.00	0.0102	0.0074
information	0.1538	0.2971	0.1509	1.7422	0.5656	0.5720	0.9141
score	*0.0030*	*0.0425*	*0.0029*	*0.0000*	*0.0000*	*0.0058*	*0.0068*
13 performance variables							
calibration	3.27E-05	3.27E-05	3.27E-05	6.04E-13	1.46E-07	1.15E-04	2.43E-07
information	0.2352	0.3319	0.2485	1.5642	0.4731	0.6025	0.7910
score	**7.68E-06**	**1.08E-05**	**8.12E-06**	**9.45E-13**	**6.92E-08**	**6.95E-05**	**1.92E-07**

M_{DM}, C_{DM}: performance based linear pools using resp. moment and classical model.

Table C.9: Comparison of MM and CM for radionuclide transport in soil data ($\boldsymbol{r_2 = 0.5}$)

	M_{DM}	C_{DM}	Eq. weights	Exp. 1	Exp. 2	Exp. 3	Exp. 4
Weights							
Expert 1	0.25	0	0.25	1	0	0	0
Expert 2	0.25	0	0.25	0	1	0	0
Expert 3	0.25	0.4614	0.25	0	0	1	0
Expert 4	0.25	0.5386	0.25	0	0	0	1
Moment model score							
13 seed variables							
penalty ϕ_1	*3788*	*5350*	*3788*	*7504*	*7056*	*6822*	*7495*
13 performance variables							
penalty ϕ_1	**41647**	**126812**	**41647**	**21322**	**15731**	**14840**	**379020**
Classical model score							
13 seed variables							
calibration	0.0195	0.1431	0.0195	0.0000	0.00	0.0102	0.0074
information	0.1509	0.2971	0.1509	1.7422	0.5656	0.5720	0.9141
score	*0.0029*	*0.0425*	*0.0029*	*0.0000*	*0.0000*	*0.0058*	*0.0068*
13 performance variables							
calibration	3.27E-05	3.27E-05	3.27E-05	6.04E-13	1.46E-07	1.15E-04	2.43E-07
information	0.2485	0.3319	0.2485	1.5642	0.4731	0.6025	0.7910
score	**8.12E-06**	**1.08E-05**	**8.12E-06**	**9.45E-13**	**6.92E-08**	**6.95E-05**	**1.92E-07**

M_{DM}, C_{DM}: performance based linear pools using resp. moment and classical model.

Table C.10: Comparison of MM and CM for radionuclide transport in soil data ($\boldsymbol{r_2 = 0.9}$)

	M_{DM}	C_{DM}	Eq. weights	Exp. 1	Exp. 2	Exp. 3	Exp. 4
Weights							
Expert 1	0.2589	0	0.25	1	0	0	0
Expert 2	0.2631	0	0.25	0	1	0	0
Expert 3	0.2420	0.4614	0.25	0	0	1	0
Expert 4	0.2360	0.5386	0.25	0	0	0	1
Moment model score							
13 seed variables							
penalty ϕ_1	*21609*	*31731*	*21627*	*33422*	*31386*	*41685*	*44582*
13 performance variables							
penalty ϕ_1	**277537**	**1066189**	**300979**	**104443**	**77828**	**69806**	**3341316**
Classical model score							
13 seed variables							
calibration	0.0195	0.1431	0.0195	0.0000	0.00	0.0102	0.0074
information	0.1525	0.2971	0.1509	1.7422	0.5656	0.5720	0.9141
score	*0.0030*	*0.0425*	*0.0029*	*0.0000*	*0.0000*	*0.0058*	*0.0068*
13 performance variables							
calibration	3.27E-05	3.27E-05	3.27E-05	6.04E-13	1.46E-07	1.15E-04	2.43E-07
information	0.2483	0.3319	0.2485	1.5642	0.4731	0.6025	0.7910
score	**8.11E-06**	**1.08E-05**	**8.12E-06**	**9.45E-13**	**6.92E-08**	**6.95E-05**	**1.92E-07**

M_{DM}, C_{DM}: performance based linear pools using resp. moment and classical model.

Table C.11: Summary of Comparison of MM and CM for radionuclide transport in soil data

	$r_2 = 0.1$		$r_2 = 0.5$		$r_2 = 0.9$	
	M_{DM}	C_{DM}	M_{DM}	C_{DM}	M_{DM}	C_{DM}
Moment model score						
13 seed variables						
penalty ϕ_1	*1790*	*2564*	*3788*	*5350*	*21609*	*31731*
13 performance variables						
penalty ϕ_1	**15670**	**27590**	**41647**	**126812**	**277537**	**1066189**
Classical model score						
13 seed variables						
calibration	0.0195	0.1431	0.0195	0.1431	0.0195	0.1431
information	0.1538	0.2971	0.1509	0.2971	0.1525	0.2971
score	*0.0030*	*0.0425*	*0.0029*	*0.0425*	*0.0030*	*0.0425*
13 performance variables						
calibration	3.27E-05	3.27E-05	3.27E-05	3.27E-05	3.27E-05	3.27E-05
information	0.2352	0.3319	0.2485	0.3319	0.2483	0.3319
score	**7.68E-06**	**1.08E-05**	**8.12E-06**	**1.08E-05**	**8.11E-06**	**1.08E-05**

C.5 Case 5: Atmospheric Deposition

The atmospheric deposition application of the classical model was conducted as a pilot study for the Joint EU-USNRC Project on Uncertainty Analysis of Probabilistic Accident Consequence Codes (Harper et al. 1995). The application involved 4 experts assessing 24 seed variables. Three seed variables were judged to be of a different scale and left out of the current evaluation. These are the variables with the identifiers: 'el. Iod. Trees', '1-2 walls' and '3-4 trees'. The first 11 seed variables of the 21 left we used to derive the performance based weights for both the classical model, C_{DM}, and the moment model, M_{DM}. The remaining 10 were used to test the performance of the classical and moment model linear pools. Coefficient $c_1 = 10^8$ has been used for the moment model. The moment model results in the same DM weights for $r_2 = 0.1$, 0.5 and 0.9. The results for $r_2 = 0.5$ are displayed in Table C.12.

Table C.12: Comparison of MM and CM for atmospheric deposition data ($\boldsymbol{r_2 = 0.5}$)

	M_{DM}	C_{DM}	Eq. weights	Exp. 1	Exp. 2	Exp. 3	Exp. 4
Weights							
Expert 1	0	0	0.25	1	0	0	0
Expert 2	0	1	0.25	0	1	0	0
Expert 3	0	0	0.25	0	0	1	0
Expert 4	1	0	0.25	0	0	0	1
Moment model score							
11 seed variables							
penalty ϕ_1	*1670*	*17986*	*92284*	*831245*	*17986*	*76614*	*1670*
10 performance variables							
penalty ϕ_1	**802615**	**42764**	**110402**	**22140**	**42764**	**14298**	**802615**
Classical model score							
11 seed variables							
calibration	0.0011	0.8525	0.3696	0.3696	0.85	0.3696	0.0011
information	0.6257	0.6138	0.1158	0.3473	0.6138	0.3499	0.6257
score	*0.0007*	*0.5233*	*0.0428*	*0.1283*	*0.5233*	*0.1293*	*0.0007*
10 performance variables							
calibration	0.2441	0.2281	0.4926	0.6828	0.2281	0.4926	0.2441
information	0.4759	0.2787	0.0749	0.1572	0.2787	0.2929	0.4759
score	**0.1162**	**0.0636**	**0.0369**	**0.1073**	**0.0636**	**0.1443**	**0.1162**

M_{DM}, C_{DM}: performance based linear pools using resp. moment and classical model.