Expert judgement for dependence in probabilistic modelling: A systematic literature review and future research directions

Werner, Christoph; Bedford, Tim; Cooke, Roger M.; Hanea, Anca; Morales Napoles, Oswaldo

DOI
10.1016/j.ejor.2016.10.018

Publication date
2017

Document Version
Final published version

Published in
European Journal of Operational Research

European Journal of Operational Research 258 (2017) 801–819

Contents lists available at ScienceDirect

European Journal of Operational Research

journal homepage: www.elsevier.com/locate/ejor

Invited Review

Expert judgement for dependence in probabilistic modelling: A systematic literature review and future research directions

Christoph Werner[a,*], Tim Bedford[a], Roger M. Cooke[b], Anca M. Hanea[c], Oswaldo Morales-Nápoles[d]

[a] Department of Management Science, University of Strathclyde, Glasgow, United Kingdom
[b] Resources for the Future, Washington, DC, USA
[c] Centre of Excellence for Biosecurity Risk Analysis, University of Melbourne, Melbourne, Australia
[d] Faculty of Civil Engineering and Geosciences, Delft University of Technology, Delft, The Netherlands

ARTICLE INFO

Article history:
Received 17 March 2016
Accepted 9 October 2016
Available online 22 October 2016

Keywords:
Risk analysis
Uncertainty modelling
Dependence elicitation
Structured expert judgement
Dependence modelling

ABSTRACT

Many applications in decision making under uncertainty and probabilistic risk assessment require the assessment of multiple, dependent uncertain quantities, so that in addition to marginal distributions, interdependence needs to be modelled in order to properly understand the overall risk. Nevertheless, relevant historical data on dependence information are often not available or simply too costly to obtain. In this case, the only sensible option is to elicit this uncertainty through the use of expert judgements. In expert judgement studies, a structured approach to eliciting variables of interest is desirable so that their assessment is methodologically robust. One of the key decisions during the elicitation process is the form in which the uncertainties are elicited. This choice is subject to various, potentially conflicting, desiderata related to e.g. modelling convenience, coherence between elicitation parameters and the model, combining judgements, and the assessment burden for the experts. While extensive and systematic guidance to address these considerations exists for single variable uncertainty elicitation, for higher dimensions very little such guidance is available. Therefore, this paper offers a systematic review of the current literature on eliciting dependence. The literature on the elicitation of dependence parameters such as correlations is presented alongside commonly used dependence models and experience from case studies. From this, guidance about the strategy for dependence assessment is given and gaps in the existing research are identified to determine future directions for structured methods to elicit dependence.

1. Introduction

In decision making under uncertainty it is vital that dependencies between uncertain variables are appropriately modelled, as otherwise the model may not be fit for purpose. Dependent uncertainty may arise either directly because variables in the model are correlated, or indirectly when an uncertainty analysis of model parameters is carried out to explore model robustness. Both cases exhibit complex interrelations and dependencies which need to be considered if assumptions such as independence are not justifiable.

However, it is often not straightforward to either model or quantify dependence. In particular whenever no relevant historical data are available, the only sensible way to achieve uncertainty quantification is through eliciting expert judgements. When performed rigorously, the elicited quantities, often aggregated from multiple experts, offer reliable information for model quantification. Nevertheless, there are several different broad approaches and many choices to be made by the analyst, all of which can affect the elicitation burden for experts and ultimately also the reliability of the outcome.

While research and reviews that offer guidance exist for methods addressing the elicitation of univariate quantities (Cooke, 1991; European Food and Safety Authority (EFSA), 2014; French, 2011; Jenkinson, 2005; O'Hagan et al., 2006; Ouchi, 2004), and while dependence modelling is an active research area (Kurowicka & Cooke, 2006), little guidance exists about the elicitation of dependencies. The exceptions are Bayesian (Belief) nets (BNs), though also for these modelling and elicitation challenges remain, as shown later. In fact, developing defensible elicitation processes for multivariate quantities is still much under development despite its fundamental importance for decision as well as risk

* Corresponding author.
E-mail addresses: christoph.werner@strath.ac.uk (C. Werner), tim.bedford@strath.ac.uk (T. Bedford), Cooke@rff.org (R.M. Cooke), anca.hanea@unimelb.edu.au (A.M. Hanea), O.MoralesNapoles@tudelft.nl (O. Morales-Nápoles).

http://dx.doi.org/10.1016/j.ejor.2016.10.018
0377-2217/

analysis (Moskowitz & Bunn, 1987; Smith & Von Winterfeldt, 2004). Some of the first studies that elicit dependence are Cooke and Kraan (1996), Keeney and von Winterfeldt (1991), Kunda and Nisbett (1986), Gokhale and Press (1982) and Kadane, Dickey, Winkler, Smith, and Peters (1980). Since then more ways for quantifying multivariate distributions and models through experts have been investigated, yet on the actual elicitation only little discussion and guidance is available. References that introduce some aspects are Daneshkhah and Oakley (2010), Kurowicka and Cooke (2006), O'Hagan et al. (2006) and Garthwaite, Kadane, and O'Hagan (2005). However, a complete and systematic way of comparing different dependence parameters as elicited quantities, and reflecting their use in dependence models in the form of a literature review has been non-existent so far. Therefore, research and applications of several dependence measures in models and their elicitation methods are presented. With a practical focus, case studies are discussed whenever available. This paper addresses elicitation processes for dependence information and aims at providing understanding of their use in applications. It offers guidance on making robust choices about which summary of expert knowledge on multivariate distributions should be elicited, and how they might be used within a dependence modelling context, as these are key decisions within the overall elicitation process. This is achieved by outlining how much is understood about the complexity of approaches to dependence modelling and the cognitive assessment burden for experts.

Throughout this paper we use the word "dependence" in a general sense (in contrast to specific association measures) to refer to situations where there are multiple uncertain quantities and gaining information about one would change uncertainty assessments for some others. More formally, two unknown quantities X and Y, are independent (for me) if I do not change my beliefs about X when given information about Y. For higher dimensions I regard all quantities independent of one another if knowledge of one group of variables does not change my belief about other variables. Dependence is simply the absence of independence. It is a property of an expert's belief about the quantities. This definition relates to Lad (1996) who reminds us that in a subjective probability context one expert's (in-) dependence assessment might not be shared with another expert possessing a different state of knowledge.

The definition of dependence as we use it here relates directly to the scope of this review. A first comment on the scope is that the word "dependence" is used in many ways within Operational Research (OR) and related fields, and it is worth clarifying how its use here differs from its meaning in other OR contexts. The underlying framework adopted is that of subjective probability (as aforementioned), which plays a key role within expected utility maximisation for decision making. Dependence then, refers to the way we model and assess the probability dependence structure required for such decision support processes. We do not consider non-probabilistic representations of uncertainty, nor do we consider approaches to represent dependence between criteria used to model the preferences of the decision maker as discussed widely in the multi-criteria decision analysis (MCDA) literature.

The foundations of subjective probability are drawn from a wide literature, in which Savage (1954) provides one of the most sophisticated accounts. In this account, probabilities can be assessed through preferences over lotteries, and there are implied consistency rules for preferences which can be empirically validated. It is well known that there is a distinction between normative and empirical validation, so the degree to which researchers choose to be led by normative or empirical consistency has led to many different approaches. For instance, Dubois, Prade, and Sabbadin (2001) provide a theoretical framework which attempts to tie these strands together in the context of possibility theory, and the implications of this are discussed in detail by Cooke (2004).

The modelling of dependence between attributes in MCDA is the subject of a wide literature, and as discussed above, is outside the scope of this review. Facilitative approaches within multi-attribute utility theory provide a variety of models, for which (whenever possible) problem structuring is used to ensure preference independence (Von Winterfeldt & Fasolo, 2009; Wallenius et al., 2008), while other approaches have been inspired by issues such as assessing the range of preferences within a stakeholder group (Flari, Chaudhry, Neslo, & Cooke, 2011; Neslo & Cooke, 2011), or trying to model preferences based on a limited number of attributes or limited resolution of attribute measurement. For the latter, in particular interaction among criteria in complex systems and dependence of attributes is modelled. This is done for instance to assess the aggregated importance of correlated criteria or further investigate dependent attributes for predictive modelling. Common methods in the OR literature are: non-additive aggregation models such as Choquet and Sugeno integrals (Angilella, Greco, Lamantia, & Matarazzo, 2004; Grabisch, 1996; Marichal, 2004), Robust Ordinal Regression (Figueira, Greco, & Słowiński, 2009; Greco, Mousseau, & Słowiński, 2014) and (Dominance-Based) Rough Set Approaches which use decision rules in the form of *if* [condition] *then* [consequent] (Błaszczyński, Greco, & Słowiński, 2007; Greco, Matarazzo, & Słowiński, 2001; 2004).Another interesting approach in this regard is Abbas (2009) who constructs a multi-attribute utility function through a copula, a dependence model that is introduced later for modelling probabilistic dependence. A frequently considered empirical area for MCDA-based approaches is financial portfolio optimisation (Ehrgott, Klamroth, & Schwehm, 2004).

A last comment on the scope is that while we discuss the cognitive complexity of assessing dependence in various ways, such as already considered by Kruskal (1958), and while insights from psychological studies are mentioned, corresponding research streams for causal and association judgements are not reviewed exhaustively. Normative and descriptive models for causal reasoning or mental conceptualisation of correlations, which origin is often attributed to Smedslund (1963), are found for instance in Mitchell, De Houwer, and Lovibond (2009), Gredebäck, Winman, and Juslin (2000), Beyth-Marom (1982) and Allan (1980). An overview and introduction to these areas is given in Hastie (2016) and Shanks (2004).

The paper is organised as follows. Section 2 discusses the extent to which findings from eliciting univariate quantities apply to the elicitation of multivariate ones in order to provide the reader with an indication for the scope of the overall topic. Section 3 introduces the modelling context which shows how modelling and eliciting dependence are related. This offers an overall structure to the research problem. Then, Section 4 discusses how elicitation is approached for quantifying various dependence models. Section 5 presents dependence parameters that are commonly elicited together with its implications for experts' assessment burden before Section 6 briefly reviews findings on mathematical aggregation of dependence assessments. Section 7 provides an overview of the empirical contributions in the literature based on which Section 8 formulates directions for future research and concludes the paper. We refer to Appendix B (Supplementary material) whenever a technical term needs a more detailed explanation, however the original references should be considered for an extended introduction.

2. Generalisations of univariate elicitation processes for eliciting dependence

Structured processes for the elicitation of dependence follow historically from findings made when eliciting univariate quantities. In the early days of uncertainty modelling, formal processes

for eliciting univariate uncertainties, such as marginal probabilities, were developed to ensure a methodologically robust approach to parameter quantification in the face of lacking relevant historical data. From these, methods to elicit dependence followed given the need of accounting for relationships between uncertainties. Cooke (2013) discusses the historical development of expert judgement in uncertainty analysis and its achievements in more detail.

This development is not surprising as univariate quantities are (typically) more intuitive to experts and their specification is required (at least implicitly) prior to eliciting dependent distributions for two or more uncertain quantities.

In this section we discuss some main foci of structured expert judgement studies and evaluate the extent to which findings for univariate quantities are generalisable in the multivariate case. This discussion outlines where in a process adjustments are necessary when eliciting multivariate uncertainty and therefore provides an indication for the scope of dependence elicitation. Given the overall focus of the paper, we outline only the relevant considerations for the elicited dependence parameters and the aggregation of judgements. However, it should be noted that an elicitation process is much more complex and other decisions in it, such as how to design the statistical training for experts prior to an elicitation, might vary as well considerably when eliciting multivariate uncertainty.

Already the earliest expert judgement studies for univariate quantities have shown that assessment outcomes can differ greatly depending on the use of directly or indirectly elicited query formats (Spetzler & Stael von Holstein, 1975). As a result, an extensive literature on heuristics and biases is available on the matter of framing elicitation questions and choosing a form for the query variable. Further, recommendations are made on the theoretical suitability of the elicited format, e.g. objections are made to non-observable quantities (Kadane & Wolfson, 1998). For eliciting multivariate quantities on the other hand, the same conclusions are not readily applicable. As will be seen, the effect of direct and indirect elicitation approaches is less well-understood and findings are often conflicting. The objection to non-observable quantities is less clear and indeed we show later that eliciting non-observable quantities performs well in terms of empirical accuracy and mathematical coherence. Similarly, for heuristics and biases only some extensions for the multivariate case exist, such as "illusory correlation" (Chapman & Chapman, 1969), stemming from the availability bias, and "confusion of the inverse", originating with the representativeness bias (O'Hagan et al., 2006) (for both see Appendix B). While these findings indicate an overlap for the existence of common biases, a lack of empirical research on the effect of framing for multivariate elicitation does not allow for generalisable conclusions.

Once the dependence information has been elicited in the form of some dependence parameter (which is thoroughly addressed in the following sections), a well-researched topic for univariate uncertainty, which generalisation would be desirable for multivariate elicitation, is the use of scoring rules. Roughly, a scoring rule is a numerical evaluation of a probability assessment based on observations. In expert judgement studies, they are typically used for two reasons, first to present an incentive for truthful assessment and second to measure the quality of an assessment after the elicitation, usually to inform a weighted combination of the judgements. In other words, they are used to define desirable properties of the assessment itself and they serve as a reward structure when evaluating an assessment. While an incentive is given by using (strictly) proper scoring rules which ensure that experts achieve their maximal expected score if and only if stating their true belief, a main property of measuring the quality of an assessment is its calibration, i.e. the statistical accuracy after observing an event of interest. Suppose an expert provides a probability distribution

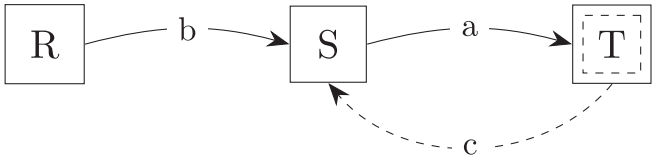

Fig. 1. Schematic representation of modelling and elicitation context.

P over a set of n mutually exclusive events i. Then, after observing the events of interest, we can construct the sample distribution S with $S(i)$ equal to the number of times that i is observed divided by n. While it appears reasonable to state at first thought that an expert is not well calibrated if $S \neq P$, this might be false if we suppose that true values represent independent samples from a random variable with distribution P. In this case, P relates to "reality" but we will never have $S = P$ due to statistical fluctuations. Loosely, an expert is therefore said to be well-calibrated if the true values of the uncertain quantities can be regarded as independent samples of a random variable with distribution P (Cooke, 1991).

When evaluating experts' performance, we have to distinguish between scoring rules for individual variables and scoring rules based on sets of assessments together with sets of realisations. The first, assigning scores to each individual assessment and summing these scores over a set of variables, is often suggested in the literature for the purpose of rewarding, yet it is not a sensible approach. A main issue is that the resulting scores cannot be interpreted in a meaningful way without knowing the number of quantities assessed and their overall sample distribution. This is due to the possible additive decomposition of these types of scores into a "calibration" and "resolution" term (DeGroot & Fienberg, 1983). Resolution measures how well experts partition the variables into statistically distinct categories while not considering whether the distributions assigned to these categories correspond to the experts' assessment. This becomes problematic when high resolution overpowers low statistical accuracy. A more detailed presentation of this drawback and some intuitive examples are given in Cooke (1991); 2014). Therefore, scoring rules for average probabilities are highly encouraged for evaluating and combining experts. While some main properties of scoring rules are applicable in the multivariate case, others cannot be readily used.

Jose, Nau, and Winkler (2009) discuss (for the univariate case) the inclusion of order information (requiring an ordered state space). Ordered events allow for rewarding that takes account of nearness to an event's realisation. In the multivariate case the lack of natural ordering means that this approach is not possible. Further, Jose, Nau, and Winkler (2008) discuss a wide class of scoring rules, called generalised divergence scores, that allow for any baseline distribution (rather than a uniform by default), and which reward according to a measure of distance between the assessed distribution and the baseline distribution. Of interest for multivariate elicitation is the derivation of a weighted scoring rule that is closely related to the *Hellinger distance* which is a measure of divergence that has been used in the calibration of experts' multivariate assessments (Section 6).

3. Guide to modelling and elicitation context

The main purpose of eliciting dependence is to quantify a multivariate stochastic model when this cannot be done wholly by conventional statistical estimation (which, in our view is a common situation). This section discusses broad approaches to dependence modelling in order to provide a clear structure for the next sections by highlighting the link between dependence modelling and expert judgement. Fig. 1 shows this general view on the modelling context with three different broad approaches to assessing

dependence and illustrates the relationships between model input and output variables.

In this general context, S represents the vector of stochastic variables in the model, and T the vector of output variables which depends deterministically on S. R represents another set of auxiliary variables used to evaluate the uncertainty on S. The solid arrows show deterministic relationships between the variables, and hence the direction in which uncertainty can be propagated.

It is not uncommon for there to be dependence between the output variables T. This can arise simply as a result of the functional dependence represented in arrow a, even if the stochastic variables in S are modelled as being stochastically independent. In many applications, however, it is not appropriate to model the variables in S as independent, and so we should find a way to model and assess dependence in S.

Approach a. In Approach (a) we model the dependence relations between the variables in S directly. The main techniques are BNs, copulas, parametric families of multivariate distributions (e.g. the multivariate Gaussian distribution), and Bayes Linear methods. We provide examples for each method in the next section. Having assessed the dependence and hence having specified the distribution of the variables in S, uncertainty is then propagated through the model (arrow a) to the output variable (or variables) T. As we shall see later, direct assessment of dependence on the variables S is most predominant in the literature. However, two other approaches are also important and worth discussing.

Approach b. In Approach (b) we introduce a new set of auxiliary variables R, which are not directly part of the model variables (though may in practice have some overlap with the variables S). The variables R are chosen so that their uncertainty is easier to quantify—in particular one might choose these variables so that they can be considered stochastically independent, with the dependence in the variables S arising as a result of the complex relationship between the "explanatory" variables R and those in S. This is shown in Fig. 1 as arrow b. This approach is of interest particularly when change of variables methods (frequently used in multivariate statistics) can be used to simplify the variable set from S. A common model type used in this context is a regression model and an example of introducing and assessing auxiliary variables is given in Section 4.2.

Approach c. In Approach (c) we "calibrate" the uncertainties on S through considering some set of output variables T on which the uncertainties can be assessed. Obviously, to be useful, this would have to be a different situation than the one in which the overall model is to be used (see dashed node inside T), as we would otherwise be simply directly assessing the uncertainty in the variables of interest. This calibration of uncertainties relies on the backwards propagation of uncertainty from T back to S, shown by arrow c. The dotted arrow is used to indicate a key difference with the solid arrows a and b. In general, more than one distribution on S will forward-propagate to the given distribution on T, that is, the inverse problem has no unique solution (or even worse, it has no solution). Other criteria (such as max entropy) are then used to select a unique inverse. That solution then defines a dependence structure on S, which can be propagated back through arrow a to look at other output contexts. This is called *Probabilistic Inversion* (PI) (Cooke, 1994; Kraan & Bedford, 2005; Kurowicka & Cooke, 2006) and we show an example in Section 4.3.

This approach is of interest when the dependence structure in S is difficult to determine directly, but must satisfy reasonable conditions on output variables that are easier to understand and hence easier to quantify.

A common theme in the latter two approaches is the model boundary. In both cases we choose to extend the model to include other input or output variables in addition to those which are strictly necessary for direct modelling. Indeed it may happen that the auxiliary variables represent simplifications of more complex issues which are insufficiently understood to be included explicitly in the model but which are known to collectively impact the behaviour of the system significantly. An example of this is the modelling of common cause events in risk analysis (Bedford & Cooke, 2001) where the range of underlying causes is too wide to be modelled individually, but which together have a substantial effect in inducing dependencies in the overall system behaviour.

We illustrate the dependence structures shown in Fig. 1 with the following simplified project risk management example which shows how choices can be made in the various modelling contexts. We are managing a project which has an overall cost (model output variable T). The cost is determined by individual activities with associated costs (variables in S) that are of importance for the project completion. If we want to model the stochastic dependence between activities in order to obtain information about the overall cost, a first option is to do so directly by specifying the dependencies directly between the cost elements. The dependence models used here are part of modelling context a. If modelling the dependence between the individual activities directly does not produce a satisfactory model output, we have the choice to include explanatory variables (R) that help us to understand the relationship better. For instance, we can include factors like environmental uncertainties if we belief that our project's activity costs are (partly) influenced by them. The techniques used here are part of modelling context b. Recall that we are choosing to extend the model which relates to the earlier discussion on the model boundary. The reason for modelling dependency in this way is because it may be easier to consider the impact of certain factors explicitly rather than implicitly when only using approach a. If the model output resulting from the inclusion of additional factors is still not satisfactory, we might choose to model some systemic impacts of the project. For instance, factors like the availability of qualified staff might be present and result in a subtle dependence relationship, leading to the distribution for the overall cost (the model output variables T) being incorrectly assessed. With methods used in c, we would have a separate assessment of the distribution (or at least for features of this distribution) for the overall cost which would lead to a changed model for the joint distribution of the activity costs (modelling context a or b). We could also consider modelling a more complex situation in which we manage several projects. In this case, the overall cost becomes multivariate instead of univariate (i.e. T becomes a vector of variables). Then, we can use methods (from c) that allow propagating our uncertainty from one project about which we have information backwards in order to make inference about the distribution of the activities (S) and hence the distribution for overall costs (T). The common objective is to find a good model for the uncertainties relating S and T. Conceptually, we can only ever specify part of the required information for this model, so that in practice our model is always underspecified (though this point is often not appreciated because modellers often adopt low-dimensional parametric families of models early on). Approaches b and c provide complementary approaches to specify further information about the model.

4. Dependence models and expert judgement

Before presenting and reviewing dependence parameters as elicited quantities explicitly, in this section we first discuss expert judgement for common dependence models. This includes main challenges when using experts to quantify models as well as the applicability of elicited forms for a satisfactory representation of

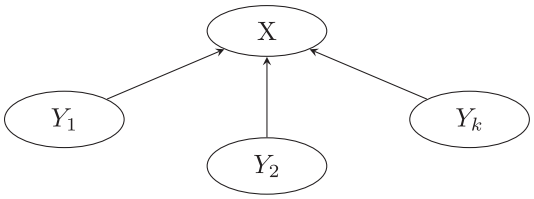

Fig. 2. Example Bayesian network with one child and three independent parents.

the experts' information in the model. We present the modelling aspects first given that decisions here precede and strongly affect the choice of which dependence parameter to elicit. In accordance with the earlier framework (see Fig. 1), BNs and copulas together with probabilistic and non-probabilistic parametric models are introduced for context (*a*), regression models for (*b*) and Probabilistic Inversion for (*c*).

4.1. Elicitation for direct modelling

4.1.1. Bayesian (belief) networks

In (*a*), a common way of integrating high dimensional uncertainty in a probabilistic model is by specifying a multivariate distribution for the random variables through the product of marginal and conditional probabilities. A common modelling framework is a BN (Darwiche, 2009; Pearl, 2000). A random variable is described by a node in the graph while arcs represent the qualitative dependence relationships amongst variables. The direct predecessors/successors of a node are called parents/children, and the BN is specified (for example) by determining for every child node its conditional probability distribution given the states of its parent nodes. Hence, it is composed of a directed acyclic graph with marginal distributions for source nodes and conditional distributions for child nodes given the parents. A simple example BN to be used throughout this review is shown in Fig. 2.

When using expert judgement, French (2011) views eliciting BNs as an obvious approach for obtaining dependence information. However, while more has been written about eliciting the qualitative dependence structure (the arrows in the BN) (Henrion, 1989; Nadkarni & Shenoy, 2004), eliciting dependence quantitatively has been recognised as a main issue when constructing BNs (Druzdel & Van Der Gaag, 2000; Renooij, 2001). Identified difficulties are the elicitation for high dimensional models and the assessment burden due to an exponentially growing number of probabilities to assess (in discrete BNs). Therefore, some alternative modelling approaches have been developed to be used in conjunction with expert judgement methods.

While in the low dimensional, discrete case, experts provide information in form of conditional probabilities to populate conditional probability tables, in higher dimensions this is intractable and too time-consuming. An alternative approach is to model continuous distributions and to elicit dependence information through (un-) conditional rank correlations. These models are known as non-parametric BNs for which a review of applications can be found in Hanea, Napoles, and Ababei (2015). For these, Morales Nápoles, Kurowicka, and Roelen (2008) developed a way of eliciting conditional exceedance probabilities for higher dimensions to derive the required rank correlations. This method is detailed in the next section when discussing elicited forms of dependence parameters explicitly.

In order to address the reduction of the assessment burden (in the discrete case), one way is to reduce the number of necessary assessments. For instance, Wisse, van Gosliga, van Elst, and Barros (2008) propose piecewise linear interpolation (see Appendix B) in order to reduce the overall number of required assessments for a full conditional probability table. Their method elicits conditional probabilities which are discussed in the next section as an elicited form. Another method that reduces the number of required assessments is through assumptions on the causal interpretation of a BN. The assumptions on the causal interpretation originate with noisy-OR gates (Pearl, 1988) which use an underlying parametric distribution that reduces necessary assessments logarithmically (see Appendix B). The functional OR relationship denotes how individual parent nodes are combined for a common effect and assumes that they are independent of each other with respect to their causal effect on the child nodes. Thus, the presence of one parent node suffices to produce an effect on the child independently of other parents (with a certain probability—hence noisy rather than deterministic). A leaky noisy-OR gate includes a background probability that represents the influence of non-modelled causes. From this, Zagorecki and Druzdzel (2004), building onto Druzdzel and Van Der Gaag (1995), introduce the elicitation of leaky and non-leaky noisy-OR parameters as alternatives to conditional probabilities. They use parameters introduced by Henrion (1989) and Diez (1993) and a potential framing (for the BN in Fig. 2) is:

> *"What is the probability that X is present when Y_1 is present and all other causes of X (addition for leaky case: including those not modelled explicitly) are absent?"*

In an experimental setting, Zagorecki and Druzdzel (2004) elicit leaky and non-leaky noisy-OR parameters together with conditional probabilities. An artificial dependence relation between three parents and one child node was determined (causes for anti-gravity of an unknown type of rock) and in a small simulation, participants could choose the influence (strength level of presence or absence) of each cause and observe what happens as an effect (anti-gravity or not). Then they assessed the conditional probability distribution with each assessment method, i.e. non-leaky and leaky noisy-OR parameters and a direct conditional probability assessment. The leaky noisy-OR parameter was assessed as most accurate (in terms of Euclidean distance to empirical distribution) while conditional probability was found least accurate. The authors claim that with an increasing number of nodes their method offers a clear advantage over conditional probability elicitation as the latter will become unmanageable. More generally, noisy-OR methods belong to the group of canonical models (Pearl, 1988). For these, assumptions on the underlying probabilistic relationship are made so that a conditional probability table can be generated algorithmically given parameters that are assessed by experts and which only grow linearly with the number of parent nodes. Usually the parameters refer to conditional assessments which are made about a number of combinations of the states of the parent nodes. An alternative to the aforementioned noisy-OR method is the noisy-MAX method (Diez, 1993). Within the same group of methods is also the ranked nodes approach (Fenton, Neil, & Caballero, 2007). Briefly, ranked nodes are random variables with discretised ordinal scales which are typically assessed by experts through verbal descriptors of the scale.

The usage of verbal classifiers to assess BNs has also been proposed more generally to counteract a high assessment burden. Here, the influence of a node is simply determined verbally rather than numerically. For instance, van der Gaag, Renooij, Witteman, Aleman, and Taal (1999) use a scale containing both, numerical and verbal anchors, and Mkrtchyan, Podofillini, and Dang (2015) conclude (in a review on the use of expert judgement for BNs in human reliability assessment) that the use of verbal labelling for inferences in BNs is common. We discuss verbal elicitation of dependence explicitly in the next section.

A last way to facilitate judgement is by providing graphical support. Hänninen, Banda, and Kujala (2014) provide experts with the pie chart probability tool available in GeNIe Bayesian Network Software to adjust assessments. Probability masses are determined and

the resulting distribution is graphically visible immediately. This procedure is repeated until the experts feel comfortable with their assessments.

As shown in Section 7, the use of expert judgement for BNs is considered in a variety of empirical areas given the popularity of this dependence model itself.

4.1.2. Copulas

In certain situations of context (a), a multivariate distribution can also be modelled by a copula rather than by the "marginal-and-conditional approach" (Clemen & Reilly, 1999), presented for BNs before. While an extensive introduction to copulas can be found in Durante and Sempi (2015) and Joe (2014), recall first that for a continuous random variable X with distribution function F_X, the random variable $U = F_X(X)$ is uniformly distributed. If we have two continuous random variables X and Y, then the distribution of the vector $(F_X(X), F_Y(Y))$ is supported on the unit square and has uniform marginals. Any such distribution is called a (bivariate) copula. This construction can be reversed: Any set of univariate distribution functions combined with a copula represents a multivariate distribution as a result of Sklar (1959). The notion of a copula is easily extended to greater than two dimensions.

Often a one–parameter copula family is used, $C_\theta(u, v)$, that can be indexed by a parameter θ related to a rank correlation such as those of Spearman or Kendall (see Appendix B). In fact, both can be expressed in terms of the copula: Spearman's correlation is

$$\rho_C = 12 \iint_{[0,1]^2} C(u, v)\mathrm{dudv} - 3$$

and Kendall's τ is

$$\tau_C = 4 \iint_{[0,1]^2} C(u, v)\mathrm{dudv} - 1$$

Within a chosen family of copulas (see Appendix B), expert elicitation can be used to determine the correlation and hence specify the dependence. Whenever the family is uncertain, information on how copulas differ for upper or lower tail concentration, i.e. tail (in-)dependence (see Appendix B), needs be elicited additionally. For this, upper (or lower) asymptotic tail dependence is of interest. The asymptotic upper tail dependence parameter is defined as:

$$\lambda_U(X, Y) = \lim_{u \to 1^-} P(Y > F_Y^{-1}(u) | X > F_X^{-1}(u))$$

when a limit $\lambda_U \in [0, 1]$ exists. In this case, X and Y are defined as dependent in the upper tail when $\lambda_U > 0$, whereas whenever $\lambda_U = 0$, they are tail independent (Joe, 2014). In other words, for the former case, it is more likely to observe high values for Y given high values for X. Following naturally from the concept of tail dependence, the tail concentration function distinguishes various copula formats and is defined for any u in (0, 1) as $\lambda_U = P(U > u, V > v)/(1-u)$. For the (upper) tail, it leads to the tail dependence coefficient in form of $\lambda_U = (1 - 2u + C(u, u))/(1-u)$.

The review results presented in Section 7 show limited experience for expert judgement within a copula modelling framework. One reason might be that copulas are distinguished on the one hand by measures of association such as rank correlations, but on the other hand also by its behaviour along the dependence function as indicated by its family. This constitutes a great deal of complexity to be integrated into an elicitation method. However, both types of information are highly important given that two different copula families exhibit a very different behaviour even for the same rank correlation (as shown in Appendix B). This is particularly crucial for copula families that model extreme joint dependence through asymptotic upper/lower tail dependence (as considered in the first elicitation approach presented below) in contrast to tail independent ones. At this point, it is important to note that the use and elicitation of measures of association related to tail dependence depends (obviously) on whether one is interested in capturing tail dependence explicitly or whether another measure might serve the modelling purpose better, given the increased cognitive complexity for experts to assess tail dependence.

Some proposed methods that aim at a sensible representation of an expert's understanding of dependence in form of a copula are outlined in the following. Arbenz and Canestraro (2012) decompose the asymptotic upper tail dependence coefficient (presented above) and query its components from experts before combining it again. They consider this as a non-asymptotic approximation of $\lambda_U(X, Y)$. The elicitation is as follows: in a first step, all non-negligible causes for X to be "extremely large" denoted as events j, $j = 1, 2, \ldots, J$, are listed. Then, experts assess $P(event\ j | X = \text{"}extremely\ large\text{"})$, so the likelihood that the chosen event is present if X is in the tail of its distribution. Lastly, experts are queried $P(Y = \text{"}extremely\ large\text{"} | event\ j)$, i.e. the probability that the corresponding event affects Y with the implied magnitude. All assessments are then combined by $\lambda_U(X, Y) \approx \sum_{j=1}^{J} P(Y = \text{"}extremely\ large\text{"} | event\ j) P(event\ j | X = \text{"}extremely\ large\text{"})$. The proposed framing is:

> *"Given that an extremely bad outcome is observed in X, what is your estimate of the probability that Y will experience an extremely bad outcome?"*

According to the authors (whose experts were actuaries) this method was perceived as cognitively easy.

Another option that is being researched further by several co-authors of this review but has not been published so far is querying conditional exceedance probabilities for chosen quantiles from experts to fit a parametric copula. This is done by plotting elicited values for each considered quantile together with candidate copula choices and after a first "eyeballing" use conventional goodness-of-fit tests for the distance to parametric families. Fig. 3 shows simulated conditional exceedance probabilities for several parametric copulas with given rank correlations. With the assessment of the probability that Y exceeds its uth quantile given that X exceeds its uth quantile for a certain number of thresholds u, a sensible copula choice that represents the experts' beliefs can be estimated. We address the details of eliciting conditional exceedance probabilities in the next section.

As a non-standard parametric alternative, Meeuwissen and Bedford (1997) discuss using a minimally informative copula with given rank correlation. A copula is modelled by asking experts to provide a dependence constraint between two random variables, and taking the copula which is minimally informative with respect to the uniform (independent) copula. This is further developed in Bedford, Daneshkhah, and Wilson (2016) and Bedford (2002). Here, experts assess the expectation of functions for the two underlying variables. From that a (min inf) joint probability is constructed which satisfies the expected value constraint. An advantage is that in this approach it is easier to relate a copula parameter to an observable quantity than it is for common parametric families. An example is given for the dependence of failure times between machine components. Minimal informativeness also served as motivation for Kotz and Van Dorp (2010) who consider a sub-family of generalised diagonal band (DB) copulas which require a dependence parameter. It is specified by experts through conditional exceedance probabilities (given the median value). Van Dorp (2005) regards DB copulas as advantageous when using expert judgement as a dependence parameter that relates to its one copula parameter can be defined. We will introduce this dependence parameter in the next section when we address forms of elicited dependence parameters explicitly.

Besides some empirical work in maintenance optimisation (Bunea & Bedford, 2002), the majority of experiences for elicit-

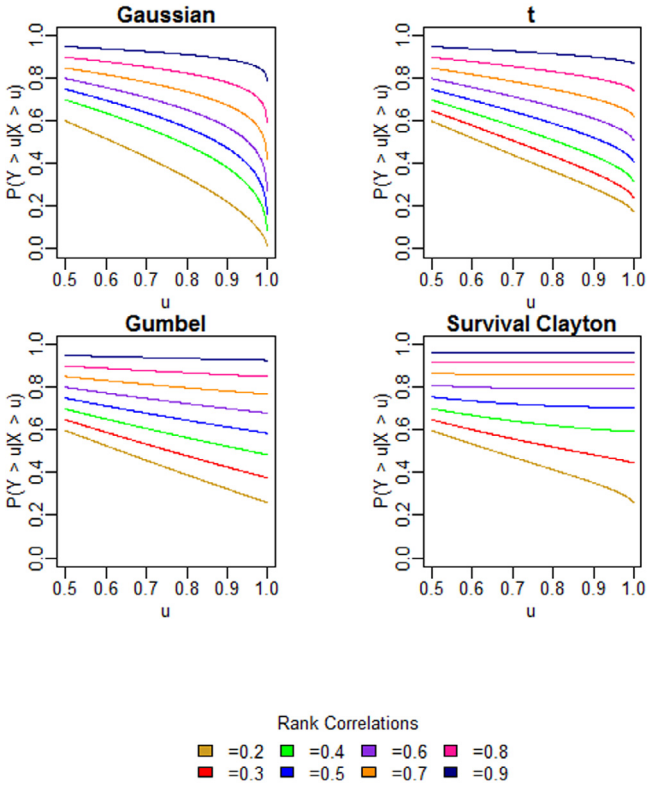

Fig. 3. Conditional exceedance probabilities at uth quantiles (rank correlations: 0.2–0.9).

ing copulas, such as the first approach presented above, comes from banking and insurance (Arbenz & Canestraro, 2012; Böcker, Crimmi, & Fink, 2010; Regis, 2011; Shen, Odening, & Okhrin, 2015), an area in which the popularity of copulas has increased lately (Genest, Gendron, & Bourdeau-Brien, 2009). Here, expert judgement is typically used to assess conditional and joint probabilities of (extreme) loss events. These studies might be helpful for other areas where copulas are gaining increased interest, such as hydrology (Genest & Favre, 2007).

4.1.3. (Probabilistic) parametric models: multivariate distributions

Another way to model dependence in (a) is by specifying a multivariate distribution. For an introduction and overview of the distributions discussed here, see Balakrishnan and Nevzorov (2004).

As a main challenge when eliciting a multivariate distribution is that its full specification would be cognitively too complex for experts, we should impose a structure on the distributional choice. While for univariate distributions it might be sufficient to assume a minimal structure such as a continuous and smooth cumulative distribution function which can be specified satisfactorily by a few quantile assessments (O'Hagan et al., 2006), in higher dimensions this is still unreasonable for practical use. Rather, a parametric multivariate distribution that represents an expert's belief sufficiently is a necessary assumption. Then, an expert's opinion is fully specified by determining a few parameters. While any distributional assumptions have to be in agreement with the experts, they should be as well in accordance with the modelling purpose. For instance, it should be suitable for its use in a specific decision problem for which a distributional form is predetermined or its use as a prior in a Bayesian modelling framework. The latter offers a probabilistic framework to complement the lack of data for some common statistical dependence models. Prior beliefs of experts (see Appendix B) for given parameters are updated once observations are available. A prior is chosen so that it can be most easily updated (O'Hagan et al., 2006). Generally, this is a different elicitation situation/purpose than using expert judgements to obtain beliefs about uncertainties without the inclusion of future observations (what is done in most of the literature reviewed here), but this is not of importance for us as with regards to dependence elicitation both methodologies have similar challenges. Hence, both methodologies contribute to the findings presented here.

In the literature on eliciting parameter information for quantifying a multivariate distribution, mainly multivariate normal (Al-Awadhi & Garthwaite, 1998, 2001; Dickey, Lindley, & Press, 1985; Garthwaite & Al-Awadhi, 2001), or t (Al-Awadhi & Garthwaite, 2001; Kadane et al., 1980) and Dirichlet distributions (Chaloner & Duncan, 1987; Elfadaly & Garthwaite, 2013; Zapata-Vázquez, O'Hagan, & Soares Bastos, 2014) are considered. A method that specifies a multivariate distribution in a more flexible way (as shown below) is given in Moala and O'Hagan (2010).

For the common parametric assumption of a multivariate normal or t distribution, the elicitation aims at quantifying the mean vector, μ, and the covariance matrix, Σ. Instead of determining the variables of interest directly, even though this has been attempted through interactive graphical methods (Chaloner, Church, Louis, & Matts, 1993), typically hyperparameters that follow from distributional assumptions on the form of μ and Σ and therefore specify (or index) the multivariate distribution of interest are determined. In other words, the values of the hyperparameters reflect the available subjective prior knowledge about the unknown model parameters. This is typically based on specifying hierarchical priors assuming exchangeability (see Appendix B) for the joint distribution in question. The variables of interest are then conditionally independent given the hyperparameters. This is known as Bayesian hierarchical modelling (see Appendix B) which is a common way to restructure dependence in order to elicit parameters as univariate quantities. Typically, the hyperparameters consist of means, scale parameters, degrees of freedom and the spread matrix which (whenever possible) are elicited through univariate quantities and conditional medians of observable variables. Percy (2002, 2004) presents how the specification of suitable prior distributions can be simplified and how values of hyperparameters can be elicited from experts through quantiles of predictive prior distributions for a variety of common distributions in the reliability context of mathematical modelling of maintenance. While we explain this approach below (for Dirichlet distributions), it is noteworthy here that a main advantage is that observable quantities can be used. Further, he proposes to elicit fewer quantiles than unknown hyperparameters and use interaction of experts for further adjustments.

A different problem for which a multivariate distribution needs to be specified is whenever an event can take one of k possible outcomes ($k > 2$) and the probability of the ith outcome, p_i, is elicited from experts. This might be denoted as eliciting the opinion about a "set of proportions" (Zapata-Vázquez et al., 2014). As the sum of all p_i must equal 1, p_i cannot be assessed in isolation. Further, with $k > 2$, a multinomial distribution models the overall outcome given that we have independent trials and the probability of each outcome is the same in each trial. The commonly chosen parametric distribution is then a Dirichlet distribution, the conjugate prior distribution of a multinomial one (O'Hagan et al., 2006). One of the earliest approaches in Chaloner and Duncan (1987) uses an elicitation strategy based on predictive distributions. When considering a specified number of draws from the population of interest, the expectation of the number that belongs to a category is in fact p_i. Given that, they ask their experts for the joint modes of the predictive distribution. Other methods assess the Dirichlet distribution by imaginary observations, i.e. by determining the extent to which experts change their belief given an observation

from a draw (O'Hagan et al., 2006). More recently, Zapata-Vázquez et al. (2014) proposed a refinement to acknowledge the strong assumptions of a Dirichlet distribution (due to the small number of parameters that determine its form) and therefore make use of over-fitting. Loosely, they ask experts for more assessments than (strictly) necessary to fit a distribution in order to reject the choice of a Dirichlet distribution if it is inappropriate.

A more flexible method that avoids experts' belief to fit a single pre-specified parametric family is presented in Moala and O'Hagan (2010). While the focus of the elicitation is laid on the analyst who seeks to identify the probability density function for a multivariate vector, the posterior distribution is based on the prior distribution as specified by an expert. In order to ensure flexibility on the parametric assumptions, the analyst's prior belief is a Gaussian process which allows the multivariate distribution to take a variety of forms given the experts' assessments. The elicited parameters are univariate quantities and a small number of joint probabilities, unless the elicitation of the latter can be reduced to querying univariate information as well, depending on assumptions for the multivariate vector's probability space.

Given that dependence information for quantifying parametric multivariate distributions is (mainly) elicited through univariate quantities, experimental studies show a similar performance to expert judgement studies with univariate variables of interest. For instance, (conditional) medians are regarded as cognitively easy and reliable to assess (Al-Awadhi & Garthwaite, 2006). Empirical findings on the elicitation of multivariate distributions are scarce however which is why no indication for a particular application area can be given (Section 7).

4.1.4. (Non-probabilistic) parametric models: Bayes linear methods

An alternative to eliciting distributional (prior) beliefs for Bayesian models in (*a*) is the Bayes linear method (BLM) (Goldstein & Wooff, 2007). It differs by using expectation as basis and is able to represent more complex problems through adjusting beliefs by linear fitting. Without distributional assumptions all required parameters are first and second moments (Farrow, 2003). Hence, eliciting dependence information concerns beliefs about the covariance of parameters (rather than joint probabilities). While not much experience on the actual elicitation is found in the literature, Revie, Bedford, and Walls (2010, 2011) and Revie (2008) address expert judgement for BLM specifically. The dependence model considered is $Y = \alpha X + R$ where X is the explanatory variable of Y. R represents the unexplained uncertainty between X and Y (with no correlation between X and R) and α is used to measure the strength of the relationship between X and Y. As a pragmatic way to elicit covariance information, the elicitation of quantiles is proposed whereas the relation between these and the moments needs to be derived. A possibility is through Pearson and Tukey (1965), further developed in Keefer and Bodily (1983), who propose eliciting from three to five percentiles to obtain means and variances. Hence, with the 5*th*, 50*th* and 95*th* quantiles specified as $x_{0.05}$, $x_{0.5}$, $x_{0.95}$ for an uncertain variable X, the mean is derived by $\mu_X = 0.63x_{0.5} + 0.185[x_{0.05} + x_{0.95}]$ and the variance by $\sigma_X^2 = ((x_{0.95} - x_{0.05})/(3.29 - 0.1(\Delta/\sigma_0))^2$ with $\Delta = x_{0.95} + x_{0.05} - 2x_{0.5}$ and $\sigma_0 = ((x_{0.95} - x_{0.05})/3.25)^2$.

In Revie et al. (2010) five elicitation techniques are compared. A first one is the direct elicitation of cross-moments which is omitted here given that it is discussed in the next section as a commonly elicited form. For the remaining methods we assume that the mean and variance of X and Y have been elicited beforehand. In the direct calculation approach, experts assess their updated belief of $E(Y)$ after the observation that $E(X)$ increased hypothetically. While α can be computed, for the uncertain variable R the experts' 5th, 50th and 95th quantiles are elicited through:

"Given that X is known to be $\bar{x}$ with complete certainty, what are the 5th, 50th and 95th quantiles of Y?"

It follows that $E(R)$ and $var(R)$ can be calculated as shown before and then $E(Y) = \alpha E(X) + E(R)$, $var(Y) = \alpha^2 var(Y) + var(R)$ and $cov(X, Y) = \alpha var(X)$. For the adjusted expectation method, experts are asked to re-assess their belief about X based on the true value of Y. When defining the true value as $\bar{y}$, the new belief for $E(X)$ is $E_Y(X) = X_Y$ with observed $\bar{y}$. The covariance can then be calculated as $cov(X, Y) = ((E_Y(X) - E(X))/(Y - E(Y)))var(Y)$. The value of α is again computed and defines the values an expert can assess for coherence reasons. The adjusted uncertainty approach works in the same way as adjusted expectation, with the only difference that the variance of X is updated based on an observation of the true Y. With the adjusted variance denoted as $var_Y(X)$, the adjusted covariance is then derived using $cov(X, Y) = \sqrt{(var(X) - var_Y(X))var(Y)}$.

In an experimental setting of the same study, experts were presented with the pairs of variables for life expectancy between males and females (in the same country), height and weight of male students, as well as mean time to failure between vehicles. All experts were familiar with basic statistical summaries, but not with BLM. The different techniques were compared for accuracy, incoherence and intuitiveness. Thereby, adjusted uncertainty was the only method that exhibited incoherent assessments and also had more inaccurate results with far more assessments of negative or no correlation when all empirical data was positively correlated. Direct calculation on the other hand had the best performance in terms of accuracy and no incoherent assessments. Direct correlation and adjusted expectation barely showed any differences for experts' performance. However, over 15% of the responses were deemed inconsistent.

While this is the first and only such complete attempt to explicitly focus on the actual elicitation of covariance in BLM, some main references for empirical studies with documented expert judgment approaches are Gosling et al. (2013), Revie, Bedford, and Walls (2011), Bedford, Denning, Revie, and Walls (2008), Farrow, Goldstein, and Spiropoulos (1997) and O'Hagan, Glennie, and Beardsall (1992).

4.2. Elicitation for indirect modelling with auxiliary variables

4.2.1. Regression models

A common dependence model in context (*b*) is a regression model. For recent overviews, see Ryan (2008) and Weisberg (2005).

Recall that here information on the dependence is modelled *indirectly* by restructuring the natural input. Technically restructuring is done using variable transformation techniques. Beliefs about parameters are then elicited while being formulated as univariate query variables. Similar to quantifying parametric multivariate distributions, elicitation here is typically done for prior beliefs in a Bayesian methodology.

The parameter of interest is a regression coefficient, β. The likelihood function $p(Y|X, \beta)$ relates observed data Y to regression coefficients β and covariates X. Experts then specify the prior distribution for $p(\beta)$ typically through hyperparameters which are the mean and the variance of the regression coefficient (James, Choy, & Mengersen, 2010). Eliciting moments of regression coefficients directly however might be cognitively too complex given that experts would need to understand the effect that a change of covariate X has on Y. Therefore, the literature on eliciting priors for regression models proposes indirect approaches. For these, experts provide a probability of the response value based on specified values of the explanatory variables or vice versa. From this, prior elicitation methods for linear models, normal (Kadane et al., 1980) and multiple (Garthwaite & Dickey, 1991), piecewise-linear (Garthwaite, Al-Awadhi, Elfadaly, & Jenkinson, 2013) as well as logistic

regression models (O'Leary et al., 2009) have been developed. For the latter, experts typically assess conditional means, $E(Y|X, \beta)$ (Bedrick, Christensen, & Johnson, 1996; James et al., 2010) for a probability of presence, p_i, with binary responses for observation i modelled as $\text{logit}(p_i) = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + ... + \beta_j x_{i,j} + \epsilon_i$ (O'Leary et al., 2009). For instance, Choy, O'Leary, and Mengersen (2009) elicit the probability of presence for a certain wallaby type at a specified location with fixed habitat characteristics in habitat modelling. Depending on distributional assumptions for the probability of presence (such as a Beta distribution) the mode rather than an arithmetic average or median might be elicited due to the potential skewness of the distribution.

In a similar manner, parameters can be elicited for (multiple) linear regression models. Garthwaite and Dickey (1991) propose a model of the form:

$$E(Y|x_1, x_2, ..., x_i) = (\beta_1 x_1, \beta_2 x_2, ..., \beta_i x_i)$$

where again β denotes the regression coefficient and $E(Y|x_1, x_2, ..., x_i)$ is the expected (average) value of Y when $X_1 = x_1, X_2 = x_2, ..., X_i = x_i$. Experts then specify the prior distribution of β by assessing hyperparameters. To do so, the authors introduce design points, values at which a prediction is made after hypothetical data are given. Likewise, Kadane et al. (1980) elicit fractiles for a predictive distribution with specified values at design points, using a bisection method (see Appendix B).

Regression elicitation is further explored in Choy et al. (2009), O'Leary et al. (2009) and Al-Awadhi and Garthwaite (2006). O'Leary et al. (2009) present three different elicitation methods with graphical support, similarly to Al-Awadhi and Garthwaite (2006) who use an interactive graphics method as well. Empirical studies for expert judgement in regression modelling are mainly found in the area of ecology for which e.g. Choy et al. (2009) summarise various approaches.

4.3. Elicitation for modelling propagation of output

4.3.1. Probabilistic inversion

In modelling context (c), a common situation is that input parameters of a dependence model are not observable. Therefore, a direct quantification of these variables is not sensible and methods such as PI (Cooke, 1994; Kurowicka & Cooke, 2006) are used. Its aim is to take the distribution representing the uncertainty on certain observables and translate it on the uncertainty of target variables. While the distribution can come from historical data, PI can be used as well as a method for transforming expert assessments of some observable model outputs into uncertainties on parameter values. A motivation for PI (that was never published as such) originated in the development of expert judgement methods and uncertainty analysis in the nuclear sector (for a historical overview, see (Cooke, 2013; Kraan & Cooke, 1997)) where experts refused to assess transfer coefficients directly. Similarly, Kraan and Bedford (2005) elicit outputs of a power law that models spread of lateral plume in atmospheric dispersion in form of $\sigma_y(x) = A_y x^{B_y}$. The output $\sigma_y(x)$ denotes the lateral (indicated as y) spread at windspeeds x and is determined by the dispersion coefficients A and B. Instead of querying the joint distribution on (A, B), which would require experts to consider all possible effects of this relationship through the model, they are asked to quantify uncertainty on the output at various downwind distances through a univariate elicitation method. In addition to modelling plume spread, the same paper discusses a case study in banking. Empirical findings of the method are however lacking which is why no indication of specific application areas can be given.

5. Forms of elicited dependence parameters

This section reviews the proposed forms of dependence parameters for elicitation, i.e. association measures or summary types of an expert's joint distribution that are used in an elicitation question. As well, the corresponding framing of elicitation questions is presented. In addition to outlining the main elicited forms, an evaluation regarding desirable properties is given whenever possible. Chosen desiderata allow for guidance on the suitability of elicited dependence parameters from different perspectives.

Desiderata for elicited dependence parameters

A first perspective concerns theoretical feasibility whereas a common desideratum for expert judgement is that the elicited forms are observable and physically measurable. This allows assessments to be credible and defensible (Cooke, 1991). With a similar objective, a rigorous foundation in probability theory is desirable.

A further perspective considers the assessment burden for experts. In this regard Kadane and Wolfson (1998) emphasise practicality, i.e. that experts feel comfortable at assessing uncertainty while their opinion is captured to a satisfactory degree. For the former, query variables should be kept intuitively understandable. For the latter, queried information should be linked as directly as possible to the specific dependence model of interest, ensuring that an expert's assessment is satisfactorily reflected in the final output of the model. As variables are often transformed into some other parameter than the one that populates a dependence model (e.g. due to a potential reduction in the assessment burden), it is important to measure and control the degree of resemblance between the resulting assessments (through the model) and the dependence information as specified by the expert (Kraan, 2002). Note that the transformation of dependence parameters is typically based on assumptions about the underlying bivariate distribution. For instance, when transforming a product moment correlation coefficient into a rank correlation, this is straightforward under the assumption of bivariate normality. However, positive definiteness is not guaranteed which relates to the next desideratum, that of mathematical coherence. Coherence means that the outcome should be within mathematically feasible bounds. For dependence measures, ensuring positive definiteness of a resulting correlation matrix might be a potential issue and methods that adjust experts' judgements might be necessary (Lurie & Goldberg, 1998). Yet, whether an expert agrees with this adjustment or not determines their confidence in the final assessment. Another solution to incoherence is to fix possible bounds for the assessment a priori, even though this can severely decrease the intuitiveness of the assessment. A last desideratum is to calibrate assessments on statistical accuracy. This means, we would like to test experts' performance (in terms of statistical accuracy) against empirical data (if available), often to inform the weighting for mathematically combining judgements.

While no elicited dependence parameter meets all desiderata, their consideration supports comparison and allows a better guidance in terms of suitability within certain modelling situations.

At a broad level, a distinction for elicited quantities can be made between *probabilistic* and *statistical* approaches (Clemen & Reilly, 1999; Kraan, 2002; Morales Nápoles et al., 2008). Whenever possible the presented findings are categorised into one of the groups. Approaches that do not fit in any of these classifications can be found in Section 5.3.

5.1. Probabilistic methods

In the selected literature popular variables to elicit are of probabilistic nature. This popularity can be attributed to the firm

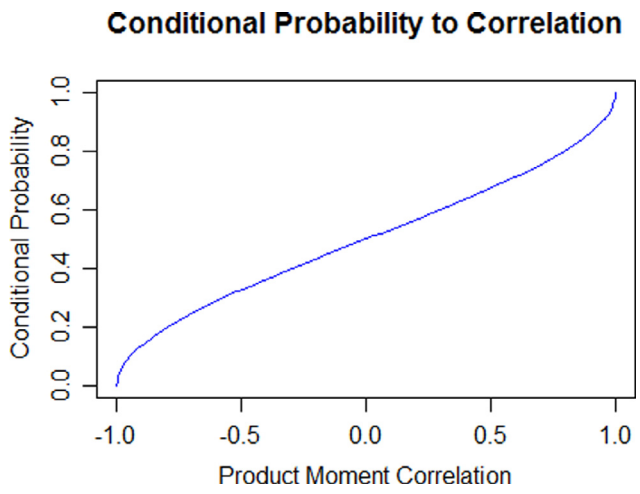

Fig. 4. Expert's conditional probability assessment as a function of the product moment correlation coefficient.

foundation (in probability theory) and the (potential) observability of the elicited variables which accompany this choice.

5.1.1. Forms of probabilistic dependence parameters

Conditional (exceedance) probabilities. In the context of probabilistic measures of dependence, *conditional probability* might be the best known one. A common way to elicit conditional probabilities is to provide an expert with the information that the conditioning variable is observed above (or below) its median value (marginal probabilities are elicited first or are known from data) before the probability that the target variable lies above (or below) its median value is enquired. A possible framing of the question is:

> *"Consider the pair of variables, X and Y. Suppose now that Y has been observed to be above its/your median value for it. What is the probability that X lies also above its/your median value for it?"*

This might be extended to any quantile defining for the pair of random variables X and Y the elicited form for a conditional probability as $P_{CP}(x_i, y_i) := P(X \geq x_i | Y \geq y_i)$ where $i = 0.5$ refers to the median value, but i might take any other quantile. Experts assess independence between X and Y as $P_{CP}(x_i, y_i) = P(X \geq x_i)$ implying that learning about $P(Y \geq y_i)$ does not add any information. For a (strong) negative relationship experts state their belief as $P_{CP} \in [0, P(X \geq x_i))$ while for a (strong) positive it is $P_{CP} \in (P(X \geq x_i), 1]$. Given the above form, a conditional probability is sometimes also called conditional exceedance probability. In contrast, another way to elicit a conditional probability is by $P_{CP}(x_i, y_i) := P(X \geq x_i | Y = y_i)$. This way can be applied similarly and its use depends strongly on context. However, O'Hagan et al. (2006) regard it as less cognitively complex.

In order to transform a conditional probability into a product moment correlation coefficient (e.g. for modelling purposes) the relation between the two can be derived as shown in Fig. 4.

The above derivation is possible only when an assumption about the underlying copula is made (Kurowicka & Cooke, 2006). Fig. 4 was obtained under the assumption of normal copula density for X and Y. The analyst finds the product moment correlation that ensures a positive definite correlation matrix and satisfies the expert's assessments (Morales Nápoles et al., 2008).

Experts' performance when eliciting conditional probabilities (in comparison to six other methods) has been investigated in Clemen, Fischer, and Winkler (2000). The assessed pairs of variables are relationships such as height–weight, as well as dependence between individual stocks, their indices and the relation between stocks and their indices. Participating experts were MBA students with some basic statistical training. In this experimental setting, conditional probability is among the worst performing methods for coherence and fourth out of six in terms of accuracy against empirical data. Similar coherence issues when assessing conditional probabilities were observed by Moskowitz and Sarin (1983) who therefore provided their experts with a Joint Probability Table which led to large improvements in performance. Generally, for this method the elicitation of several values to condition on is recommended (Cooke & Kraan, 1996).

In the case-study literature (Section 7), the elicitation of conditional probabilities is nevertheless favoured as it often serves as direct model input. Main references where this approach has been formally used stem from the Joint CEC/USNRC Uncertainty Analysis framework (Cooke & Kelly, 2010). The experts participating in these studies became familiar with this format which underlines the importance of training experts to ensure familiarity.

An alteration to the elicitation of conditional probabilities which is also closely related to concordance probabilities (see below) is presented in Fackler (1991). Experts are asked to assess the median deviation concordance probability which is also known as quadrant probability (Kruskal, 1958). It is defined as the probability of the two variables, X and Y, falling both either below or above their medians, i.e. $P_{QP}(x, y) := P((X - x_{0.5})(Y - y_{0.5}) > 0)$ with $x_{0.5}$ and $y_{0.5}$ being the respective medians. This could be asked for as follows:

> *"Consider the pair of variables X and Y. You have indicated that there is a* 50/50 *chance of X being above or below* $x_{0.5}$ *and Y being above or below* $y_{0.5}$*. What is the probability that X and Y both will either be above or below their medians?"*

The above formulation is a slightly altered version of the original reference to offer a general framing. While the conditional probability cannot be fully represented with a quadrant probability, the author claims that the dependence elicitation concentrates on events that experts "should be capable of making most informed judgements about" (Fackler, 1991). According to Kruskal (1958), this is "perhaps the simplest measure of association between two random variables" and an advantage is that it can be assessed and interpreted on the customary range. This measure is non-parametric, meaning that is has a well-defined interpretation (even) when structural assumptions, such as bivariate normality, do not hold. Further, it is ordinally invariant, i.e. it remains unchanged by monotone functional transformations of its coordinates. This has advantages with regards to modelling convenience as well as in terms of cognitive complexity to assess it. The measure is closely related to Blomqvist β (Blomqvist, 1950) which is defined as $\beta = P((X - x_{0.5})(Y - y_{0.5}) > 0) - P((X - x_{0.5})(Y - y_{0.5}) < 0)$.

Similar to Kruskal (1958) when discussing the conveniences of using the quadrant probability, Blomqvist (1950) describes his measure of association as being "valid under rather weak assumptions regarding the distribution of the population" and "easy to deal with in practice". Under the assumption of bivariate normality, a relation to the correlation coefficient, ρ, is given by $(2/\pi \arcsin \rho)$. Given the advantages from a modelling together with elicitation perspective and as pointed out by a reviewer of an earlier version of this paper, the quadrant probability and Blomqvist β deserve more attention when eliciting dependence.

Conditional (exceedance) probabilities (for higher dimensions). Eliciting higher dimensions of dependence such as in Morales Nápoles, Hanea, and Worm (2013) and Morales Nápoles et al. (2008) requires the assessment of conditional rank correlations in addition to unconditional ones. To do so, the variables of interest that are conditioned onto are ordered according to some order of preference. This corresponds for instance to the relation of parent to

child nodes in a directed acyclic graph. Once experts have assessed the unconditional rank correlation ρ_{X,Y_1} (in Fig. 2) with any of the other techniques presented here, the conditional rank correlations need to be determined ($\rho_{X,Y_2|Y_1}$ and $\rho_{X,Y_k|Y_2,Y_1}$ in Fig. 2). A probabilistic way to do so is through conditional (exceedance) probabilities for higher dimensions which directly follow from the low dimensional case discussed above. A question (according to Fig. 2) might be framed as follows:

> *"Suppose that not only Y_1 but also Y_2 has been observed above its/your median value. What is now your probability that also X will be observed above its/your median value?"*

For this the conditioning set of the unconditional case will be extended to $P_{CP}(x_i,\ y_{1,\,i},\ y_{2,\,i}) := P(X \geq x_i | Y_1 \geq y_{1,\,i},\ Y_2 \geq y_{2,\,i})$ for the ith quantile, e.g. $i = 0.5$ for the median. If experts assess (conditional) independence, the estimate will be the same as for $P_{CP}(x, y_1) = P(X \geq x_i | Y_1 \geq y_{1,i})$. Otherwise the positive/negative relationship is assessed as before. Whenever $P_{CP}(x, y_1, y_2) \neq 1$ or 0 it follows that X is not completely explained by Y_1 so that Y_2 adds to the explanation of the former. In psychological research of causal learning theory, Y_1, Y_2 and Y_k would be referred to as cues that compete for associative strength (Mitchell et al., 2009). The idea of associative strength shows a key difference to the elicitation of noisy-OR parameters presented earlier in the context of BNs.

The intuitiveness of this method might be inhibited given that the choice of the first (unconditional) correlation imposes restrictions of the possible values for the second (conditional) correlation (similar to those of positive definiteness of a correlation matrix). This introduces the necessity to compute (in real time) updated intervals (different than the unrestricted $[-1, 1]$) into which the new assessment can fall, to preserve coherence. Technical details can be found in Morales Nápoles (2010).

In order to test experts' performance when assessing a multidimensional dependence structure, (Morales Nápoles et al., 2013) compared conditional probabilities of exceedance with the direct elicitation of pairwise correlation. In their study, a group of 14 experts (with previous training on statistics) was presented with two versions of a graphical model for the relationship between sulphur dioxide emissions and fine particular matter in Alabama, USA. The experts were split into two groups so that different dependence measures could be elicited. For the first model, querying the rank correlation directly exhibited the best performance when averaging out the absolute difference of empirical data and all individual answers. Based on a performance-based measure of accuracy (detailed in Section 6), the top three most accurate experts assessed correlation directly. However, when averaging performances per elicitation technique and model, the conditional exceedance probabilities outperformed direct assessments. Nevertheless, the authors could not formulate definitive conclusions since the different model versions might have had an influence on the differences in experts' performances.

Joint probabilities. From conditional probabilities it follows naturally to consider the elicitation of joint probabilities. A joint probability, $P_{JP}(x, y) := P(X \leq x, Y \leq y)$, can be queried for two random variables, X and Y, by asking:

> *"Consider the pair of variables X and Y. What is the probability that both are within the lower (upper) k_{th} percentage of their respective distributions?"*

If an expert assesses independence between X and Y, the joint probability corresponds to $P_{JP}(x, y) = F_X(x)F_Y(y)$, where F_X and F_Y represent the marginal cumulative distributions of the corresponding elicitation variables. A positive relationship is assessed by either $P_{JP}(x, y) = F_X(x)$ or $P_{JP}(x, y) = F_Y(y)$. For a negative relationship $P_{JP}(x, y)$ approximates 0.

A relation to the (product moment) correlation coefficient is derived similarly as in the case of conditional probability. For medians, conditional probabilities are derived by using the relation $2P(X \geq x_{0.5}, Y \geq y_{0.5}) = P(X \geq x_{0.5} | Y \geq y_{0.5})$ (O'Hagan et al., 2006).

Daneshkhah and Oakley (2010) mention a modification to elicit joint probabilities. It is presented in Moala and O'Hagan (2010), where the elicited probability takes the form $P_{JP}(x,\ y) := P(x_{\mathrm{i}} \leq X \leq x_{\mathrm{j}},\ y_{\mathrm{i}} \leq Y \leq y_{\mathrm{j}})$. It is concluded that this alternative is able to capture the most important features of an expert's distribution with a good accuracy and by just making use of a small amount of data.

Eliciting joint probability directly however is seen as rather cognitively complex and (even) assessing independence in such a way is regarded as non-intuitive (Garthwaite et al., 2005). A systematic bias for this kind of assessment is that experts tend to overestimate the probability of conjunctive events and underestimate that of disjunctive ones (O'Hagan et al., 2006). This might be due to the requirement that certain knowledge of probability theory is necessary for this method. (Clemen et al., 2000) found that when elicited joint probabilities are transformed to correlations, the obtained values tend to be out their feasible bounds rather frequently. Further, it was the least accurate method when compared to empirical data.

Concordance probabilities. A further way to think probabilistically about dependence is by considering concordance (and discordance) of random variables. The concept of concordance probabilities is closely related to the earlier introduced quadrant probability and it is limited to a frequency or cross-sectional interpretation for the pair of variables in question, i.e. it requires a population to draw from (Clemen & Reilly, 1999). The question can be framed as:

> *"Consider two independent draws, (x_a, y_a) from their common underlying population a and (x_b, y_b) from population b. Given that $x_a > y_a$ holds for population a, what is your probability that the relation $x_b > y_b$ holds for population b?"*

Exemplary populations for a and b might be height and weight of some specified group of people. Formally, the probability of concordance between two random variables, X and Y, considering n independent draws (x_a, y_a) to (x_b, y_b) is given by:

$$P_C(x, y) = \frac{\sum_{a=1}^{n-1} \sum_{b=a+1}^{n} 1_{C^*}((x_a, y_a), (x_b, y_b))}{\binom{n}{2}}$$

with $C^* = (x_a - x_b)(y_a - y_b) > 0$. It can be assessed by an expert on [0, 1]. A value of (or close to) 0 indicates a strong negative relationship, 0.5 represents independence, and 1 refers to a strong positive relationship. The transformation to a rank correlation such as Kendall's tau, τ, is defined as $\tau = 2P_C - 1$. With the assumption that X and Y can be approximated by a bivariate normal distribution, the relation from τ to other correlation measures, such as Pearson's product moment correlation, ρ^*, or Spearman's rank correlation, ρ, can be inferred through $\rho^* = \sin(\pi\tau/2)$ and $\rho^* = 2\sin(\pi\rho/6)$ (Kruskal, 1958). Nevertheless, a (transformed) product moment correlation matrix that is positive definite is not guaranteed (Kraan, 2002).

Within the psychological literature of causal learning, the concordance probability relates to the term *degree of relatedness*. In the classical experimental design, participants are presented with information about the presence or absence of an input variable, representing a candidate cause, as well as the presence or absence of an effect/outcome. For instance, medical experts assess the likelihood of a disease from the (non-) occurrence of a symptom. Based on their assessments of discordant and concordant observations the aim is to formulate descriptive rules for inferring causal strength (Shanks, 2004).

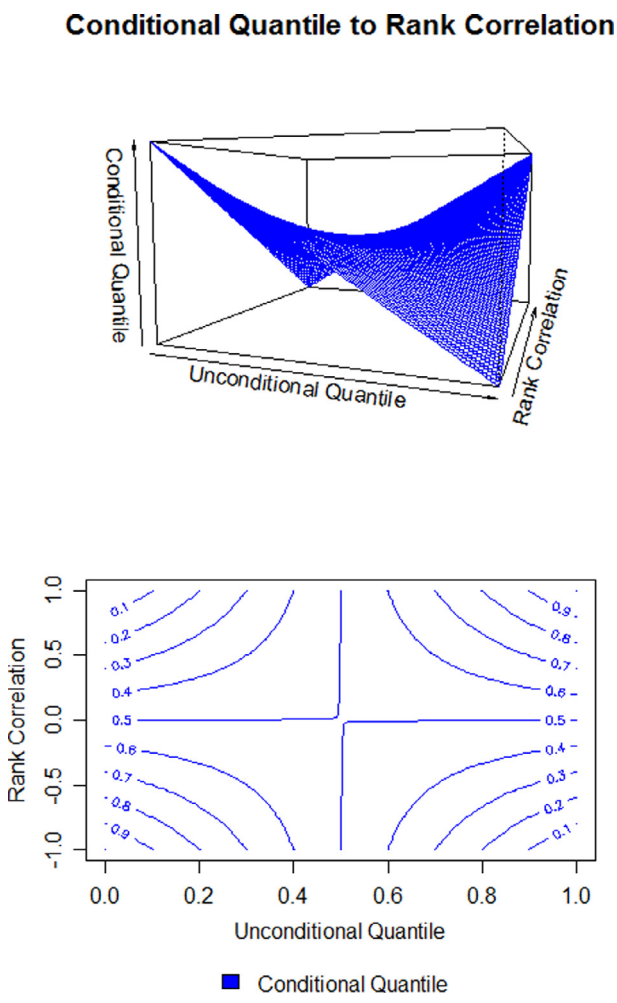

Fig. 5. Conditional quantiles to rank correlations.

In Clemen et al. (2000), this technique performed reasonably accurate in comparison to other methods and only rarely incoherent assessments were made. Similarly, Garthwaite et al. (2005), Kunda and Nisbett (1986) and Gokhale and Press (1982) come to the conclusion that this method is reasonably accurate and might be preferred if a population is given. Yet the importance of an expert's familiarity with the population is emphasised.

Expected conditional quantiles (fractiles/percentiles). The quantile (fractile/percentile) method requires conditional estimates and therefore shares certain characteristics with eliciting conditional probabilities. Experts are presented with information that the conditional value corresponds to a certain quantile (or fractile/percentile) and given that information, the experts assess which expected quantile the other variable takes. A possible framing might be:

> *"Consider variables X and Y. Given the value Y has been observed at its ith quantile, q_i. What is your expectation of X's value in terms of its quantile?"*

For the pair of random variables, X and Y, this is defined as $E(F_X(x)|Y = y(q_i))$ where $F_X(x)$ is the corresponding distribution function of X and $y(q_i)$ is the value that Y takes at its ith quantile. The relation to rank correlation is given through the standard non-parametric regression function of $E(F_X(x)|Y = y(q_i)) = \rho_{X,Y}(F_Y(y) - 0.5) + 0.5$ (Fig. 5). The conditional quantile is bounded by $\mu_{\min} \leq E(F_X(x)|Y = y(q_i)) \leq \mu_{\max}$ where $\mu_{\min} = \min[F_Y(y), 1 - F_Y(y)]$ and $\mu_{\max} = \max[F_Y(y), 1 - F_Y(y)]$. If $F_Y(y)$ is above its median, the values close to the minimum refer to a (strong) negative relationship, and the values close to the maximum indicate a (strong) positive one. For independence, experts assess $E(F_X(x)|Y = y(q_i)) = 0.5$. A closely related method is predictive assessment which was mentioned in the context of hyperparameters.

It should be noted that this dependence parameter has certain characteristics which would have similarly justified listing it among the statistical approaches which are presented in Section 5.2.1, after the general discussion on the assessment burden of probabilistic methods.

5.1.2. Assessment burden of probabilistic methods

Despite the limited empirical evidence available for experts' intuitive understanding of different assessment methods, Morales Nápoles et al. (2008) and Clemen et al. (2000) conclude that probabilistic statements are not perceived as cognitively easy. Conditional as well as joint probability assessments were rated by experts as most difficult among all other methods presented to them. In particular, when moving towards higher dimensions, the growing conditioning sets for conditional exceedance probabilities were met with accordingly growing concern. Additionally, for conditional quantiles (fractiles/percentiles) the expert must understand these location properties of distributions quite well together with the notion of regression towards the mean which might induce cognitive difficulties (Clemen & Reilly, 1999). A possible advantage of these techniques is that the assessment burden can be decreased for most probabilistic methods by re-framing the questions. For instance, it is often possible to express their forms as relative frequencies which are a more natural way of thinking about probabilities. Such framings were found to have a positive effect both on assessment burden and accuracy in the univariate case (Hoffrage, Lindsey, Hertwig, & Gigerenzer, 2000). Recognition of the cognitive burden of assessing dependence has existed at least since Kruskal (1958), who supports probabilistic methods, in particular the quadrant probability, due to its intuitive decision analytic interpretation in comparison to statistical methods.

5.2. Statistical methods

Despite some objections to the direct elicitation of moments of distributions or even cross moments, such as non-observability (Kadane & Wolfson, 1998), the literature offers some interesting findings and conclusions about the direct assessment of statistical measures of association (and alternative formulations).

5.2.1. Forms of statistical dependence parameters

Direct (rank) correlation. Directly asking experts for the natural input of a dependence model is seen by some as a natural way of eliciting dependence. Often, this is a correlation coefficient. One option is to ask experts for an estimate of the (rank) correlation between pairs of variables X and Y. A framing might be simply:

> *"Consider variables X and Y. What is the (rank) correlation between them?"*

This usually refers to the Spearman's rank correlation coefficient (see Appendix B) which is defined on the interval of $[-1, 1]$. A value of $\rho = -1$ denotes the strongest possible negative correlation, $\rho = 0$ expresses that X and Y are uncorrelated while $\rho = 1$ refers to the strongest possible positive relation. An advantage of eliciting rank correlations over product moment ones is that the interpretation of the former is independent of its marginal distributions implying that its values are always in the aforementioned interval. Nevertheless, for choosing the appropriate correlation coefficient, an analyst has to take into account what kind of relationship is assessed. Rank correlations, such as Spearman's version,

assume monotonicity while Pearson's product moment coefficient (see Appendix B) can only be meaningful for linear relationships (Reilly, 2000).

An obvious precondition for this type of dependence parameter to be intuitive is a certain level of familiarity with statistical measures. Therefore, several (conflicting) conclusions have been made from research on this query variable. Some studies, such as Kadane and Wolfson (1998), Morgan, Henrion, and Small (1992), as well as Gokhale and Press (1982), view a direct method as unreliable. The latter for instance conclude that even trained statisticians will have difficulties with this method even when being presented with graphical output in form of scatterplots. This is in agreement with Meyer, Taieb, and Flascher (1997) who conclude that experts judge the degree to which variables deviate from perfect correlation rather than directly assessing dependence of variables when shown a scatterplot. Yet according to other research, a direct elicitation has performed better in comparison with other assessment methods. Revie et al. (2010), Clemen et al. (2000) and Clemen and Reilly (1999) concluded that eliciting a correlation coefficient is more accurate than other dependence variables (in relation to empirical data) as well as more coherent. The better performance in comparison to other methods is primarily attributed to sufficient normative expertise of the experts.

Ratios of (rank) correlation. When considering higher orders of dependence, a direct way to elicit this information from experts is through ratios of (unconditional) rank correlations. In this method, experts assess the "relative strength" of each rank correlation (Morales Nápoles, 2010). (Morales Nápoles, Delgado-Hernández, De-León-Escobedo, & Arteaga-Arcos, 2014) and (Delgado-Hernández, Morales-Nápoles, De-León-Escobedo, & Arteaga-Arcos, 2014) present it as an alternative to conditional exceedance probabilities for higher dimensions which have the requirement to assess large conditioning sets that make the elicitation exercise rather unintuitive.

When defining unconditional rank correlations in the exemplary BN of Fig. 2 as r_{X,Y_1} and r_{X,Y_2}, then for the first conditional rank correlation, $\rho_{X,Y_2|Y_1}$, the ratio $R = r_{X,Y_2}/r_{X,Y_1}$ would be elicited. The corresponding question might be framed as:

Given your previous estimate, what is the ratio of r_{X,Y_2} to r_{X,Y_1}?

Similar to the conditional probabilistic techniques, the values that an expert can assess are restricted for each subsequent ratio. Imposing bounds ensures coherence but makes the elicitation less intuitive. Empirical comparisons to probability of exceedance have neither shown a superior nor an inferior performance. Nevertheless, the proponents of this method found that experts often think in terms of unconditional correlations rather than ratios. The intention of the ratio framing is to prompt experts to think in terms of relative influence between variables. However, there is no way of ensuring the experts will follow the proposed path.

Verbal. An indirect statistical approach to elicit experts' beliefs about dependence is through the use of a pre-defined scale. The most common way to do so is by using verbal descriptions that correspond to certain correlation coefficient values. For instance, Clemen et al. (2000) use a scale of seven points on which the relationship between X and Y is measured as $S_{X,Y}$. The points range from 1 describing a very strong negative relationship up to 7 which denotes a very strong positive relationship. Accordingly, 4 refers to no relationship. The transformation to Spearman's rank correlation is done through $\rho = (S_{X,Y} - 4)/3$. Despite its obvious subjectivity in determining the scale due to the rather informal translation of verbal qualifiers, a good performance in terms of coherence and accuracy can be observed in empirical studies using this method. Moreover, the method is intuitive which makes it popular. In the area of human reliability analysis, Swain and Guttmann (1983) introduce the Technique for Human Error Rate Prediction (THERP) which uses a verbal scale for assigning the dependence level between human errors. The conditional probability for failure between tasks A and B is computed as $P(B|A) = (1 + K \cdot P(B))/(K + 1)$ where K is assessed via verbal qualifiers of complete dependence ($K = 0$) to high ($K = 1$), medium ($K = 6$), low ($K = 19$) and zero dependence ($K = \infty$). The dependence assessment method in THERP is the foundation of various further developments of dependence modelling efforts in this area.

Coefficient of determination. A method that has been used rather rarely but that is still possible is to elicit the coefficient of determination. For this, Clemen and Reilly (1999) propose to ask for the percentage of variance explained as it would result from regressing one variable on another (R^2). Van Dorp (2005) uses this idea to construct a dependence measure which can be used in the elicitation of copula parameters. It is proposed for a common risk factor model within the context of the Program Evaluation and Review Technique (PERT) for which dependence is modelled with a DB copula (see previous section). PERT is an operational research technique for analysing and scheduling projects whereas the uncertainty in completion time is typically of interest. For modelling the dependence between the (aggregated) common risk factor Y (factors influencing project completion time) and random variable X (completion time), first $R(X) = b - a$, i.e. the range where realisations of X can be observed, is defined. Next, the range of the conditional distribution, $R(X|Y = y, \phi)$, is specified where the state of different common risk factors that result in the aggregate risk of Y as well as the dependence parameter of the DB copula, ϕ, are known. From this, the dependence measure $\xi(X|Y, \phi) = (1 - R(X|Y, \phi)/R(X))100\%$ is derived (see reference for full elaboration). This measure can be thought of as the average percent reduction in the range of X when the state of common risk factor,Y, is given. Suppose Y defines the set of possible risk factors, $Y = \{rain, no\ rain\}$, and the range of X is the length of an activity, e.g. a project's duration in days. Then the query question is asked as follows:

> *"Not knowing the state of the common risk factor, Y, a value of x has been assessed for X. Suppose you knew the state of the common risk factor, Y, on average within a spread of how many days could you now assess the completion of this activity, X?"*

An expert's assessment of 5 days would then correspond to 50%, i.e. this is the percentage of uncertainty that is explained by knowing the state of the risk factor. The author highlights that the elicitation question is framed in terms of X which is an observable quantity. While an intuitive appeal for the method is mentioned, no empirical results in terms of performance or cognitive burden for experts have been reported. Extensions for use with different copula families are achieved by slightly altering the formulation of $R(X)$.

5.2.2. Assessment burden for statistical methods

Overall, the statistical methods are seen as intuitively accessible for experts and enjoy favourable feedback in terms of assessment burden (Clemen et al., 2000; Revie et al., 2010). Especially verbal scales are seen as directly applicable and have therefore enjoyed further consideration. Clemen et al. (2000) report that for statistical methods training and feedback for follow-up studies improved accuracy. This is confirmed by expert studies with frequent feedback on correlation assessments, such as weather forecasters (Bolger & Wright, 1994).

Similarly, neurological experiments in which experts get frequent feedback on correlation coefficients find evidence for a hu-

man ability to "learn" the effect of varying correlation coefficients (Wunderlich, Symmonds, Bossaerts, & Dolan, 2011). Even though not conclusive, there are reasons to believe that statistical methods for dependence elicitation are more intuitively understandable, or at least "learnable", when compared to other approaches. This is nevertheless a signal rather than a strong conclusion also due to the fact that statistical methods have often been tested (only) for simple examples (e.g. height–weight relationships) rather than complex elicitation problems.

With regards to the complexity of problems for which experts might assess a correlation directly, Kruskal (1958) offers perhaps one of the most detailed discussions. He addresses the cognitive complexity required for assessing correlation coefficients directly in terms of their operational, decision-analytic and intuitive interpretation. From this perspective, according to him the necessary level of cognitive processing for assessing a correlation coefficient can be rather high. For instance, when interpreting a (rank) correlation in terms of concordance and discordance of hypothetical observations of a population (which has a clear and intuitive meaning) experts might have to assume (the rather unintuitive idea of) an infinite population (see Appendix B for the definition of rank correlations). The product moment coefficient is seen as (even) more difficult to assess as it is not ordinally invariant which (as aforementioned) inhibits a simple, intuitive understanding given that any assessment is interpreted with regards to the transformations made to the marginal distributions.

5.3. Other methods

In the following, methods that do not fit the categories above (for reasons which will be explained) are considered.

One such method is proposed by Abbas, Budescu, and Gu (2010) who elicit joint probabilities through univariate distributions and isoprobability contours. In other words, dependence is elicited indirectly. We present this approach separately because experts express preferences over binary gambles with identical payoffs rather than providing probabilistic (or numerical) responses directly.

Loosely, an isoprobability contour is a collection or set of points which have the same cumulative probability. In order to elicit the 50th percentile of a contour for two variables of interest, X and Y, experts assess first the common quantiles for X, e.g. the median, $x_{0.5}$, the 75th quantile, $x_{0.75}$, and so forth. Then, the experts are offered two gambles, for which the authors propose the framing of:

A: *You receive a fixed amount, z, if the outcome of variable X is less than* $x_{0.5}$ *and variable Y takes any value* (short: $(x_{0.5}, y_{\max})$).

B: *You receive the same fixed amount, z, if the outcome of variable X is less than* $x_{0.75}$ *and the outcome of variable Y is less than* y_1 *(with* $y_1 < y_{\max}$; short: $(x_{0.75}, y_1)$).

The formulation has been altered to fit the wording of the earlier framings for elicitation questions in this review. The value for y_1 is specified and depending on the response of an expert, y_1 is adjusted until the expert is indifferent between the two gambles. If no indifference is achieved, the process ends after a predetermined number of iterations and upper and lower bounds for y_1 are specified to choose the midpoint. With the same framing, the experts continue choosing between binary deals while varying the quantiles for X and values of y_n, such as **A:**$(x_{0.75}, y_1)$ and **B:**$(x_{0.9}, y_2)$ and so forth. Through enough iterations, i.e. a sufficient number of indifferent choices that determine the points on the contour, its 50th percentile is assessed. Once this is achieved, the joint cumulative distribution of any point, $(x, y) \in [x_{\min}, x_{\max}] \times [y_{\min}, y_{\max}]$, can be derived with one additional assessment of a univariate quantity such as a marginal probability for any of the variables of interest, $F_x(x)$, by finding the point $(x_1, y_{\max})$ lying on its isoprobability contour. The joint probability assessment reduces then to a univariate problem through $F(x, y) = F(x_1, y_{\max}) \triangleq F_x(x_1)$.

This approach was tested with graduate students who assessed the joint probability of weight and height relationships within their university cohort. A monetary incentive was offered for obtaining honest and accurate answers. The authors conclude that this method is sensible with respect to difficulty, monotonicity and accuracy, but still discuss some possible assumptions that might ease the assessment burden. As a main advantage over conventional methods they mention the flexibility in analysing the results by deriving various dependence measures from the elicited outcomes.

Another method that has been proposed for specifying dependence through expert judgements and which fits into this sub-section is Papathomas and O'Hagan (2005). They consider a Bayesian updating procedure for dependent binary random variables. Again, dependence assessments are not made directly, but a threshold copula approach is used to fully determine the dependence structure.

6. Aggregation of dependence assessments

As we typically elicit judgements from more than one expert in order to obtain a broader perspective on the uncertainties of interest, concerns around the aggregation of multiple expert opinions also influence the decision of which dependence parameter to elicit. Broadly, two groups of aggregation methods exist, behavioural and mathematical ones. Behavioural ways seek consensus among the experts while mathematical methods use a weighting scheme for the combination. Typically, mathematical aggregation is preferred to avoid shortcomings of the first, such as individual experts dominating (or even dictating) the assessment result. A potential issue that might occur with mathematical aggregation in dependence elicitation is however that not all dependence assessments are preserved. While for instance a linear combination of correlation matrices still is a correlation matrix, conditional independencies, such as specified in a BN, will not be preserved.

When combining experts' assessments mathematically, mainly two methods are considered: Bayesian aggregation which might account for biases (e.g. overconfidence) and pooling methods which are seen as more robust and easier to use (Hora & Kardeş, 2015). The latter are discussed in more detail given their explicit consideration when aggregating dependence judgements. Generally, a pooling function is a weighted combination of individual judgements. Experts are assigned weights either equally or so that the weights reflect their competence (all weights are non-negative and sum to one). The most common types of pooling functions are linear and geometric. In the theoretical literature, both types are justified on axiomatic grounds (Dietrich and List, in press; McConway, 1981). However, in the context of aggregating dependence assessments, it might be considered problematic that these pooling methods are not compatible with probabilistic independence preservation. This independence property ensures that if all experts agree for two variables to be (conditionally) independent, then this is reflected in the combined assessment. Yet, unless independence is justified on structural grounds as well (e.g. through a graphical dependence representation) and is therefore not purely accidental, this normative requirement is questionable (Bradley, Dietrich, & List, 2014). As shown, often dependence parameters are elicited in a modelling process in which structural judgements, such as directed acyclic graphs, are included and therefore we take the position that both sources of information are respected and pooling methods can be regarded as valid combination functions. For other models, the structural information in form of functional

dependence might be assessed separately and prior to the quantitative assessment in the elicitation process.

Linear pooling: equal weighting. One way of pooling experts' assessments is by equally weighting their estimates (i.e. averaging them). Equal weighting of several (directly) elicited correlations was found to increase statistical accuracy when distance to empirical data was measured (Winkler & Clemen, 2004). The authors tested the robustness of their conclusions by removing/adding experts from/to the pool and found that the mean average error (MAE) decreased as the number of experts increased.

Linear pooling: performance-based weighting. In the same study, Winkler and Clemen (2004) show that taking the average of only the top performing cohort of experts as measured by the MAE reduces the overall error considerably (calculated when averaging the entire set of estimates). This finding is consistent with expert judgement studies for univariate quantities (Cooke & Goossens, 2008) and motivated the idea of developing a measure of calibration to assess experts' performance in terms of statistical accuracy for multivariate assessments. Note that there is some indication that a common calibration method for univariate expert judgements (Cooke, 1991) was shown not to be feasible for aggregating dependence assessments (Morales Nápoles et al., 2013).

The first and only calibration score for multivariate assessments (to the authors' knowledge) is the dependence calibration score introduced in Morales Nápoles and Worm (2013) which makes use of the *Hellinger distance*. In order to assess this score, seed variables known to the facilitator/analyst but not the experts are elicited in addition to the target variables. This is similar to Cooke's Classical model (Cooke, 1991). For two bivariate copulas, f_C (a copula model used for calibration purposes) and f_E (a copula as estimated by expert opinions), the Hellinger distance H is then:

$$H(f_C, f_E) = \iint_{[0,1]^2} \sqrt{\frac{1}{\sqrt{2}}(\sqrt{f_C(u,v)} - \sqrt{f_E(u,v)})^2}\mathrm{dudv}$$

In Abou-Moustafa, De La Torre, and Ferrie (2010) an overview of different distances between distributions is given. If the distributions are Gaussian, these distances can be written in terms of the mean and covariance matrix, i.e. the parameters of the Gaussian distribution. Under the Gaussian copula assumption, H might be parameterised by two correlation matrices:

$$H_G(\Sigma_C, \Sigma_E) = \sqrt{1 - \frac{\det(\Sigma_C)^{1/4}\det(\Sigma_E)^{1/4}}{(1/2\det(\Sigma_C) + 1/2\det(\Sigma_E))^{1/2}}}$$

where Σ_C is a correlation matrix used for calibration purposes and Σ_E the matrix derived from experts' assessments. The dependence calibration score is then:

$$D = 1 - H$$

The score is 1 if an expert's assessment corresponds to the calibration model exactly. Conversely, it differs from 1 as the expert's assessment differs from the calibration model. Under the Gaussian assumption, i.e. when using H_G , the score approaches 1 as Σ_E approximates Σ_C elementwise and the score decreases as H_G differs from H_C elementwise. A score equal to 0 means that at least two variables are linearly dependent in the correlation matrix used for calibration purposes and the expert fails to express this. Or contrary to this, an expert expresses perfect linear dependence between two variables when this is not the case. For more details, see Morales Nápoles, Worm, Hanea, and Kalkman (2016). In the same study (Morales Nápoles et al., 2016), the authors extend the method discussed in Morales Nápoles and Worm (2013). They use the Hellinger distance to compare a Gumbel copula generated from precipitation data with a copula constructed from experts' assessments of tail dependence between rain amount and duration (the way to obtain these estimates is discussed in Morales Nápoles et al. (2008)). For that study, a combination of expert opinions based on the dependence calibration score outperformed individual expert opinions. Further, it is shown that experts with highest calibration scores for univariate assessments were not the experts with the highest dependence calibration score.

In order to combine dependence assessments, experts are weighted according to their dependence calibration score. Similar to the univariate case, a cut-off level is established, either chosen by the analyst or by optimising the performance of the combination. If an individual expert falls below this level, their score will be unweighted for the pooling function.

7. Dependence elicitation in the empirical literature

Following the previous discussions about elicitation in various modelling contexts and about forms of elicited dependence parameters, this section provides an overview of the common approaches in practice that are prevalent in the case study literature.

While a complete outline of our review methodology can be found in Appendix A, we briefly present how the literature on eliciting dependence has been reviewed. The objective for this literature review is two-fold:

1. Assess the application areas and approaches to dependence modelling that are used in case studies published in the literature, in order to evaluate the reach of the different elicitation methods.
2. Ensure that the theoretical review is complete and includes a broad variety of perspectives.

As a first step, a search strategy was formulated that defined the key words used in order to ensure a thorough search of potential references of interest. For this, we started combining common key words of expert judgement studies such as "expert judgement (British English)/judgment (American English)" or "elicitation" itself, with general key words of dependence elicitation and modelling. This was refined by including key words for specific dependence modelling techniques and dependence parameters. Next, appropriate databases were identified, again starting generally before searching explicitly in archives of the topic's research areas, such as Operational Research and Decision as well as Risk Analysis. For evaluating the relevance of references under equal principles, criteria that specify the fit to this review (and which are outlined completely in Appendix A) had to be defined. The candidate references were then filtered and lastly, the selected findings were distinguished between theoretical and practical contributions as the latter were categorised for the overview in this section.

In total 53 references have been identified in which dependence has been elicited within decision analysis/risk analysis case studies (in some, more than one dependence parameter was elicited). The elicited dependence parameters are categorised as conditional (exceedance) probabilities (CP/CEP), point estimates as well as quantiles, joint probabilities, statistical parameters such as correlation coefficients, verbal and other methods (whereas other methods here differ from the ones presented in Section 5.3). A detailed list of the identified case studies can be found in the additional Supplementary material. The empirical references were investigated from different perspectives and Fig. 6 summarises how the empirical literature is clustered.

In the upper-left corner it can be seen that the predominant dependence model for which dependence is elicited is a BN (61.02%). For that, the main dependence parameters elicited are conditional (exceedance) probabilities (point estimate) and verbal scales. Dependence is elicited much less frequently for copulas, BLM approaches or parametric multivariate distributions.

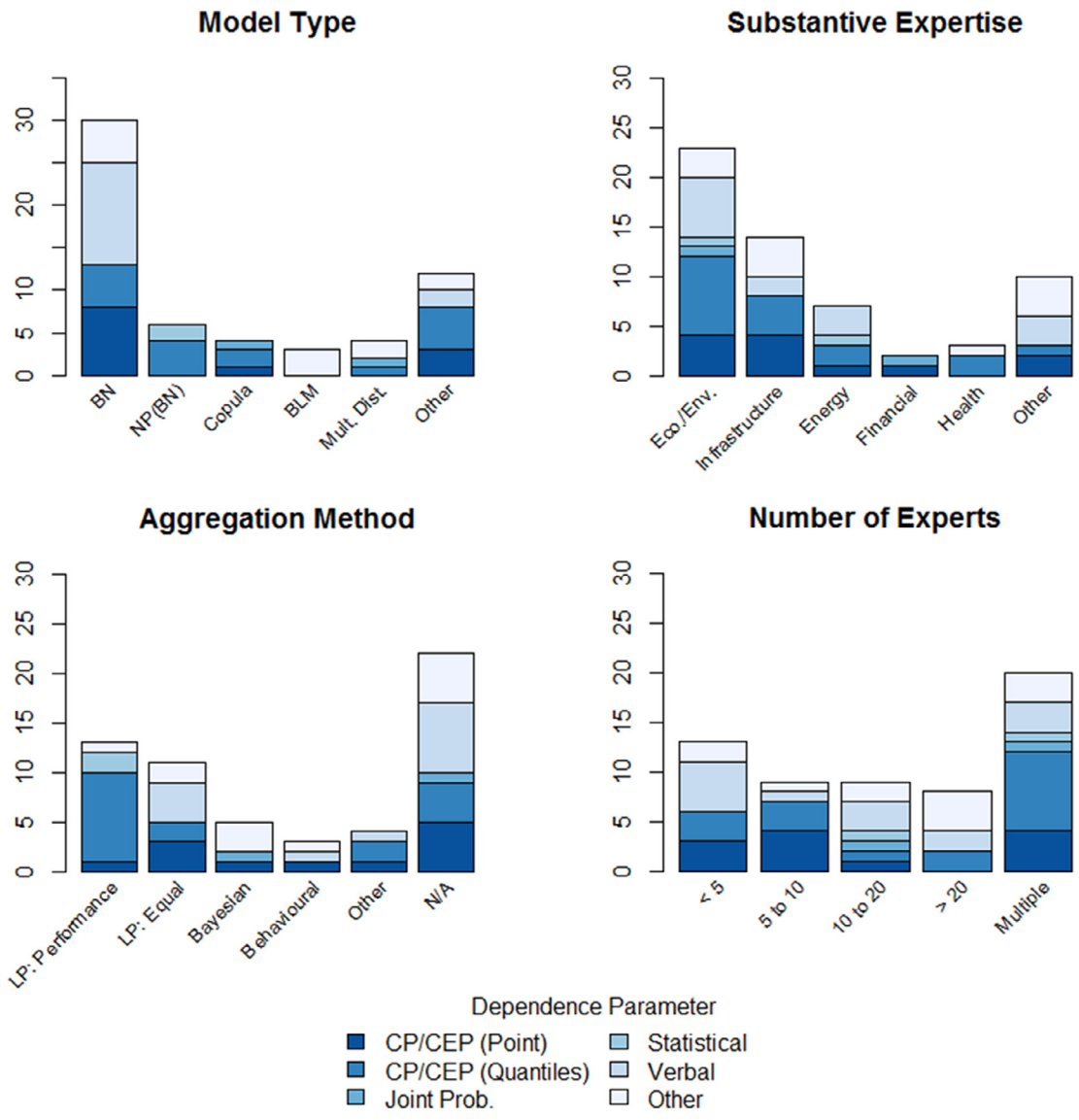

Fig. 6. Different perspectives on elicited dependence parameters' use in the case study literature.

For dependence parameters per aggregation method an apparent finding is that performance-based methods are used mainly together with conditional (exceedance) probabilities (through quantile assessments). This might not be surprising given that the authors for these studies come from the same expert judgement school that emphasises the use of performance-based combination and quantile (rather than point) assessment. In total performance-based weighting is used in 22.03% of all case studies, just more than equal weighting which is used in 18.64% of all references. Most significant however is that for 37.28% of all case studies the aggregation method is not described or mentioned at all.

When clustering the experts' domains and substantive expertise (upper-right corner), it is shown that in particular for environmental and ecological studies as well as in risk analyses for infrastructure problems, dependence is elicited through probabilistic variables (CP/CEP), point and quantile assessments, together with verbal methods. Overall, the main domains that experts have substantive expertise in are environmental/ecological (38.98%), infrastructure (23.72%) and energy decision analysis/risk analysis (11.86%). In this context, it is an interesting observation that the relevant case studies (see Supplementary material) are mostly published in domain-specific journals rather than journals with a focus on the modelling and hence elicitation methodology. This gives a few indications about the status quo of the empirical side of the research problem addressed in this review. It shows that modelling dependence together with expert judgement for quantification is a research problem that is (actually) recognised in the identified domains. Interestingly, the domains have an established tradition of applying rigorous risk analysis methods, often stemming from the area of probabilistic risk analysis (Bedford & Cooke, 2001). Further, this finding indicates that due to a focus on the application in the fields, there is less focus on developing new theory for dependence modelling and elicitation which would be found in journals with a methodological focus. This allows for cross-fertilisation of various findings discussed in the previous sections and our review aims to establish a contribution for this.

While a recommended number of experts from marginal elicitation protocols is between 5 and 10 experts (see aforementioned references on guidance for univariate elicitation), for dependence elicitation this is taken into consideration only in 15.25% of the cases. Slightly more often (22.03%), less than five experts are used. Again, the predominant percentage (33.89%) for "Multiple" implies a less clear documentation.

While these findings are not conclusive they offer an indication on the predominant approaches in the case study literature.

8. Conclusions and further research

We have argued that multivariate decision models under uncertainty are becoming more and more prevalent' whether as BNs (continuous or discrete), as parametric multivariate models, or as separate specifications of univariate distributions together with copulas to model the dependencies. We also argued that this immediately leads to the need for elicitation techniques to quantify these models.

The biggest challenge in the use of expert judgement to quantify dependence is in the way we manage the elicitation burden for experts. Implicit in our discussion above is that the elicitation burden has two key dimensions:

- The required quantity of information—there is a danger that large amounts of information required from experts will burden them too much in terms of time and the prolonged intensity of the task.
- The complexity of the required information—there is a danger that the experts might not be able to hold all the required information in the forefront of their minds while considering complex scenarios in which (conditional) probabilities are required.

Both considerations should guide the analyst to choose between ways to reduce the elicitation burden, by: simplifying the parameterisations of models, by considering the qualitative and quantitative steps of elicitation separately, or by finding ways of explaining in practical terms the quantities that are being elicited. However, there is a clear trade-off between easing the elicitation burden and building models that replicate the important behaviour of real world systems. Satisfying both the above requirements is challenging and under research.

The qualitative structure provided by a Bayesian network is one example in this direction. However, often it is difficult to decide on a particular form of network. We may have situations, for example, where a multivariate distribution can be estimated from data for moderate values of the variables, but where qualitatively different behaviour can occur in the tails. Expert judgement may be more appropriate in this context, as stochastic behaviour is then driven by different relationships between variables.

The literature review illustrates clearly the challenge faced in finding better ways to elicit multivariate uncertainties: In many cases the reported studies use students instead of (costly) experts. Often, when experts are used, they are asked to only provide guidance on parameters, but the justification for the chosen parametric family is not given. Clearly, for purposes of validity and verification we need to evolve better practices in selecting such families. Otherwise we are not in a strong position to challenge poor operational practice, such as the prevalence of the Gaussian copula used widely in financial modelling prior to the recent crash, and almost certainly still in equally wide use (Salmon, 2009).

Finally, in the paper we have focused on the use of expert assessment in quantifying multivariate distributions. However, the revolution in data analytics is using machine-learning and expert systems rather than human experts. It is therefore worth reflecting on the relative benefits, similarities and complementarities of these approaches. An individual human expert may be considered analogous to a particular machine-learning model, and the empirical result that machine-learning model averaging typically gives better results than any one of the models on their own, reflects older observations in the use of expert judgement that weighted averages of expert assessments are better calibrated than individual experts. However, the human expert may be able to provide simplifications through parametric model choices, and insights into model "phase changes" that the machine-learning models struggle with, because the data does not go far enough into the tail. The research challenges we have set out above will help us find a more satisfactory approach to combining human and machine expert judgements for uncertainty modelling.

Acknowledgements

The authors would like to thank the European Cooperation in Science and Technology, COST Action IS1304 - Expert Judgement Network, which allowed them to meet in person on various research meetings and which supported the first author to spend a Short Term Scientific Mission at TU Delft.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at 10.1016/j.ejor.2016.10.018.

References

Abbas, A. E. (2009). Multiattribute utility copulas. *Operations Research, 57*(6), 1367–1383.
Abbas, A. E., Budescu, D. V., & Gu, Y. (2010). Assessing joint distributions with isoprobability contours. *Management Science, 56*(6), 997–1011.
Abou-Moustafa, K. T., De La Torre, F., & Ferrie, F. P. (2010). Designing a metric for the difference between Gaussian densities. In *Brain, body and machine* (pp. 57–70). Springer.
Al-Awadhi, S. A., & Garthwaite, P. H. (1998). An elicitation method for multivariate normal distributions. *Communications in Statistics—Theory and Methods, 27*(5), 1123–1142.
Al-Awadhi, S. A., & Garthwaite, P. H. (2001). Prior distribution assessment for a multivariate normal distribution: An experimental study. *Journal of Applied Statistics, 28*(1), 5–23.
Al-Awadhi, S. A., & Garthwaite, P. H. (2006). Quantifying expert opinion for modelling fauna habitat distributions. *Computational Statistics, 21*(1), 121–140.
Allan, L. G. (1980). A note on measurement of contingency between two binary variables in judgment tasks. *Bulletin of the Psychonomic Society, 15*(3), 147–149.
Angilella, S., Greco, S., Lamantia, F., & Matarazzo, B. (2004). Assessing non-additive utility for multicriteria decision aid. *European Journal of Operational Research, 158*(3), 734–744.
Arbenz, P., & Canestraro, D. (2012). Estimating copulas for insurance from scarce observations, expert opinion and prior information: A Bayesian approach. *Astin Bulletin, 42*(01), 271–290.
Balakrishnan, N., & Nevzorov, V. B. (2004). *A primer on statistical distributions*. USA: John Wiley & Sons.
Bedford, T. (2002). *Interactive expert assignment of minimally-informative copulae*. Citeseer.
Bedford, T., & Cooke, R. M. (2001). *Probabilistic risk analysis: Foundations and methods*. UK: Cambridge University Press.
Bedford, T., Daneshkhah, A., & Wilson, K. J. (2016). Approximate uncertainty modeling in risk analysis with vine copulas. *Risk Analysis, 36*(4), 792–815.
Bedford, T., Denning, R., Revie, M., & Walls, L. (2008). Applying Bayes linear methods to support reliability procurement decisions. In *Reliability and maintainability symposium, 2008. RAMS 2008. Annual* (pp. 341–346). IEEE.
Bedrick, E. J., Christensen, R., & Johnson, W. (1996). A new perspective on priors for generalized linear models. *Journal of the American Statistical Association, 91*(436), 1450–1460.
Beyth-Marom, R. (1982). Perception of correlation re-examined. *Memory and Cognition, 10*(6), 511–519.
Błaszczyński, J., Greco, S., & Słowiński, R. (2007). Multi-criteria classification–a new scheme for application of dominance-based decision rules. *European Journal of Operational Research, 181*(3), 1030–1044.
Blomqvist, N. (1950). On a measure of dependence between two random variables. *The Annals of Mathematical Statistics, 21*(4), 593–600.
Böcker, K., Crimmi, A., & Fink, H. (2010). Bayesian risk aggregation: Correlation uncertainty and expert judgement. In K. Böcker (Ed.), *Rethinking risk measurement and reporting—Volume I*. UK: Risk Books.
Bolger, F., & Wright, G. (1994). Assessing the quality of expert judgment: Issues and analysis. *Decision Support Systems, 11*(1), 1–24.
Bradley, R., Dietrich, F., & List, C. (2014). Aggregating causal judgments. *Philosophy of Science, 81*(4), 491–515.
Bunea, C., & Bedford, T. (2002). The effect of model uncertainty on maintenance optimization. *IEEE Transactions on Reliability, 51*(4), 486–493.
Chaloner, K., Church, T., Louis, T. A., & Matts, J. P. (1993). Graphical elicitation of a prior distribution for a clinical trial. *Journal of the Royal Statistical Society: Series D (The Statistician), 42*(4), 341–353.
Chaloner, K., & Duncan, G. T. (1987). Some properties of the Dirichlet-multinomial distribution and its use in prior elicitation. *Communications in Statistics—Theory and Methods, 16*(2), 511–523.

Chapman, L. J., & Chapman, J. P. (1969). Illusory correlation as an obstacle to the use of valid psychodiagnostic signs. *Journal of Abnormal Psychology, 74*(3), 271.

Choy, S. L., O'Leary, R., & Mengersen, K. (2009). Elicitation by design in ecology: Using expert opinion to inform priors for Bayesian statistical models. *Ecology, 90*(1), 265–277.

Clemen, R. T., Fischer, G. W., & Winkler, R. L. (2000). Assessing dependence: Some experimental results. *Management Science, 46*(8), 1100–1115.

Clemen, R. T., & Reilly, T. (1999). Correlations and copulas for decision and risk analysis. *Management Science, 45*(2), 208–224.

Cooke, R. M. (1991). *Experts in uncertainty: opinion and subjective probability in science*. USA: Oxford University Press.

Cooke, R. M. (1994). Parameter fitting for uncertain models: Modelling uncertainty in small models. *Reliability Engineering & System Safety, 44*(1), 89–102.

Cooke, R. M. (2004). The anatomy of the squizzel: The role of operational definitions in representing uncertainty. *Reliability Engineering & System Safety, 85*(1), 313–319.

Cooke, R. M. (2013). Uncertainty analysis comes to integrated assessment models for climate change... and conversely. *Climatic change, 117*(3), 467–479.

Cooke, R. M. (2014). Validating expert judgment with the classical model. In *Experts and consensus in social science* (pp. 191–212). Springer.

Cooke, R. M., & Goossens, L. L. (2008). Tu Delft expert judgment data base. *Reliability Engineering & System Safety, 93*(5), 657–674.

Cooke, R. M., & Kelly, G.-N. (2010). Climate change uncertainty quantification: Lessons learned from the joint EU-USNRC project on uncertainty analysis of probabilistic accident consequence codes. *Resources for the Future Discussion Paper, 5*, 10–29.

Cooke, R. M., & Kraan, B. (1996). Dealing with dependencies in uncertainty analysis. In *Probabilistic safety assessment and management* (pp. 625–630). Springer.

Daneshkhah, A., & Oakley, J. (2010). Eliciting multivariate probability distributions. In K. Böcker (Ed.), *Rethinking risk measurement and reporting—Volume i*. UK: Risk Books.

Darwiche, A. (2009). *Modeling and reasoning with Bayesian networks*. USA: Cambridge University Press.

DeGroot, M. H., & Fienberg, S. E. (1983). The comparison andevaluation of forecasters. *Journal of the Royal Statistical Society. Series D (The Statistician), 32*(1/2), 12–22.

Delgado-Hernández, D.-J., Morales-Nápoles, O., De-León-Escobedo, D., & Arteaga-Arcos, J.-C. (2014). A continuous Bayesian network for Earth dams' risk assessment: An application. *Structure and Infrastructure Engineering, 10*(2), 225–238.

Dickey, J., Lindley, D. V., & Press, S. J. (1985). Bayesian estimation of the dispersion matrix of a multivariate normal distribution. *Communications in Statistics—Theory and Methods, 14*(5), 1019–1034.

Dietrich, F., & List, C. (2016). Probabilistic opinion pooling. In A. Hajek, & C. Hitchcock (Eds.), *The Oxford Handbook of Probability and Philosophy* (pp. 179–207). UK: Oxford Handbooks.

Diez, F. J. (1993). Parameter adjustment in Bayes networks. The generalized noisy or-gate. In *Proceedings of the ninth international conference on uncertainty in artificial intelligence* (pp. 99–105). USA: Morgan Kaufmann Publishers Inc.

Druzdel, M., & Van Der Gaag, L. C. (2000). Building probabilistic networks: "Where do the numbers come from?". *IEEE Transactions on Knowledge and Data Engineering, 12*(4), 481–486.

Druzdzel, M. J., & Van Der Gaag, L. C. (1995). Elicitation of probabilities for belief networks: Combining qualitative and quantitative information. In *Proceedings of the eleventh conference on uncertainty in artificial intelligence* (pp. 141–148). USA: Morgan Kaufmann Publishers Inc.

Dubois, D., Prade, H., & Sabbadin, R. (2001). Decision-theoretic foundations of qualitative possibility theory. *European Journal of Operational Research, 128*(3), 459–478.

Durante, F., & Sempi, C. (2015). *Principles of copula theory*. USA: CRC Press.

Ehrgott, M., Klamroth, K., & Schwehm, C. (2004). An MCDM approach to portfolio optimization. *European Journal of Operational Research, 155*(3), 752–770.

Elfadaly, F. G., & Garthwaite, P. H. (2013). Eliciting Dirichlet and Connor–Mosimann prior distributions for multinomial models. *Test, 22*(4), 628–646.

European Food and Safety Authority (EFSA) (2014). Guidance on expert knowledge elicitation in food and feed safety risk assessment. *EFSA Journal, 12*(6), 1–278.

Fackler, P. L. (1991). Modeling interdependence: An approach to simulation and elicitation. *American Journal of Agricultural Economics, 73*(4), 1091–1097.

Farrow, M. (2003). Practical building of subjective covariance structures for large complicated systems. *Journal of the Royal Statistical Society: Series D (The Statistician), 52*(4), 553–573.

Farrow, M., Goldstein, M., & Spiropoulos, T. (1997). Developing a Bayes linear decision support system for a brewery. In S. French, & J. Q. Smith (Eds.), *The practice of Bayesian analysis* (pp. 71–106).

Fenton, N. E., Neil, M., & Caballero, J. G. (2007). Using ranked nodes to model qualitative judgments in Bayesian networks. *Transactions on Knowledge and Data Engineering, 19*(10), 1420–1432.

Figueira, J. R., Greco, S., & Słowiński, R. (2009). Building a set of additive value functions representing a reference preorder and intensities of preference: Grip method. *European Journal of Operational Research, 195*(2), 460–486.

Flari, V., Chaudhry, Q., Neslo, R., & Cooke, R. (2011). Expert judgment based multi-criteria decision model to address uncertainties in risk assessment of nanotechnology-enabled food products. *Journal of Nanoparticle Research, 13*(5), 1813–1831.

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

Garthwaite, P. H., & Al-Awadhi, S. A. (2001). Non-conjugate prior distribution assessment for multivariate normal sampling. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 63*(1), 95–110.

Garthwaite, P. H., Al-Awadhi, S. A., Elfadaly, F. G., & Jenkinson, D. J. (2013). Prior distribution elicitation for generalized linear and piecewise-linear models. *Journal of Applied Statistics, 40*(1), 59–75.

Garthwaite, P. H., & Dickey, J. M. (1991). An elicitation method for multiple linear regression models. *Journal of Behavioral Decision Making, 4*(1), 17–31.

Garthwaite, P. H., Kadane, J. B., & O'Hagan, A. (2005). Statistical methods for eliciting probability distributions. *Journal of the American Statistical Association, 100*(470), 680–701.

Genest, C., & Favre, A.-C. (2007). Everything you always wanted to know about copula modeling but were afraid to ask. *Journal of Hydrologic Engineering, 12*(4), 347–368.

Genest, C., Gendron, M., & Bourdeau-Brien, M. (2009). The advent of copulas in finance. *The European Journal of Finance, 15*(7–8), 609–618.

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

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

Gosling, J. P., Hart, A., Owen, H., Davies, M., Li, J., MacKay, C., et al. (2013). A Bayes linear approach to weight-of-evidence risk assessment for skin allergy. *Bayesian Analysis, 8*(1), 169–186.

Grabisch, M. (1996). The application of fuzzy integrals in multicriteria decision making. *European Journal of Operational Research, 89*(3), 445–456.

Greco, S., Matarazzo, B., & Słowiński, R. (2001). Rough sets theory for multicriteria decision analysis. *European Journal of Operational Research, 129*(1), 1–47.

Greco, S., Matarazzo, B., & Słowiński, R. (2004). Axiomatic characterization of a general utility function and its particular cases in terms of conjoint measurement and rough-set decision rules. *European Journal of Operational Research, 158*(2), 271–292.

Greco, S., Mousseau, V., & Słowiński, R. (2014). Robust ordinal regression for value functions handling interacting criteria. *European Journal of Operational Research, 239*(3), 711–730.

Gredebäck, G., Winman, A., & Juslin, P. (2000). Rational assessments of covariation and causality. In *Proceedings of the 22nd annual conference of the cognitive science society* (pp. 190–195).

Hanea, A., Napoles, O. M., & Ababei, D. (2015). Non-parametric Bayesian networks: Improving theory and reviewing applications. *Reliability Engineering & System Safety, 144*, 265–284.

Hänninen, M., Banda, O. A. V., & Kujala, P. (2014). Bayesian network model of maritime safety management. *Expert Systems with Applications, 41*(17), 7837–7846.

Hastie, R. (2016). Causal thinking in judgments. In G. Keren, & W. G (Eds.), *The Wiley Blackwell handbook of judgment and decision making* (pp. 590–628). USA: John Wiley & Sons.

Henrion, M. (1989). Some practical issues in constructing belief networks. In *Proceedings of the third conference on uncertainty in artificial intelligence* (pp. 161–173). Elsevier Science Publishing Company.

Hoffrage, U., Lindsey, S., Hertwig, R., & Gigerenzer, G. (2000). Communicating statistical information. *Science, 290*(5500), 2261–2262.

Hora, S. C., & Kardeş, E. (2015). Calibration, sharpness and the weighting of experts in a linear opinion pool. *Annals of Operations Research, 229*(1), 429–450.

James, A., Choy, S. L., & Mengersen, K. (2010). Elicitator: An expert elicitation tool for regression in ecology. *Environmental Modelling & Software, 25*(1), 129–145.

Jenkinson, D. (2005). The elicitation of probabilities: A review of the statistical literature. *Technical report*. Citeseer.

Joe, H. (2014). *Dependence modeling with copulas*. USA: CRC Press.

Jose, V. R. R., Nau, R. F., & Winkler, R. L. (2008). Scoring rules, generalized entropy, and utility maximization. *Operations Research, 56*(5), 1146–1157.

Jose, V. R. R., Nau, R. F., & Winkler, R. L. (2009). Sensitivity to distance and baseline distributions in forecast evaluation. *Management Science, 55*(4), 582–590.

Kadane, J., & Wolfson, L. J. (1998). Experiences in elicitation. *Journal of the Royal Statistical Society: Series D (The Statistician), 47*(1), 3–19.

Kadane, J. B., Dickey, J. M., Winkler, R. L., Smith, W. S., & Peters, S. C. (1980). Interactive elicitation of opinion for a normal linear model. *Journal of the American Statistical Association, 75*(372), 845–854.

Keefer, D. L., & Bodily, S. E. (1983). Three-point approximations for continuous random variables. *Management Science, 29*(5), 595–609.

Keeney, R. L., & von Winterfeldt, D. (1991). Eliciting probabilities from experts in complex technical problems. *IEEE Transactions on Engineering Management, 38*(3), 191–201.

Kotz, S., & Van Dorp, J. R. (2010). Generalized diagonal band copulas with two-sided generating densities. *Decision Analysis, 7*(2), 196–214.

Kraan, B. (2002). *Probabilistic inversion in uncertainty analysis: And related topics* (Ph.D. thesis).. Delft University of Technology.

Kraan, B., & Bedford, T. (2005). Probabilistic inversion of expert judgments in the quantification of model uncertainty. *Management Science, 51*(6), 995–1006.

Kraan, B., & Cooke, R. (1997). Post-processing techniques for the joint CEC/USNRC uncertainty analysis of accident consequence codes. *Journal of Statistical Computation and Simulation, 57*(1–4), 243–259.

Kruskal, W. H. (1958). Ordinal measures of association. *Journal of the American Statistical Association, 53*(284), 814–861.

Kunda, Z., & Nisbett, R. E. (1986). The psychometrics of everyday life. *Cognitive Psychology, 18*(2), 195–224.

Kurowicka, D., & Cooke, R. M. (2006). *Uncertainty analysis with high dimensional dependence modelling*. UK: John Wiley & Sons.

Lad, F. (1996). *Operational subjective statistical methods: A mathematical, philosophical, and historical introduction*: Vol. 315. USA: Wiley-Interscience.

Lurie, P. M., & Goldberg, M. S. (1998). An approximate method for sampling correlated random variables from partially-specified distributions. *Management Science, 44*(2), 203–218.

Marichal, J.-L. (2004). Tolerant or intolerant character of interacting criteria in aggregation by the Choquet integral. *European Journal of Operational Research, 155*(3), 771–791.

McConway, K. J. (1981). Marginalization and linear opinion pools. *Journal of the American Statistical Association, 76*(374), 410–414.

Meeuwissen, A. M. H., & Bedford, T. (1997). Minimally informative distributions with given rank correlation for use in uncertainty analysis. *Journal of Statistical Computation and Simulation, 57*(1–4), 143–174.

Meyer, J., Taieb, M., & Flascher, I. (1997). Correlation estimates as perceptual judgments. *Journal of Experimental Psychology: Applied, 3*(1), 3.

Mitchell, C. J., De Houwer, J., & Lovibond, P. F. (2009). The propositional nature of human associative learning. *Behavioral and Brain Sciences, 32*(02), 183–198.

Mkrtchyan, L., Podofillini, L., & Dang, V. N. (2015). Bayesian belief networks for human reliability analysis: A review of applications and gaps. *Reliability Engineering & System Safety, 139*, 1–16.

Moala, F. A., & O'Hagan, A. (2010). Elicitation of multivariate prior distributions: A nonparametric Bayesian approach. *Journal of Statistical Planning and Inference, 140*(7), 1635–1655.

Morales Nápoles, O. (2010). *Bayesian belief nets and vines in aviation safety and other applications* (Ph.D. thesis).. Delft University of Technology.

Morales Nápoles, O., Delgado-Hernández, D. J., De-León-Escobedo, D., & Arteaga-Arcos, J. C. (2014). A continuous Bayesian network for earth dams' risk assessment: Methodology and quantification. *Structure and Infrastructure Engineering, 10*(5), 589–603.

Morales Nápoles, O., Hanea, A., & Worm, D. (2013). Experimental results about the assessments of conditional rank correlations by experts: Example with air pollution estimates. In *Esrel 2013: Proceedings of the 22nd European safety and reliability conference" safety, reliability and risk analysis: Beyond the horizon"*. The Netherlands: CRC Press/Balkema-Taylor & Francis Group.

Morales Nápoles, O., Kurowicka, D., & Roelen, A. (2008). Eliciting conditional and unconditional rank correlations from conditional probabilities. *Reliability Engineering & System Safety, 93*(5), 699–710.

Morales Nápoles, O., & Worm, D. (2013). Hypothesis testing of multidimensional probability distributions. TNO report.

Morales Nápoles, O., Worm, D., Hanea, A., & Kalkman, I. (2016). Calibration and combination of expert's dependence estimates. (under review).

Morgan, M. G., Henrion, M., & Small, M. (1992). *Uncertainty: A guide to dealing with uncertainty in quantitative risk and policy analysis*. USA: Cambridge University Press.

Moskowitz, H., & Bunn, D. (1987). Decision and risk analysis. *European Journal of Operational Research, 28*(3), 247–260.

Moskowitz, H., & Sarin, R. K. (1983). Improving the consistency of conditional probability assessments for forecasting and decision making. *Management Science, 29*(6), 735–749.

Nadkarni, S., & Shenoy, P. P. (2004). A causal mapping approach to constructing Bayesian networks. *Decision Support Systems, 38*(2), 259–281.

Neslo, R., & Cooke, R. (2011). Modeling and validating stakeholder preferences with probabilistic inversion. *Applied Stochastic Models in Business and Industry, 27*(2), 115–130.

O'Hagan, A., Buck, C. E., Daneshkhah, A., Eiser, J. R., Garthwaite, P. H., Jenkinson, D. J., … Rakow, T. (2006). *Uncertain judgements: eliciting experts' probabilities*. UK: John Wiley & Sons.

O'Hagan, A., Glennie, E., & Beardsall, R. (1992). Subjective modelling and Bayes linear estimation in the UK water industry. *Journal of the Royal Statistical Society. Series C (Applied Statistics), 41*(3), 563–577.

O'Leary, R. A., Choy, S. L., Murray, J. V., Kynn, M., Denham, R., Martin, T. G., & Mengersen, K. (2009). Comparison of three expert elicitation methods for logistic regression on predicting the presence of the threatened brush-tailed rock-wallaby *Petrogale penicillata*. *Environmetrics, 20*(4), 379.

Ouchi, F. (2004). A literature review on the use of expert opinion in probabilistic risk analysis, *World Bank Policy Research Working Paper: 3201* (pp. 1–17).

Papathomas, M., & O'Hagan, A. (2005). Updating beliefs for binary variables. *Journal of Statistical Planning and Inference, 135*(2), 324–338.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. USA: Morgan Kaufmann.

Pearl, J. (2000). *Causality: Models, reasoning and inference*. USA: Cambridge University Press.

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

Percy, D. F. (2002). Bayesian enhanced strategic decision making for reliability. *European Journal of Operational Research, 139*(1), 133–145.

Percy, D. F. (2004). Subjective priors for maintenance models. *Journal of Quality in Maintenance Engineering, 10*(3), 221–227.

Regis, L. (2011). A Bayesian copula model for stochastic claims reserving, *University of Torino Discussion Paper: 227* (pp. 1–26).

Reilly, T. (2000). Sensitivity analysis for dependent variables. *Decision Sciences, 31*(3), 551–572.

Renooij, S. (2001). Probability elicitation for belief networks: issues to consider. *The Knowledge Engineering Review, 16*(03), 255–269.

Revie, M. (2008). *Evaluation of Bayes linear modelling to support reliability assessment during procurement* (Ph.D. thesis). University of Strathclyde.

Revie, M., Bedford, T., & 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.

Revie, M., Bedford, T., & Walls, L. (2011). Supporting reliability decisions during defense procurement using a Bayes linear methodology. *IEEE Transactions on Engineering Management, 58*(4), 662–673.

Ryan, T. P. (2008). *Modern regression methods*: Vol. 655. USA: John Wiley & Sons.

Salmon, F. (2009). The formula that killed wall street. *Wired, 3*(17), 16–20.

Savage, L. J. (1954). *The foundations of statistics*. USA: John Wiley & Sons.

Shanks, D. R. (2004). Judging covariation and causation. In D. J. Koehler, & N. Harvery (Eds.), *Blackwell handbook of judgment and decision making* (pp. 220–239). USA: Blackwell Publishing.

Shen, Z., Odening, M., & Okhrin, O. (2015). Can expert knowledge compensate for data scarcity in crop insurance pricing? *European Review of Agricultural Economics, 43*(2), 237–269.

Sklar, M. (1959). *Fonctions de répartition à n dimensions et leurs marges*. Université Paris 8.

Smedslund, J. (1963). The concept of correlation in adults. *Scandinavian Journal of Psychology, 4*(1), 165–173.

Smith, J. E., & Von Winterfeldt, D. (2004). Anniversary article: Decision analysis in management science. *Management Science, 50*(5), 561–574.

Spetzler, C. S., & Stael von Holstein, C.-A. S. (1975). Exceptional paper-probability encoding in decision analysis. *Management Science, 22*(3), 340–358.

Swain, A. D., & Guttmann, H. E. (1983). Handbook of human-reliability analysis with emphasis on nuclear power plant applications. Final report. *Technical report*. Albuquerque, NM (USA): Sandia National Labs.

van der Gaag, L. C., Renooij, S., Witteman, C. L., Aleman, B. M., & Taal, B. G. (1999). How to elicit many probabilities. In *Proceedings of the fifteenth conference on uncertainty in artificial intelligence* (pp. 647–654). USA: Morgan Kaufmann Publishers Inc.

Van Dorp, J. R. (2005). Statistical dependence through common risk factors: With applications in uncertainty analysis. *European Journal of Operational Research, 161*(1), 240–255.

Von Winterfeldt, D., & Fasolo, B. (2009). Structuring decision problems: A case study and reflections for practitioners. *European Journal of Operational Research, 199*(3), 857–866.

Wallenius, J., Dyer, J. S., Fishburn, P. C., Steuer, R. E., Zionts, S., & Deb, K. (2008). Multiple criteria decision making, multiattribute utility theory: Recent accomplishments and what lies ahead. *Management Science, 54*(7), 1336–1349.

Weisberg, S. (2005). *Applied linear regression*: Vol. 528. USA: John Wiley & Sons.

Winkler, R. L., & Clemen, R. T. (2004). Multiple experts vs. multiple methods: Combining correlation assessments. *Decision Analysis, 1*(3), 167–176.

Wisse, B., van Gosliga, S. P., van Elst, N. P., & Barros, A. I. (2008). Relieving the elicitation burden of Bayesian belief networks. In *BMA*.

Wunderlich, K., Symmonds, M., Bossaerts, P., & Dolan, R. J. (2011). Hedging your bets by learning reward correlations in the human brain. *Neuron, 71*(6), 1141–1152.

Zagorecki, A., & Druzdzel, M. J. (2004). An empirical study of probability elicitation under noisy-or assumption. In *Flairs conference* (pp. 880–886).

Zapata-Vázquez, R. E., O'Hagan, A., & Soares Bastos, L. (2014). Eliciting expert judgements about a set of proportions. *Journal of Applied Statistics, 41*(9), 1919–1933.