Analysis of Correlated Expert Judgments from Extended Pairwise Comparisons

Jason R. W. Merrick	Virginia Commonwealth University
J. Rene van Dorp	The George Washington University
Amita Singh	The George Washington University

Abstract

We develop a Bayesian multivariate analysis of expert judgment elicited using an extended form of pairwise comparisons. The method can be used to estimate the effect of multiple factors on the probability of an event and can be applied in risk analysis and other decision problems. The analysis provides predictions of the quantity of interest that incorporate dependencies amongst the various experts. In this form we may learn about the dependencies between the experts from their responses. The analysis is applied to a real data set of expert judgments elicited during the Washington State Ferries Risk Assessment. The effect of the statistical dependence amongst experts is compared to an analysis assuming independence amongst them.

Keywords: Expert judgment; Pairwise Comparisons; Bayesian statistics; Multivariate analysis.

1. Introduction

Many applications of expert judgment elicitation involve the estimation of the probabilities of events, with these probabilities sometimes affected by multiple factors. Merrick et al. (2000) propose an expert judgment elicitation method that estimates the effect of multiple factors on the probability of an event. This form of elicitation has been applied in the Prince William Sound (PWS) Risk Assessment (Merrick et al. 2002) and the Washington State Ferries Risk Assessment (WSF) (van Dorp et al. 2001) to estimate the probability of human error given organizational factors, such as the experience and training of the crew, and to estimate the probability of an accident given situational factors, such as the proximity and type of nearby vessels and the environmental conditions at the time. While the elicitation method was proposed for use in risk analysis, it can be applied in other decision situations where applicable data is lacking.

Clemen and Winkler (1999) review several models for combining experts' judgments about probabilities with the decision maker's prior information under the Bayesian aggregation framework developed in Morris (1974, 1977, 1983). It would seem natural to extend one of these techniques to incorporate the relevant factors. However, empirical research has shown that experts overestimate probabilities near zero (Cooke, 1991). In our previous risk assessment work in the maritime domain, we have found that experts are more comfortable assessing the relative probability of an event in two situations when these probabilities are low. Thus, the form of the elicitation in Merrick et al. (2000) asks the experts to assess the ratio of the probabilities of the event for the two scenarios. The multiple factors describe two scenarios to the expert in a meaningful manner and in each comparison one factor is changed between the two scenarios. The

method is akin to that in Bradley and Terry (1952), but the aim is to estimate the effect of the multiple factors rather than developing a ranking scale.

The occurrence and non-occurrence of the events is modeled by exchangeable Bernoulli trials with an unknown probability that depends on the factors describing the scenario. To link the probabilities to the factors, Merrick et al. (2000) assume a log-linear relationship. This assumption implies that the decision maker is interpreting the experts' responses through her assumed model. Merrick et al. (2000) use a classical multiple regression to assess the parameters of this relationship. Szwed et al. (2004) develop a Bayesian analysis of these expert judgments. However, the Bayesian analysis in Szwed et al. assumes that the responses of the experts are independent. While each of these analyses only provides predictions of the ratios of the probabilities of the event in two scenarios, actual probabilities of the event can be obtained if the probability of one scenario can be assessed, a reference scenario; the probability for another scenario is then found by multiplying the probability for the reference scenario by the ratio of the probabilities for the required scenario and the reference scenario.

There is a fundamental difference, however, between most Bayesian aggregation methods and our development. In most such methods, the experts are assessing an unknown quantity that is potentially, but not currently, observable by the decision maker. More formally Morris's framework requires that the likelihood for the experts' assessments should be conditioned on the observable quantities of interest. These assessments are then aggregated directly with the decision maker's prior beliefs. In our approach, the decision maker first assumes a log-linear model and then interprets the experts' assessments through this model. The decision maker aggregates estimates of the

parameters of the model based on each expert's assessments, but does not aggregate their assessments directly. Thus our development is more similar to that in Dawes (1979) in intention, where the behavior of the experts is considered in taking factors as cues in their assessments of probabilities or other unknown quantities. Thus we and Dawes model the experts' responses with a linear model that Dawes refers to as an improper, first-order approximation of the true relationship.

In this paper, we offer an extension of the Bayesian analysis in Szwed et al. (2004) to relax the assumption of independence between the experts. It is well accepted that the judgments of multiple experts can be correlated and that the treatment of these correlations is necessary for proper analysis of such data (Winkler 1981; French 1980 1981; Lindley 1983 1985; Mosleh et al. 1988; Clemen 1987; Clemen and Reilly, 1999; Jouini and Clemen 1996). Such correlation is often introduced, in the language of Clemen (1987), by overlapping information available to the experts and thus used in determining their responses to the questionnaires. Winkler (1981) and Clemen (1987) use a normal assessment error model to aggregate expert judgment concerning a continuous quantity.

We develop a Bayesian aggregation method (Morris, 1974) for this type of elicitation, but at the level of the parameters of our log-linear relationship. Combining the log-linear assumption with Winkler's multivariate normal error structure, our analysis takes the form of a multivariate regression analysis of the experts' responses and the factors in the questions. This form allows for correlation between the experts' responses. While the analysis mirrors the development of Bayesian multivariate regression (Press 1982), it is a special case as each expert is providing judgments on the same quantities, not different quantities as in the case of a full multivariate regression.

The development herein is part of an extension of the methodology for maritime risk assessment discussed in Merrick et al. (2000) to assess the uncertainty of accident frequency estimates. The expert judgment method is one part of this methodology; the other part is a simulation of the maritime transportation system that estimates the frequency of the occurrence of the factors that affect the accident probability. Merrick et al. (2004a) develop a Bayesian simulation method and a meta-model on the outputs of the simulation. Merrick et al. (2004b) combine the Bayesian simulation and its output meta-model with the development herein; the methodology is applied therein to a case study of proposed ferry service expansions in San Francisco Bay. This is one application of our expert judgment method and the Bayesian aggregation analysis developed here.

The outline of the paper is as follows. In Section 2, we illustrate the type of question used in our elicitation technique with an example drawn from a maritime risk study. We then illustrate the form of the underlying probability model assumed and show how this leads to regression as a suitable analysis methodology. Our multivariate extension is justified and developed in Section 3. The expert judgment data collected in the WSF Risk Assessment and the process used to collect it is described in Section 4 and this data is used to illustrate the use of our analysis method. We compare the analysis herein incorporating dependencies with one that assumes independence analogous to Szwed et al. (2004). Some concluding remarks are drawn in Section 5.

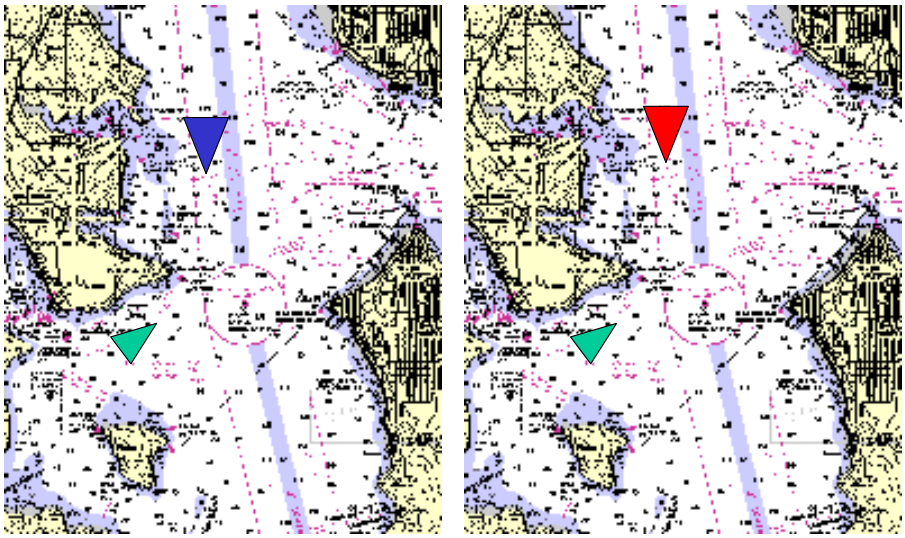
2. The Elicitation Method

2.1 The Questions

The following discussion will be based on the questionnaire used in the WSF Risk Assessment, although the same technique was used in the PWS Risk Assessment and can

be used in many risk analysis and general decision problems. As an example, we shall examine the questionnaire for the likelihood of a collision between a ferry and another vessel given that the ferry has suffered a navigational aid (radar) failure. To assess the probability of an accident, experts were asked to compare two situations, as shown in Figure 1.

Issaquah class ferry on the Bremerton to Seattle route in a crossing situation within 15 minutes, no other vessels around, good visibility, negligible wind.



Other vessel is a navy vessel Other vessel is a product tanker

Figure 1. An example of the type of question used in the expert judgement

This is essentially a pairwise comparison type of question (Bradley and Terry 1952). However, the questionnaires are used to estimate the effect of several factors, rather than the single factor in standard pairwise comparisons.

The questions ask the expert to consider two situations between which only one factor has changed. The basic situation in Figure 1 is an Issaquah class ferry traveling from Bremerton to Seattle on a clear day with no wind. There is another vessel crossing the bow of the ferry less than 1 mile away. In the situation on the left-hand side, the other

vessel is a Navy vessel, while on the right-hand side it is a product tanker. The questions were asked in the format of Figure 2. The responses were given on the scale at the bottom of Figure 2, which is taken from Saaty (1977).

Situation 1	**Attribute**	**Situation 2**
Issaquah	Ferry Class	-
SEA-BRE(A)	Ferry Route	-
Navy	1st Interacting Vessel	Product Tanker
Crossing	Traffic Scenario 1st Vessel	-
< 1 mile	Traffic Proximity 1st Vessel	-
No Vessel	2nd Interacting Vessel	-
No Vessel	Traffic Scenario 2nd Vessel	-
No Vessel	Traffic Proximity 2nd Vessel	-
> 0.5 Miles	Visibility	-
Along Ferry	Wind Direction	-
0	Wind Speed	-
Likelihood of Collision		
9 8 7 6 5 4 3 2 1 2 3 4 5 6 7 8 9		

Figure 2. An example of the question format

We instruct the experts to interpret their response as the ratio of the probability of an accident in the two situations pictured. If the expert circled a "1", the two probabilities would be equal. We assume that if the expert circled the "9" on the right (left) then the ratio of the probabilities would be 9 (1/9). The questionnaires are designed to obtain as much information as possible in the minimum number of questions.

Our design approach is similar to the idea of factorial designs, but with a difference due to our use of pairwise comparisons. A simple two-factor, two-level

factorial design can be pictured as a square. Each corner of the square is an experimental run (using the language of experimental design). Measuring the response at the bottom-left corner and the bottom-right corner allows us the estimate the size of the effect for one factor. Measuring the response at the top-left corner and comparing this to the bottom-left corner allows estimation of the effect for the other factor. Measuring the response at the top-right corner as well allows the estimation of an interaction term between the two factors. However, as we are performing pairwise comparisons, each question in our approach represents the *sides* of the square, asking for the ratio of the response for two corners. Thus rather than the requirement for four experimental runs to estimate two main effects and their interaction, we require three questions. These three questions compare (1) the bottom-left corner to the bottom-right corner; (2) the bottom-left corner to the top-left corner; and (3) the bottom-left corner to the top-right corner. Using this approach, we design our questionnaire to estimate the main effect of each individual factor and specific second-order interactions which we believe may exist. We can ask additional questions to assess center points as well.

2.2 Analyzing the Experts' Responses

The model assumed in the PWS and WSF Risk Assessments takes the form of a proportional probabilities model, based on the idea of the proportional hazards model (Cox 1972). Let $X = \left(x_1, \ldots, x_q\right)^T$ denote the q factors describing a situation in which the event of interest could occur. The conditional probability of the event, given the situation defined by X, is assumed to be

$$P(Event \mid X, p_0, \beta) = p_0 \exp\left(X^T \beta\right), \tag{1}$$

where $\beta = (\beta_1, \ldots, \beta_q)^T$ is a vector of q parameters and p_0 is a baseline probability parameter. Consider two situations defined by the factor vectors L and R. The ratio of the probabilities that the experts are assessing is then given by

$$\frac{P(Event \mid R, \beta)}{P(Event \mid L, \beta)} = \frac{p_0 \exp(R^T \beta)}{p_0 \exp(L^T \beta)} = \exp\left((R - L)^T \beta\right), \tag{2}$$

where (X_1 - X_2) denotes the difference vector between the two factor vectors. Thus, for this probability model, the ratio of the probabilities of the event given the two situations depends solely upon the difference between the two situations and the parameter vector β.

Each question asked the experts to assess the ratio of probabilities of the event (a collision) given the two situations. Multiple experts complete each questionnaire, so there are multiple responses to each question. Let the experts be indexed by $j(= 1, \ldots, p)$ and the questions be indexed by $i(= 1, \ldots, N)$, so the experts' responses can be denoted $z_{i,j}$. We now have that $z_{i,j}$ is the j-th expert's estimate of the ratio of probabilities for the i-th question, while the model gives this relative probability as $\exp(X_i^T \beta)$, where X_i is a vector representing the difference between the two situations in question i $(R_i - L_i)$ and L_i and R_i are the factor vectors for the left and right scenarios in the i-th question. This gives the basis for the regression equation used, specifically

$$\ln(z_{i,j}) = X_i^T \beta + u_{i,j} \tag{3}$$

where $u_{i,j}$ is the residual error term representing the variation between the experts' responses around the model.

Assuming that the errors $u_{i,j}$ are independent and normally distributed with zero mean and variance σ^2, this equation is a standard linear regression, where $y_{i,j} = \ln(z_{i,j})$ is the dependent variable, X_i is the vector of independent variables, β is a vector of regression parameters and $u_{i,j}$ is the error term. Clemen and Reilly (1999) observe that it is often necessary in expert judgment analysis to use such transformations to arrive at the normal distribution. Kadane et al. (1980) develop a method for assessing prior hyperparameters on a linear regression model; however, their approach is based on direct assessments rather than pairwise comparisons. A conjugate Bayesian analysis of (3) is developed in Szwed et al. (2004) assuming conditional independence of the experts' responses given the model parameters. Pulkkinen (1993 1994a 1994b) was first to introduce, to the best of our knowledge, a Bayesian analysis of pairwise comparisons, but his Bayesian paired comparison inference model also assumed independence amongst experts.

The pairwise comparisons made by the experts are used to assess a distribution for β only. This raises an interesting question. This method can only be used to estimate ratios of probabilities for two scenarios. How then does a decision maker obtain the actual probability of an event with particular values for the factors for use in decision making? The decision maker can assess the probability for one reference scenario, denoted X_0. Suitable techniques for aggregation of probability assessments are reviewed in Clemen and Winkler (1999). The probabilities for another scenario, X_*, can be found by multiplying

$$P(Event \mid X_0) \times \frac{P(Event \mid X_*, \beta)}{P(Event \mid X_0, \beta)} = P(Event \mid X_0) \times \exp\left((X_* - X_0)^T \beta\right).$$

Caution should be used in estimating $P(Event \mid X_0)$ from expert judgment directly though. As absolute assessments of probabilities by experts are less calibrated for probabilities near zero or one (see Cooke, 1991 for discussion), it might be preferable to assess the probability of the event for the value of X_0 that makes the probability of the event as near as possible to 0.5. One should also note that it is possible to calculate probabilities above one using this method (but not below zero), thus requiring truncation. However, while the support of the distributions would allow incoherent values, they are extremely unlikely as we are dealing with low probability events in this context.

3. Analysis for Correlated Experts

3.1 A Multivariate Model

Clemen (1986 1987), Winkler (1981) and Mosleh et al. (1988) discuss the need for the representation of correlation between the experts in the analysis of expert judgment data. Winkler (1981) develops an aggregation technique for experts' assessments of a single, continuous quantity θ using the multivariate normal distribution, although here we follow more the form and notation of Clemen and Winkler (1985). If we denote the experts' point estimates of θ as $\mu = (\mu_1, \ldots \mu_p)$ and let $e_i = \mu_i - \theta$ be their judgment errors around the parameter θ, then Winkler's likelihood is formed by assuming that

$$e = \begin{pmatrix} e_1 \\ \vdots \\ e_p \end{pmatrix} \sim MVNormal(\underline{0}, \boldsymbol{\Sigma}),$$

where $MVNormal(\underline{0}, \boldsymbol{\Sigma})$ denotes a multivariate normal distribution with mean vector $\underline{0}$, a vector of p zeros, and covariance matrix $\boldsymbol{\Sigma}$. Winkler specifies the decision maker's prior distribution on θ as diffuse and updates using the multivariate normal likelihood

$L(\theta;\mu_1,\ldots,\mu_p,\boldsymbol{\Sigma})$. Winkler's initial set-up requires the decision maker to specify the covariance matrix $\boldsymbol{\Sigma}$ as a hyperparameter of the analysis. Winkler shows that the posterior distribution of θ can then be re-written as

$$\pi(\theta;\mu,\boldsymbol{\Sigma}) \propto \exp\left(-\left(\theta-\mu^*\right)^2/2\sigma^{*2}\right) \tag{4}$$

where

$$\mu^* = \underline{1}^T\boldsymbol{\Sigma}^{-1}\mu/\underline{1}^T\boldsymbol{\Sigma}^{-1}\underline{1} \tag{5}$$

$$\sigma^{*2} = 1/\underline{1}^T\boldsymbol{\Sigma}^{-1}\underline{1} \tag{6}$$

and $\underline{1}^T = (1,\ldots,1)$ is a vector of p 1's. The mean tem in (5) is in fact a linear combination of the experts' assessments based on their covariance $\boldsymbol{\Sigma}$. Winkler's second set-up allows the decision maker to specify a prior distribution on $\boldsymbol{\Sigma}$, specifically an inverted Wishart distribution.

In our case, the single quantity θ is replaced by the multiple assessments of $\exp\left(X_i^T\beta\right)$ ($i=1,\ldots,N$) that are linked by the common parameter vector β. Note, however, that the term $\exp\left(X_i^T\beta\right)$ is a function of the decision maker's model parameter and is not an observable quantity. Thus our aggregation model is defined at the level of β, the parameters of this model, and is used purely to find a posterior distribution for β given the experts' responses. We are essentially modeling the experts' responses to form a new prediction as in Dawes (1979).

There are multiple assessments made by multiple experts, which we denoted by $y_{i,j} = \ln\left(z_{i,j}\right)$. We may mirror Winkler's development by defining $u_{i,j} = y_{i,j} - X_i^T\beta$ and letting

$$u_i^T = \begin{pmatrix} u_{i,1} \\ \vdots \\ u_{i,p} \end{pmatrix} \sim MVNormal(\underline{0}, \boldsymbol{\Sigma}). \tag{7}$$

Here $\boldsymbol{\Sigma}$ is a parameter of the expert aggregation model in (7), while β is a parameter of the decision maker's assumed log-liner model defined in (1). We may re-write this model in matrix form to obtain

$$\begin{pmatrix} y_{1,1} & \cdots & y_{1,p} \\ \vdots & \ddots & \vdots \\ y_{N,1} & \cdots & y_{N,p} \end{pmatrix} = \begin{pmatrix} x_{1,1} & \cdots & x_{1,q} \\ \vdots & \ddots & \vdots \\ x_{N,1} & \cdots & x_{N,q} \end{pmatrix} \begin{pmatrix} \beta_1 & \cdots & \beta_1 \\ \vdots & \ddots & \vdots \\ \beta_q & \cdots & \beta_q \end{pmatrix} + \begin{pmatrix} u_{1,1} & \cdots & u_{1,p} \\ \vdots & \ddots & \vdots \\ u_{N,1} & \cdots & u_{N,p} \end{pmatrix}$$

or

$$\mathbf{Y} = \mathbf{X}\beta\underline{1}^T + \mathbf{U} \tag{8}$$

This equation is similar to a full multivariate regression model

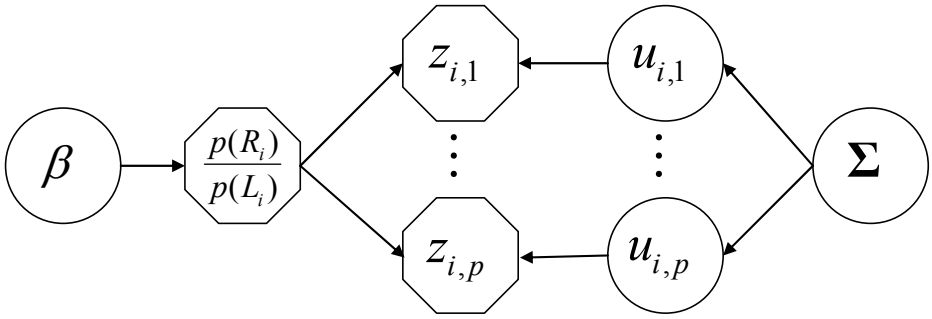
$$\mathbf{Y} = \mathbf{XB} + \mathbf{U}, \tag{9}$$

where $\mathbf{X}$ is a $(N \times q)$-matrix of differences between the q covariates for the N questions, $\mathbf{B}$ is a $(q \times p)$-matrix where each column represents the covariate effect parameters for an expert and $\mathbf{U}$ is a $(N \times p)$-vector of residual errors. The difference between (8) and (9) is that in (8) columns of the regression parameter matrix $\mathbf{B}$ are restricted to be equal as each expert is providing estimates of the same quantity.

The form in (8) suggests that we follow the analysis of a multivariate regression model, such as that developed in Press (1982). Equation (7) implies that the rows of $\mathbf{U}$ are independent vectors distributed according to a multivariate normal with a zero mean vector and covariance matrix $\boldsymbol{\Sigma}$. The rows of $\mathbf{U}$ are assumed to be independent as they are responses to the individual questions, but the columns are dependent as they represent

the responses of the experts to each question. Analyzing the model in (7) will make our analysis different from Winkler's, as the prior distribution on $\boldsymbol{\Sigma}$ will be updated by the judgments of the experts.

The overall model can be best displayed as a Bayesian belief net (Figure 3).



$$\frac{p(R_i)}{p(L_i)} = \mathrm{e}^{(R_i - L_i)^T \beta} \qquad \ln z_{i,j} = \ln \frac{p(R_i)}{p(L_i)} + u_{i,j} \qquad \left(u_{i,1},...,u_{i,p}\right) \sim MVN(\underline{0}, \boldsymbol{\Sigma})$$

Figure 3. The aggregation model for the i-th question.

In Figure 3, we show deterministic relationships with a hexagon. The assumed functional forms are shown as equations below. Note that we do not need to show the $u_{i,j}$, but include them to show the multivariate normal error structure that gives us the multivariate regression format.

3.2 Posterior Analysis

While the following analysis mirrors the Bayesian analysis in Press (1982), the likelihood and posterior distributions for the column restricted form in (8) requires additional development. Adapting Press's likelihood for (9) to our form in (8), we may write

$$p\left(\mathbf{Y} \mid \mathbf{X}, \beta, \boldsymbol{\Sigma}\right) \propto \left|\boldsymbol{\Sigma}\right|^{-\frac{N}{2}} \exp\left\{-\tfrac{1}{2} tr\left(\mathbf{V}\boldsymbol{\Sigma}^{-1}\right)\right\} \quad \exp\left\{-\tfrac{1}{2} tr\left(\left(\hat{\mathbf{B}} - \beta \underline{1}^T\right)^T \mathbf{X}^T \mathbf{X} \left(\hat{\mathbf{B}} - \beta \underline{1}^T\right) \boldsymbol{\Sigma}^{-1}\right)\right\}$$

where $\mathbf{V} = \left(\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}}\right)^T \left(\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}}\right)$ is the usual sufficient statistic for the unrestricted model and $\hat{\mathbf{B}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Y}$ is the least squares estimates of $\mathbf{B}$ from the unrestricted model in (9), providing estimates of each parameter for each expert. Completing the square in β for the second exponential term, we may re-write the likelihood for (8) as

$$p\left(\mathbf{Y} \mid \mathbf{X}, \beta, \mathbf{\Sigma}\right) \propto \exp\left\{-\tfrac{1}{2}\left(\left(\beta - \mu_*^\beta\right)^T \left(\Sigma_*^\beta\right)^{-1} \left(\beta - \mu_*^\beta\right)\right)\right\}. \tag{10}$$

where

$$\mu_*^\beta = \frac{\hat{\mathbf{B}}\mathbf{\Sigma}^{-1}\underline{1}}{\underline{1}^T\mathbf{\Sigma}^{-1}\underline{1}} \tag{11}$$

and

$$\mathbf{\Sigma}_*^\beta = \frac{\left(X^T X\right)^{-1}}{\underline{1}^T\mathbf{\Sigma}^{-1}\underline{1}} \tag{12}$$

Note the similarity of (12) and (6). The similarity of (11) and (5) is more obvious were (8) defined on the transpose, whereas we follow the convention in Press. The mean term in (11) is a linear combination of the least-squares estimates of the parameters for each expert in $\hat{\mathbf{B}}$.

A natural conjugate analysis is made possible by the following distributional assumptions,

$$\left(\mathbf{\Sigma}\right) \sim Inv - Wishart\left(\mathbf{G}, m\right), \tag{13}$$

which defines an inverse Wishart distribution of dimension p with parameter matrix $\mathbf{G}$ and *m* degrees of freedom, and

$$\left(\beta \mid \mathbf{\Sigma}\right) \sim MVNormal\left(\phi, \frac{\mathbf{A}}{\underline{1}^T\mathbf{\Sigma}^{-1}\underline{1}}\right). \tag{14}$$

ϕ, $\mathbf{A}$, $\mathbf{G}$ and m are prior hyperparameters determined by the decision maker. Given the experts' responses to the questionnaires, the posterior distributions are

$$\left(\boldsymbol{\Sigma} \mid \mathbf{Y}, \mathbf{X}\right) \sim Inv - Wishart\left(\mathbf{G} + \mathbf{V}, m + N\right) \tag{15}$$

and

$$\left(\beta \mid \mathbf{Y}, \mathbf{X}, \boldsymbol{\Sigma}\right) \sim MVNormal\left(\left(\mathbf{A}^{-1} + \mathbf{X}^T\mathbf{X}\right)^{-1}\left(\mathbf{X}^T\mathbf{X}\frac{\hat{\mathbf{B}}\boldsymbol{\Sigma}^{-1}\underline{1}}{\underline{1}^T\boldsymbol{\Sigma}^{-1}\underline{1}} + \mathbf{A}^{-1}\phi\right), \frac{\left(\mathbf{A}^{-1} + \mathbf{X}^T\mathbf{X}\right)^{-1}}{\underline{1}^T\boldsymbol{\Sigma}^{-1}\underline{1}}\right). \tag{16}$$

Thus the analysis is conjugate, making calculation, and therefore application, easier.

3.3 Prediction

Once the Bayesian update has been performed, the next step is to develop the predictive distribution. In this case, we wish to predict how much more likely an accident is in one scenario compared to the other, or the ratio of the probabilities of an accident in the two scenarios. Actual probabilities in a given scenario can then be assessed by comparison to the reference scenario as discussed in Section 2.2.

One would imagine that the development of a predictive distribution would mirror the development for multivariate regression, but with posterior distributions drawn from our parameter restricted model form. However, we are not attempting to predict what the experts would assess for the two situations. We have used their assessments to update the decision maker's prior on β in a Bayesian expert aggregation method. We may then use the decision maker's posterior distribution for β and the first level of the model defined in (2) to estimate the ratios of probabilities. The natural logarithm of the ratio to be predicted for two scenarios with difference vector x^* conditioned on $\boldsymbol{\Sigma}$ will be a multivariate normal distribution defined by

$$\left(x^{*T}\beta \mid \mathbf{Y},\mathbf{X},\boldsymbol{\Sigma}\right) \sim MVNormal\left(x^{*T}\left(\mathbf{A}^{-1}+\mathbf{X}^{T}\mathbf{X}\right)^{-1}\left(\mathbf{X}^{T}\mathbf{X}\frac{\hat{\mathbf{B}}\boldsymbol{\Sigma}^{-1}\underline{1}}{\underline{1}^{T}\boldsymbol{\Sigma}^{-1}\underline{1}}+\mathbf{A}^{-1}\phi\right), x^{*T}\frac{\left(\mathbf{A}^{-1}+\mathbf{X}^{T}\mathbf{X}\right)^{-1}}{\underline{1}^{T}\boldsymbol{\Sigma}^{-1}\underline{1}}x^{*}\right) \tag{17}$$

We may then integrate out $\boldsymbol{\Sigma}$ using (15).

4. Example Results

4.1 Elicitation for the WSF Risk Assessment

Expert judgment was used in the WSF Risk Assessment to estimate the effect of risk factors on the probability of a collision given the occurrence of some triggering incident. The risk factors are listed in Table 1 and include the ferry class and route, the type, proximity and angle of interaction of the closest two vessels, the visibility conditions and wind speed and direction. For a discussion of the derivation of the scales used for these risk factors see Szwed et al. (2004). Six potential interactions are also included in Table 1. The supposition in the inclusion of these interactions is that the types of the ferries and the other vessel, their manner of interaction and the visibility conditions are potentially dangerous in certain combinations. Other interactions were not considered as large and so were not included to minimize the total number of questions that each expert had to respond to.

Experts may be classified in three categories (DeWispelare et al. 1995):

- **generalists** who have a thorough understanding of the project and play a role in defining the issues addressed and communicating with the experts;
- **substantive experts** who have the deep knowledge and experience of a system that allow them to provide information about the functioning of that system; and

- **normative experts** who have the analysis background to quantify the judgments of the substantive experts and combine their judgments.

Table 1. The risk factors included in the expert judgment questionnaires.

Description	Notation	Values
Ferry route and class	FR_FC	26
Type of 1st interacting vessel	TT_1	13
Scenario of 1st interacting vessel	TS_1	4
Proximity of 1st interacting vessel	TP_1	Binary
Type of 2nd interacting vessel	TT_2	5
Scenario of 2nd interacting vessel	TS_2	4
Proximity of 2nd interacting vessel	TP_2	Binary
Visibility	VIS	Binary
Wind direction	WD	Binary
Wind speed	WS	Continuous
Interaction between FR_FC and TT_1	FR_FC * TT_1	26×13
Interaction between FR_FC and TS_1	FR_FC * TS_1	26×4
Interaction between FR_FC and VIS	FR_FC * VIS	26×2
Interaction between TT_1 and TS_1	TT_1 * TS_1	13×4
Interaction between TT_1 and VIS	TT_1 * VIS	13×2
Interaction between TS_1 and VIS	TS_1 * VIS	4×2

Certain members of the risk assessment team were normative experts, with knowledge of decision theory, probabilistic reasoning and expert elicitation techniques.

Other members were generalists with both maritime experience, knowledge of maritime risk issues and systems engineering techniques. The substantive experts used in the study were the ferry captains that worked relief, filling in for captains on vacation or sick leave across all ferry routes. This ensured that the experts had a thorough knowledge of the entire system, not just a specific route. Each of the experts used had over 10 years of experience with the WSF.

The elicitation team first provided to the substantive experts some background on the project followed by an explanation of the questionnaires and their purpose. Example questions were presented similar to Figure 1, but in the context of driving a car on the highway. This context was also explained in terms of several risk factors. The highway transportation mode was chosen over maritime examples to avoid biasing the experts before beginning the questionnaires and because everyone was familiar with the situations defined. The experts were then given an example question to consider in the driving example and discussion encouraged between the experts to ensure the idea was understood. It was important to remind the experts to look at all the risk factors in the question, rather than just the one that changed between the two situations as there can be interactions between the risk factors.

Each questionnaire consisted of sixty comparisons of the type shown in Figure 2. The questionnaires were designed to collect the maximum amount of information from the sixty questions and to ensure that sufficient information was elicited to ensure the estimation of the main ten risk factors and six pre-defined interactions between risk factors. The questions were asked in random order. The randomization of the questions

meant that deliberate attempts to bias the results were difficult. Tests on the responses were performed to ensure that the experts' responses were not affected by fatigue.

4.2 Prior Distributions

The first step in analyzing the expert judgment data is the specification of the prior hyperparameters. Clemen (1986) discusses the concept of aggregation of the decision maker's beliefs with those of the experts. In our applications the decision makers have claimed ignorance of the effect of the factors in Table 1 on the probability of a collision and wished for the experts' beliefs to dominate the predictions. In the Bayesian sense, this means specifying suitably vague priors. These assumptions are conservative and in other applications a decision maker might have more specific prior information on which to base the prior specifications.

We assumed that ϕ, the vector of the prior means on β, is a vector of zeros, which indicates that a priori all covariates have on average no effect on the probability of an accident. The prior matrix $\mathbf{A}$ is assumed to be an identity matrix to indicate no prior covariance between the parameters in β. The prior matrix $\mathbf{G}$ is assumed to be an identity matrix, indicating no prior knowledge of correlations between the experts, while m is assumed to be 0.380341 calculated by Szwed et al. (2004) to represent a priori that all expert respond to the N questions completely at random.

These prior assumptions are diffuse. Figure 4 shows the prior predictive distribution of the natural logarithm of $\frac{P(Collision \mid Nav.Fail, L)}{P(Collision \mid Nav.Fail, R)}$, where L is the scenario on the left and R is the scenario on the right of Figure 1. Note that the distribution in Figure 4 has median of zero, which is equivalent to one on the non-

logarithmic scale implying that collisions are equally likely in each situation. There is wide variability with a 90% credibility interval for the ratio of probabilities between $1.88x10^{-35}$ and $5.32x10^{34}$.

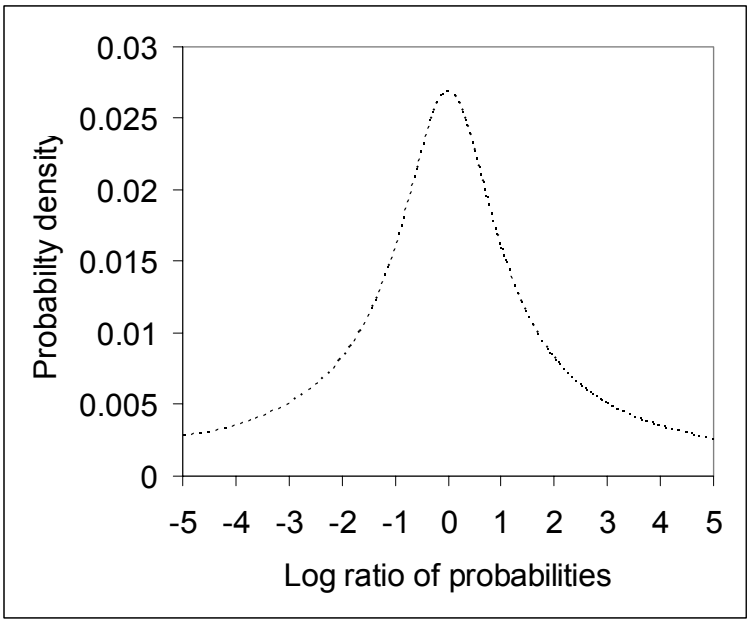


Figure 4. The prior distribution of the logarithm of the ratio of probabilities for the scenarios pictured in Figure 1.

4.2 Posterior Distributions

After updating with the experts' responses using the formulae developed in Section 3.2, the marginal posterior distributions of the β parameters are as shown in Figure 5, represented by a circle for their mean and whiskers showing their prior 90% credibility interval. To demonstrate the advantage of our model including dependence, we compare the results to the independent experts model in (9) developed in Szwed et al. (2004). Figure 6 shows the marginal posterior distributions of the β parameters obtained using the independent experts model from Szwed et al. (2004), on the same scale as Figure 5. Note that while the mean values are similar, some slight differences can be observed

amongst them in Figures 5 and 6. More noticeable, however, is that the posterior variance of every parameter is less in the dependent experts model (Figure 5) than that observed in the independent one by Szwed et al. (Figure 6). This is counter to most expert dependence examples, where positive correlations actually increase the uncertainty for a given number of experts. We will discuss this further in Section 4.4 and 4.5.

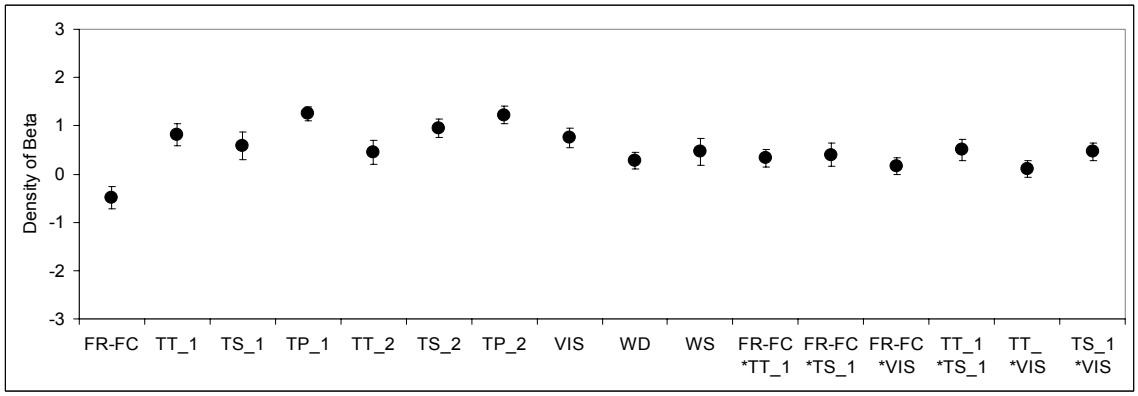


Figure 5. The marginal posterior distribution of β assuming dependent experts.

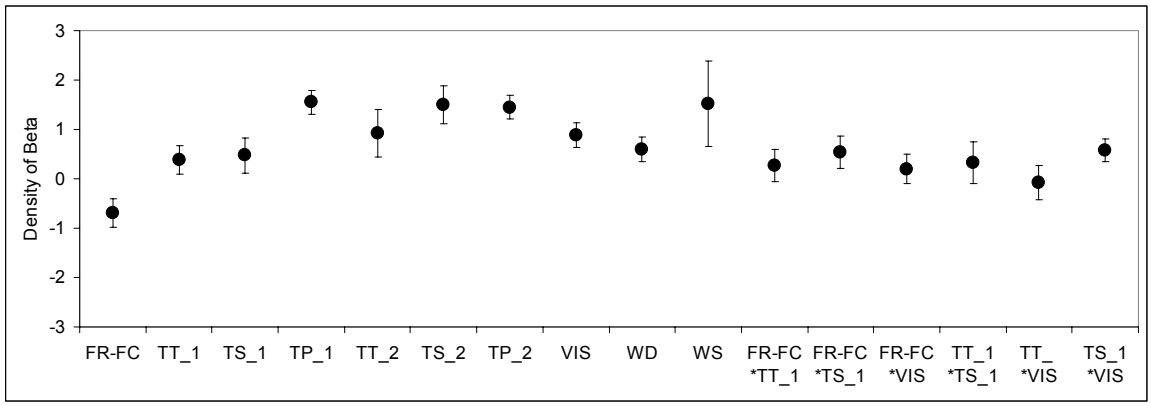


Figure 6. The marginal posterior distribution of β assuming independent experts.

Of particular interest in this analysis, is the posterior distribution of the covariance matrix $\boldsymbol{\Sigma}$ representing the updated dependencies between the experts. Table 2 shows the posterior expected value of the correlations corresponding to $\boldsymbol{\Sigma}$. Careful examination reveals some apparent groupings of the experts. Specifically experts 1, 3 and 7 appear to

be positively correlated, as do experts 2, 4 and 6. However, Table 2 does not show the uncertainty in these estimates.

Table 2. The expected correlations between the experts.

	Expert 1	Expert 2	Expert 3	Expert 4	Expert 5	Expert 6	Expert 7	Expert 8
Expert 1	1	-0.23	0.68	0.33	-0.54	-0.33	0.33	-0.26
Expert 2	-0.23	1	-0.11	0.29	-0.25	0.56	-0.08	0.14
Expert 3	0.68	-0.11	1	0.36	-0.30	-0.06	0.52	-0.31
Expert 4	0.33	0.29	0.36	1	-0.59	0.35	0.17	-0.18
Expert 5	-0.54	-0.25	-0.30	-0.59	1	0.01	0.04	0.11
Expert 6	-0.33	0.56	-0.06	0.35	0.01	1	0.08	0.24
Expert 7	0.33	-0.08	0.52	0.17	0.04	0.08	1	-0.23
Expert 8	-0.26	0.14	-0.31	-0.18	0.11	0.24	-0.23	1

Figure 7 shows the posterior distribution of the correlation matrix with the correlation between the i-th and j-th experts indicated by the notation i,j on the top left of each histogram. Only lower triangular elements are shown to reduce clutter in the figure. A vertical line is drawn at 0, indicating no dependence between the two experts. Thus a histogram showing samples to the right of the line indicates a posterior probability that the two experts have positive dependence or overlapping information, while samples to the left indicates negative dependence or different information.

The experts are re-ordered in Figure 7 to show groupings that appeared in the expected correlations in Table 2. Experts 1, 3 and 7 do have a tendency to agree in their responses, as do Experts 2, 4 and 6. These groups are not completely disparate, however. Experts 2 and 6 tend to disagree with experts 1, 3 and 7, but expert 4 tends to agree with them. Expert 5 tends to disagree with experts 1, 2, 3, 4 and 6, but does not conclusively

agree or disagree with expert 7. Expert 8 disagrees with the group of experts 1, 3 and 7, but does not conclusively agree or disagree with the group 2, 4 and 6 or with 5.

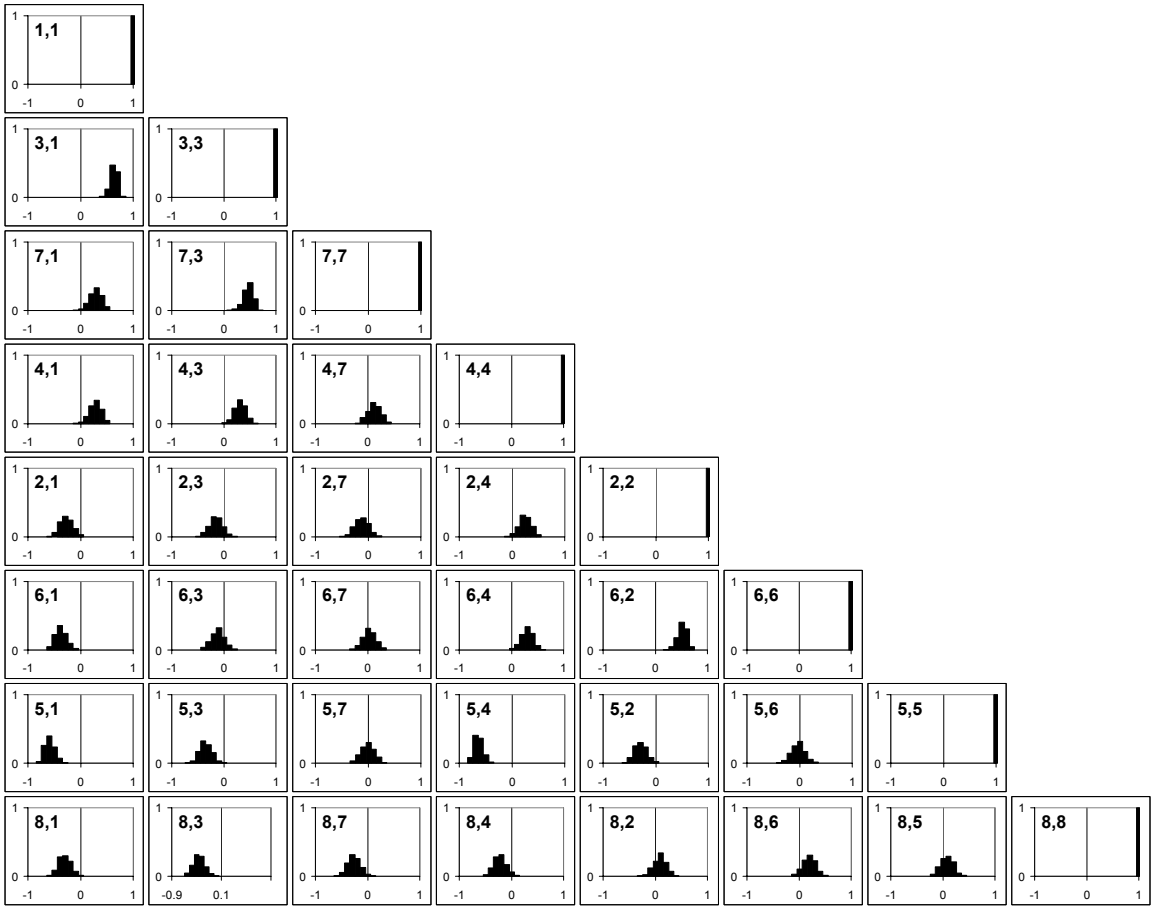


Figure 7. The posterior distribution of the expert correlation matrix.

4.3 Posterior Predictions

We have seen the difference these correlations make in the variance of the β parameters and to the precision of the residuals, but of more interest is their effect on the predictive distribution. Using the result in (14), Figure 8 shows the posterior predictive distribution of natural logarithm of $\frac{P(Collision \mid Nav.Fail, L)}{P(Collision \mid Nav.Fail, R)}$, where L is the scenario on the left and R is the scenario on the right of Figure 1. The prior distribution is also shown as a dotted

line to show the effect of updating. The posterior 90% credibility interval on the ratio of probabilities is 4.38 to 5.84, with a half-width of 0.73.

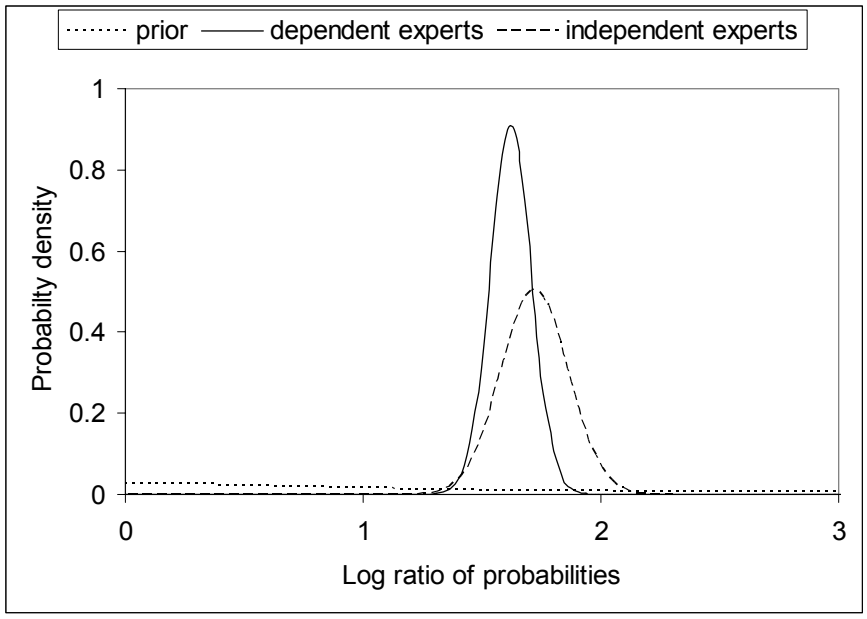


Figure 8. The prior and posterior density of the logarithm of the ratio of probabilities for the scenarios pictured in Figure 1.

The reader should note that the question in Figures 1 and 2 is different from that illustrated in Szwed et al. (2004) so we give the comparison using their method on a common question here.

Figure 8 also compares the posterior distribution of the ratio of probabilities for the same prediction obtained using the independent experts model of Szwed et al. (2004). The independent expert model gives a higher posterior expected value of the ratio of probabilities, 5.59 as opposed to 5.11 with the dependent experts model, and a larger variance. The posterior 90% credibility interval is 4.43 to 7.04, with a half-width of 1.3 compared to 0.73 for the dependent expert model.

Clemen (1986) suggests that experts will have overlapping information implying positive correlations in their assessments. Clemen and Winkler (1985) show that such positive correlations will actually reduce the precision of forecasts when compared to an equivalent number of independent experts. However, in our case we have both positive and negative correlations between the experts (Table 2 and Figure 7) and we have seen increases in the level of precision in both posterior parameter distributions and posterior predictive distributions (Figure 8).

5. Conclusions

We have developed an analysis of an extended form of pairwise comparisons introduced in Merrick et al. (2000) from a Bayesian analysis that assumes that experts' responses are independent (Szwed et al. 2004) to one that allows for correlations between experts. The analysis was set up using the theory of normal errors approach of Winkler (1981) to assess the parameters of a log-linear relationship between the probabilities and the defining factors. However, as the aggregation is not performed on the experts' direct assessments, but on estimates of parameters of the decision maker's model based on the experts' assessments. Thus our approach is akin to that of Dawes (1979) where the experts' assessments are modeled, not directly aggregated. The model itself takes the form of a special case of Bayesian multivariate regression.

The method was applied to expert judgment data elicited during the WSF Risk Assessment. The empirical results show that there were correlations between the experts in this data and that allowing for these correlations decreases the posterior variance in the predictions made using the model compared to those obtained in Szwed et al. (2004). This reduction in uncertainty is counter to the commonly assumed increases in posterior

variance due to positive correlations representing overlapping information and could be critical in determining whether to apply proposed risk interventions when such risk interventions are evaluated using the output of this expert judgment methodology. This result is made possible as our set-up allows updating of prior knowledge about dependences between the experts.

For our example prediction, an analysis assuming independence between the experts would say that a collision with a navy vessel is anywhere from 4.43 to 7.04 times as likely as a collision with a product tanker where each is in the same situation pictured in Figure 1. Our analysis that learns about the dependencies between the experts from their responses, would predict that the navy vessel is anywhere from 4.38 to 5.84 times as risky as the product tanker. This is a reduction of about 45% in the width of the prediction interval. While this is only demonstrated for this specific example, such a reduction in uncertainty could have a major impact on a decision or risk analysis. An alternative that would otherwise not clearly dominate another could be shown to be clearly superior. In other cases where positive dependence dominates the prediction, the admission of increased uncertainty over the independent experts analysis would also be helpful in making good decisions based on such pairwise expert judgments.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant Nos. SES 0213627 and SES 0213700. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

We would like to thank the two anonymous reviewers for extremely helpful feedback that improved both the article and the development itself. We would also like to thank Bob Clemen (the editor) and Casey Lichtendahl for extensive discussions that solved initial flaws in the work.

References

Bradley, R., M. Terry. 1952. Rank analysis of incomplete block designs. *Biometrica* **39** 324-345.

Clemen R. T. 1986. Calibration and aggregation of probabilities. *Management Science* **32**(3) 312-314.

Clemen R. T. 1987. Combining overlapping information. *Management Science* **33**(3) 373-380.

Clemen, R. T., Winkler, R. L. 1985. Limits for the precision and value of information from dependent sources. *Operations Research* **33** 427–442.

Clemen, R. T., Winkler, R. L. 1999. Combining probability distributions from experts in risk analysis. *Risk Analysis* **19**(2) 187–203.

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

Cooke, R. M. 1991. *Experts in Uncertainty: Expert Opinion and Subjective Probability in Science.* Oxford University Press, Oxford UK.

Cox, D. R. 1972. Regression models and life tables. Journal *of the Royal Statistical Society Ser. B* **34** 187-220.

Dawes, R. M. 1979. The robust beauty of improper linear models in decision making. *American Psychologist* **34** 571-582.

DeWispelare, A., L. Herren, R. Clemen. 1995. The use of probability elicitation in the high-level nuclear waste recognition program. *International Journal of Forecasting* **11**(1) 5-24.

French, S. 1980. Updating belief in the light of someone else's opinion. *Journal of the Royal Statistical Society Series A* **143** 43-48.

French, S. 1981. Consensus of opinion. *European Journal of Operations Research* **7** 332-340.

Jouini, M. N. R. T. Clemen. 1996. Copula models for aggregating expert opinion. *Operations Research* **44**(3) 444-457.

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

Lindley, D. 1983. Reconciliation of probability distributions. *Operations Research* **31** 866-880.

Lindley, D. 1985. Reconciliation of discrete probability distributions. In *Bayesian Statistics 2*, J. Bernado et al. (Eds.), North Holland, Amsterdam 375-390.

Merrick, J. R. W., J. R. van Dorp, J. Harrald, T. Mazzuchi, J. Spahn, M. Grabowski. 2000. A systems approach to managing oil transportation risk in Prince William Sound. Systems Engineering **3**(3) 128-142.

Merrick, J. R. W., J. R. van Dorp, T. Mazzuchi, J. Harrald, J. Spahn, M. Grabowski. 2002. The Prince William Sound Risk Assessment. *Interfaces* **32**(6) 25-40.

Merrick, J. R. W., J. R. van Dorp, V. Dinesh. 2004a. Assessing uncertainty in simulation based maritime risk assessment. Accepted by *Risk Analysis*.

Merrick, J. R. W., J. R. van Dorp. 2004b. Speaking the truth in maritime risk assessment. In preparation.

Morris, P. A. 1974. Decision analysis expert use. *Management Science* **20** 1233–1241.

Morris, P. A. 1977. Combining expert judgments: A Bayesian approach. *Management Science* **23** 679–693.

Morris, P. A. 1983. An axiomatic approach to expert resolution. *Management Science* **29** 24–32.

Moslesh, A., V. Bier, G. Apostolakis. 1988. Critique of the current practice for the use of Expert opinions in probabilistic risk assessment. *Reliability Engineering and System Safety* **20** 63-85.

Press, S. J. 1982. *Applied Multivariate Analysis Using Bayesian and Frequentist Methods and Inference*. 2nd Edition. Robert E. Krieger Publishing Company, Malabar, Florida.

Paté-Cornell, M. E. 1996. Uncertainties in risk analysis: six levels of treatment. *Reliability Engineering and System Safety* **54**(2-3) 95-111.

Pulkkinen, U. 1993. Methods for combination of expert judgments. *Reliability Engineering and System Safety* 40(2) 111-118.

Pulkkinen, U. 1994a. Bayesian analysis of consistent paired comparisons. *Reliability Engineering and System Safety* 43(1) 1-16.

Pulkkinen, U. 1994b. Gaussian paired comparison models. *Reliability Engineering and System Safety* 44(2) 207-217.

Saaty, T. 1977. A scaling method for priorities in hierarchical structures. *Journal of Mathematical Psychology* **15**(3) 234-281.

Szwed, P., J. Rene van Dorp, J. R. W. Merrick, T. A. Mazzuchi, A. Singh. 2004. A Bayesian paired comparison approach for relative accident probability assessment with covariate information. Accepted by *European Journal of Operations Research*.

van Dorp, J. R., J. R. W. Merrick, J. Harrald, T. Mazzuchi, M. Grabowski. 2001. A Risk Management Procedure for the Washington State Ferries. *Risk Analysis* **21**(1) 127-142.

Winkler, R. L. 1981. Combining probability distributions from dependent information sources. *Management Science* **27** 479-488.