Debiasing Expert Overconfidence: A Bayesian Calibration Model

Robert T. Clemen* and Kenneth C. Lichtendahl, Jr.
Fuqua School of Business
Duke University
Durham, NC 27708

clemen@mail.duke.edu
casey.lichtendahl@duke.edu

Presented at PSAM6, June 27, 2002

Abstract

In a decision and risk analysis, experts may provide subjective probability distributions that encode their beliefs about future uncertain events. For continuous variables, experts often provide these judgments in the form of quantiles of the distribution (e.g., 5th, 50th, and 95th percentiles). Psychologists have shown, though, that such subjective distributions tend to be too narrow, representing overconfidence on the part of the expert. We propose an approach for modeling and debiasing expert overconfidence. Based on past performance data (previous assessments and realizations for a number of uncertain variables), and using Bayesian methods to update prior distributions on the model parameters, we show how our model can be used to debias expert probabilities. We develop and demonstrate both a single-expert model and a multiple-expert hierarchical model.

KEYWORDS Calibration, expert judgment, subjective probability, debiasing, overconfidence

1 Introduction

Risk analyses often rely on experts to provide subjective probability assessments for unknown variables. A typical assessment question for a continuous uncertain variable might be, "Consider net operating profit for Delta Airlines next year. Please give a number x such that you believe there is a 5% chance that the

*Supported by the National Science Foundation under Grant SES-00-84383.

actual net operating profit for Delta Airlines next year will be above x." The answer to this question is the 95th percentile of the expert's subjective probability distribution for Delta's net operating profit. Similar questions would be asked for other quantiles. Typical practice is to obtain several such quantiles. These quantiles may then be used in a probabilistic model, typically by fitting a continuous distribution to them. The analyst might use either the continuous distribution itself or a discrete approximation.

It is well known that individuals are subject to judgment biases when assessing subjective probabilities (e.g., Kahneman et al. 1982). The use of specific protocols (Merkhofer 1987, Morgan & Henrion 1990) for expert assessment can help to reduce such biases but does not eliminate them. In this article we focus on expert overconfidence. Reviews that cover psychological models of confidence and experimental results include Budescu et al. (1997), Griffin & Varey (1996), Keren (1991), Klayman et al. (1999), Lichtenstein et al. (1982), and McClelland & Bolger (1994). Although such overconfidence has been documented for both binary and continuous variables, in this paper we consider only continuous variables. An expert's probability distribution displays overconfidence when an assessed interval (e.g., the interval between an expert's 5th and 95th percentiles) is too narrow. Detecting this phenomenon in practice requires the analyst to compare a number of such assessed intervals with the corresponding realizations. If each set of assessments/realization is exchangeable (de Finetti 1937) with respect to other assessments/realizations, then for a calibrated judge we would expect the proportion of realizations that fall within the assessed intervals to correspond to the specified probability. For example, if an expert has made many 5th-95th percentile judgments, then we would expect that 90% of realizations would fall within the corresponding intervals. Such an expert is said to be "calibrated." For many probability assessors, the proportion is on the order of 40%-60% of realizations (e.g., see Lichtenstein et al. 1982), reflecting overly narrow assessed intervals.

Attempts to debias expert judgments have generally focused on finding ways to improve the elicitation process itself so as to improve the calibration of the assessed probabilities (Fischhoff 1982). For example, the protocols mentioned above include steps in which experts learn the principles of subjective probability judgment and the associated biases.

In this paper, we take the approach of calibrating the expert's judgments after the elicitation is completed. This approach is consistent with ideas from Cox (1958), Morris (1974), and Harrison (1977). Shlyakhter (1994) (see also Shlyakhter et al. 1994) develops a model of overconfidence and uses past data to estimate an "inflation factor" for assessed distributions. Shlyakhter treats all experts as exchangeable and hence applies his inflation equally to all experts. In contrast, our Bayesian model is capable of calibrating individual experts.

Section 2 below describes a basic, single-expert model, which we demonstrate with data from Cooke (1991). Although useful, the single-expert model essentially treats all experts separately; Expert i's overconfidence characteristics (parameters) are completely unrelated to Expert j's. Section 3 describes a Bayesian hierarchical model in which each expert's parameters are randomly

drawn from a distribution characterized by hyperparameters with their own prior distributions. In this model, the experts are conditionally independent given their parameters. These parameters are uncertain but are related through the hyperparameters; this allows data from one expert to affect inferences about other experts. Section 4 concludes with a discussion of the model and further research.

2 Debiasing a Single Expert

2.1 A Judgment Model

Suppose an expert assesses three quantiles for uncertain quantity X_i. Denote the expert's quantiles by L_i, R_i, and U_i (e.g., 5th, 50th, and 95th percentiles). We want to transform L_i, R_i, and U_i into their unbiased couterparts L_i^*, R_i^*, and U_i^*. We assume that the assessed R_i is biased by parameter β, a location multiplier, so that

$$R_i^* = \beta R_i \,.$$

Our primary concern is with the potential bias in L_i and U_i. We model the bias in L_i and U_i in terms of their distances from R_i^*. Suppose that the distance $R_i^* - L_i$ should be multiplied by parameter α_L and similarly $U_i - R_i^*$ by parameter α_U in order to achieve unbiasedness. Thus, α_L and α_U can be thought of as scale multipliers. The expert is overconfident when α_L or α_U is greater than 1. Thus, we can calculate L_i^* and U_i^* according to the following equations:

$$\begin{aligned} L_i^* &= R_i^* - \alpha_L(R_i - L_i) = (\beta - \alpha_L)R_i + \alpha_L L_i \\ U_i^* &= R_i^* + \alpha_U(U_i - R_i) = (\beta - \alpha_U)R_i + \alpha_U U_i \end{aligned}$$

For random variable X_i (with corresponding realization x_i), an expert's unbiased distribution $G(x_i|L_i^*, R_i^*, U_i^*)$ is modeled as uniform between L_i^* and R_i^* and between R_i^* and U_i^* and with exponential tails below L_i^* and above U_i^*. The exponential tails are fit so that the density $g(x_i)$ is continuous at L_i^* and U_i^*. Thus, $g(x_i)$ is given by

$$g(x_i|L_i^*, R_i^*, U_i^*) = \begin{cases} G(L_i^*)\lambda_L e^{-\lambda_L(L_i - x_i)} & if \quad x_i < L_i^* \\ (G(R_i^*) - G(L_i^*))/(R_i^* - L_i^*) & if \quad L_i^* < x_i < R_i^* \\ (G(U_i^*) - G(R_i^*))/(U_i^* - R_i^*) & if \quad R_i^* < x_i < U_i^* \\ (1 - G(U_i^*))\lambda_U e^{-\lambda_U(x_i - U_i)} & if \quad x_i > U_i^* \,. \end{cases}$$

Parameters λ_L and λ_U are

$$\lambda_L = \left(\frac{G(R_i^*) - G(L_i^*)}{R_i^* - L_i^*}\right)\frac{1}{G(L_i^*)}$$

and

$$\lambda_U = \left(\frac{G(U_i^*) - G(R_i^*)}{U_i^* - R_i^*} \right) \frac{1}{1 - G(U_i^*)} \ .$$

Although this model maintains fidelity with the expert's assessments, it is admittedly ad hoc; there are many possible alternatives such as fitting a member of a parameterized family to the expert's assessment or modeling the tails in other ways. We do not expect our results to be highly sensitive to these modeling assumptions

Parameters α_L, α_U, and β are assumed to be constant for assessments of different random variables and hence are characteristics of the expert. However, these parameters are unknown; for our analysis, we assume diffuse priors for them. If the expert has made judgments on n random variables Y_i in the past for which we have observed realizations $y_i, \ldots, y_n$, and if these assessment/realization vectors are judged to be exchangeable, then we can use the data to find posterior distributions for the three parameters. Let L denote the vector $(L_1, \ldots, L_n)$, similarly for R and U, and for L^*, R^*, and U^* as well. For the model described above, the likelihood of observing realizations $y_i, \ldots, y_n$ is

$$f(y_1, \ldots, y_n | L^*, R^*, U^*) = \prod_{i=1}^{n} g(y_i | L_i^*, R_i^*, U_i^*),$$

or equivalently

$$f(y_1, \ldots, y_n | \alpha_L, \alpha_U, \beta, L, R, U) = \prod_{i=1}^{n} g(y_i | (\beta - \alpha_L)R_i + \alpha_L L_i,\ \beta R_i,\ (\beta - \alpha_U)R_i + \alpha_U U_i).$$

Using this model, we can find posterior distributions for α_L, α_U, and β using Markov Chain, Monte Carlo (MCMC) methods. In the following we use a Metropolis-Hastings algorithm. Details of the algorithm and implementation are available from the authors.

2.2 Example

The data for the demonstration come from a study on risks to manned spaceflight from collisions with space debris (Cooke 1991). In the original study, seven experts provided 5th, 50th, and 95th percentiles for a variety of variables related to space-debris risk. Among other assessments, the experts assessed their probability distributions for the number of radar-tracked objects injected into orbit for each of 26 years beginning in the early 1960s. Cooke and colleagues used the assessments and realizations for these 26 variables to evaluate and combine the experts' probability judgments. We use the same data to demonstrate our debiasing model.

We demonstrate our method using Expert 3's assessments for the 26 variables, which are reproduced in Table 1 along with realizations y_i. Note that

i	L_i	R_i	U_i	y_i	i	L_i	R_i	U_i	y_i
1	102.5	242.0	335.0	292.0	14	337.5	415.0	485.0	20.0
2	117.5	244.0	344.0	24.0	15	357.5	435.0	509.0	585.0
3	132.5	245.0	346.0	150.0	16	375.0	458.0	534.0	609.0
4	147.5	250.0	347.5	97.0	17	392.5	477.5	560.0	552.0
5	162.5	252.5	347.5	823.0	18	410.5	500.0	586.0	178.0
6	180.0	257.5	347.5	223.0	19	430.0	520.0	617.5	87.0
7	197.5	280.0	347.5	27.0	20	447.5	540.0	649.0	88.0
8	222.5	300.0	347.5	287.0	21	462.5	558.0	680.0	578.0
9	238.0	318.0	352.5	356.0	22	477.5	578.0	710.0	191.0
10	260.0	337.0	380.0	508.0	23	501.0	597.5	741.0	84.0
11	279.0	357.5	408.0	187.0	24	522.0	617.5	770.0	33.0
12	298.0	375.0	437.5	12.0	25	540.0	638.0	800.0	546.0
13	318.0	396.0	460.0	556.0	26	556.0	680.0	832.0	601.0

Table 1: Example expert assessments. L_i, R_i, and U_i correspond to the 5th, 50th, and 95th percentiles of the expert's distribution. These 26 assessments display substantial overconfidence; 18 of the 26 actual values (y_i) fall outside the lower (L_i) and upper (R_i) quantiles.

Expert 3 experienced 18 "surprises" (y_i below the 5th or above the 95th percentiles) for these 26 assessments. If the expert were well calibrated, we would expect only 2 or 3 surprises. Hence we believe that Expert 3 is overconfident in making probability assessments and that a decision maker may benefit from debiasing those assessments.

We ran our MCMC model for 100,000 iterations after a burn-in period of 25,000 iterations. The candidate acceptance rate for the Metropolis-Hastings algorithm was 23.8%. Complete details about the implementation and results from the run are available from the authors.

Figures 1 and 2 display posterior densities for Expert 3's α_L, α_U, and β. These densities have medians 2.07, 7.73, and 0.41, respectively. From these we can interpret Expert 3's characteristics. First, with median β of 0.41, Expert 3 appears to assess medians about 2.5 times greater than should be. Furthermore, after correcting the median to get R_i^*, Expert 3 appears to be extremely overconfident in assessing both upper and lower tails. Median $\alpha_L = 2.07$ implies that the assessed distances between L_i and R_i^* tend to be about half of what they should be. Likewise, median $\alpha_U = 7.73$ means that the assessed differences between U_i and R_i^* tend to be about one-eighth of what they should be.

The bimodality in the densities for α_L and β may seem surprising at first glance but is readily explained. The uncertain variables are all numbers that must be nonnegative, and parameters α_L and β must covary in a way that is consistent with these constraints. In particular, if β is small, the location of the calibrated distribution must shift toward zero, in which case α_L must also be small. The presence of several realizations near zero in the data set mean

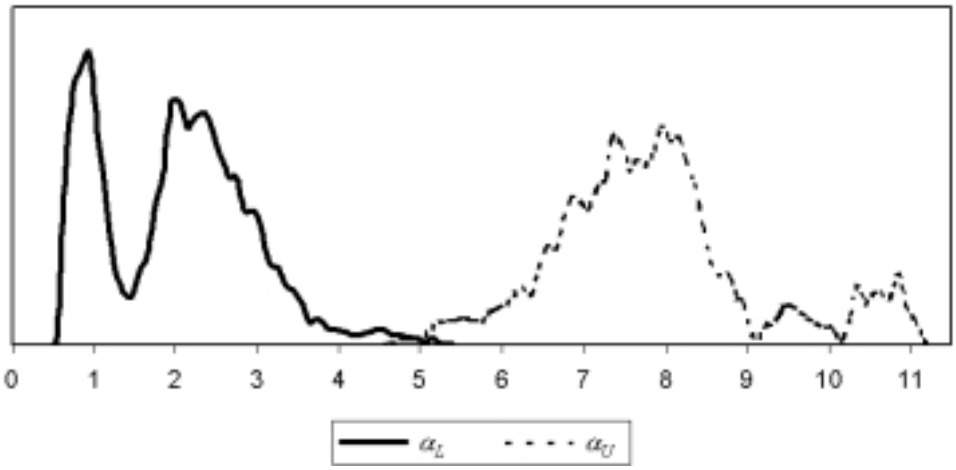

Figure 1: Posterior distributions for Expert 3's α_L and α_U.

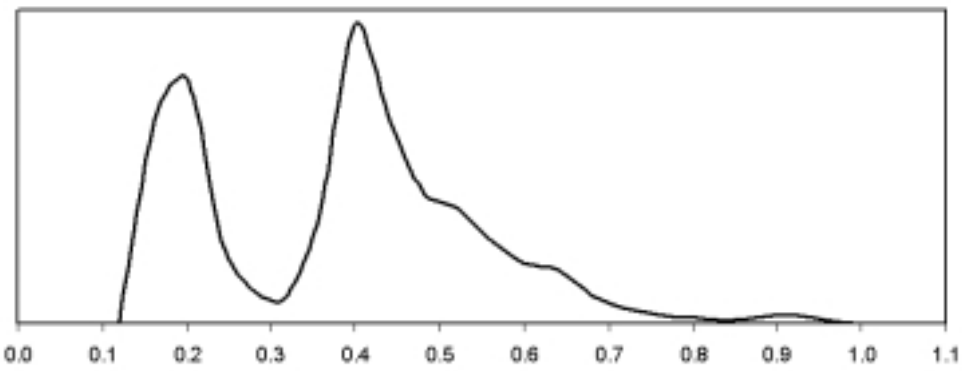

Figure 2: Posterior distribution for Expert 3's β.

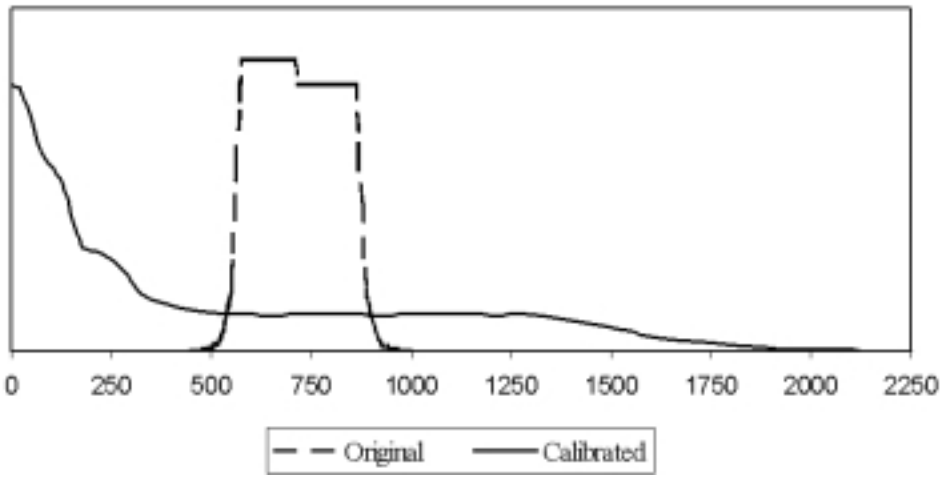

Figure 3: Expert 3's original and calibrated distributions for variable 27.

that the model must put substantial mass on very small values for the two parameters; hence the left mode in each graph. The remaining larger realizations are likewise more consistent with larger parameter values, and hence the right mode. The small modes in the right tail of the density for α_U can be explained with the same reasoning.

We can also use the posterior distributions on the parameters to calibrate a new assessment by Expert 3. For example, consider variable 27, the number of bits of radar-tracked space items placed into orbit the year after the study was done. This assessment was not included in the original set of 26 data points. For this variable, Expert 3 assessed 5th, 50th and 95th percentiles to be 572, 712, and 865, respectively. Figure 3 shows the calibrated density, for which the 5th, 50th, and 95th percentiles are 2, 335, and 1490, a dramatic change from the original assessments. The presence of substantial mass near zero reflects the non-negativity constraint. The combination of a large α_L and small β could lead to a negative L_i^*. Rather than creating a calibrated distribution that extends below zero, the algorithm concentrates mass at the boundary.

3 Debiasing Multiple Experts

3.1 A Bayesian Hierarchical Model

The model above focuses on a single expert, but common practice in risk analysis is to use multiple experts. It is not unreasonable to imagine that experts within a domain could have a tendency to display similar characteristics. In this section, we propose a Bayesian hierarchical model (e.g., Gelman et al. 1995) that captures this notion.

The essence of a hierarchical model is that an individual's parameters are modeled as if they are randomly drawn from a population distribution. For our model, we will identify an expert j's characteristics as α_{Lj}, α_{Uj}, and β_j. To begin, let α_{Lj} be distributed according to a gamma distribution:

$$\alpha_{Lj}|A_L, B_L \sim Ga(A_L + 1, B_L),$$

where A_L and B_L are hyperparameters with prior distributions

$$A_L \sim Pois(a_L)$$

and

$$B_L \sim Exp(b_L).$$

Given these specifications, the marginal density for α_{Lj} is

$$f(\alpha_{Lj}) = \frac{(a_L + 1)b_L\alpha_{Lj} + b_L^2}{(b + \alpha_{Lj})^3} e^{-a_L b_L/(b+\alpha_{Lj})} . \quad (1)$$

We let $a_L = b_L = 2$, which results in a relatively diffuse unconditional prior for α_{Lj} with mean near 1.

The hierarchical models for α_{Uj}, and β_j are specified in a corresponding manner. When implemented, this model produces posterior distributions for all hyperparameters ($A_L, B_L, A_U, B_U, A_\beta$, and B_β) as well as for all of the individual parameters for each expert (α_{Lj}, α_{Uj}, and β_j).

The hierarchical model provides a more complete environment for analyzing and calibrating expert judgments. In particular, it specifies a relationship among the experts. Data from all of the experts provides information about the hyperparameters, which in turn affects the posterior distributions for all α_{Lj}, α_{Uj}, and β_j. The example below demonstrates this effect.

3.2 Example

We continue with Cooke's space-debris data, demonstrating the hierarchical model using all seven experts. We again ran the MCMC model for 100,000 iterations after a burn-in period of 25,000 iterations. The candidate acceptance rate for the Metropolis-Hastings algorithm was 22.2%. Again, complete details are available from the authors.

We begin the discussion of the analysis by considering what these data suggest for a new, hypothetical expert for whom no calibration data are available. Assuming that this expert comes from the same population as those in the study, we can use the hierarchical model to make inferences about such a generic expert. Figure 4 shows the prior (the same for all three parameters) and posterior densities for α_L, α_U, and β for this expert. The prior is the unconditional density $f(\alpha_L)$ (Equation 1) and is implied by the prior distributions on the hyperparameters. Because their models are identical, α_U and β have the same prior density. To obtain the posterior densities, the data from the seven experts in the study have been used to update the prior densities for the hyperparameters,

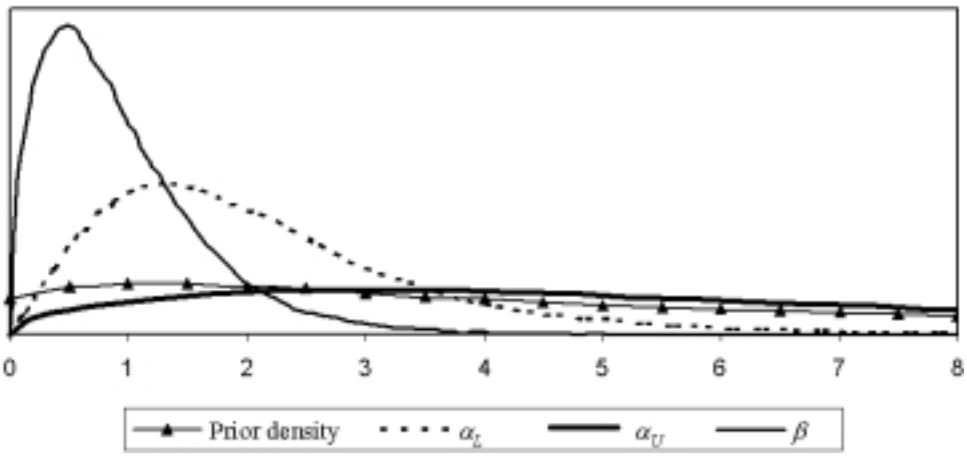

Figure 4: Prior and posterior distributions from the hierarchical model for a new, hypothetical expert's parameters α_L, α_U, and β. The prior density is the same for all three parameters.

and in turn to generate the unconditional distributions for the three parameters for the new expert. The change from the prior is substantial for α_L and β, but less so for α_U. Overall, we see that a new expert is expected to be highly overconfident; $P(\alpha_L < 1) = 0.20$ and $P(\alpha_U < 1) = 0.06$, with medians for α_L and α_U of 1.83 and 5.07, respectively. A typical expert would be expected to have a slight upward location bias, with median $\beta = 0.79$ and $P(\beta < 1) = 0.62$.

Suppose that the new, hypothetical expert provides a probability assessment for another space-debris variable, giving 5th, 50th, and 95th percentiles of 400, 450, and 500, respectively. By integrating over the posterior distributions for α_L, α_U, and β, we can produce a calibrated assessment for this new expert. Figure 5 shows the calibrated density. Note the long upper tail and the slight spike of mass at zero. This calibrated density has 5th, 50th, and 95th percentiles at 38, 417, and 1202, respectively, reflecting the overconfidence of experts in this population.

It is also instructive to compare the results of the hierarchical model with those from the single-expert model. Figures 6, 7, and 8 show the posterior densities for Expert 3's α_{L3}, α_{U3}, and β_3 using both models. In all three cases, the hierarchical model produces substantially different results. For example, for α_{L3} the hierarchical model places almost no mass below 1.5, whereas the single-expert model has 32% of the mass in this interval. Similar observations can be made for both α_U and β. In all three cases, the hierarchical model provides narrower densities, reflecting the fact that data from all seven experts have a bearing on the posterior densities for a specific expert.

Finally, we can consider how the hierarchical model would recalibrate Expert 3's assessment for Variable 27. Figure 9 shows the calibrated densities from both

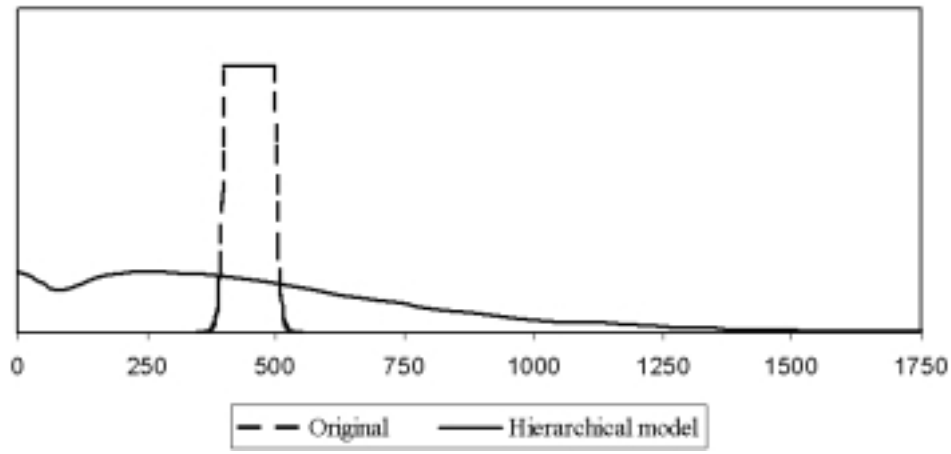

Figure 5: New expert's original and calibrated assessment using the hierarchical model.

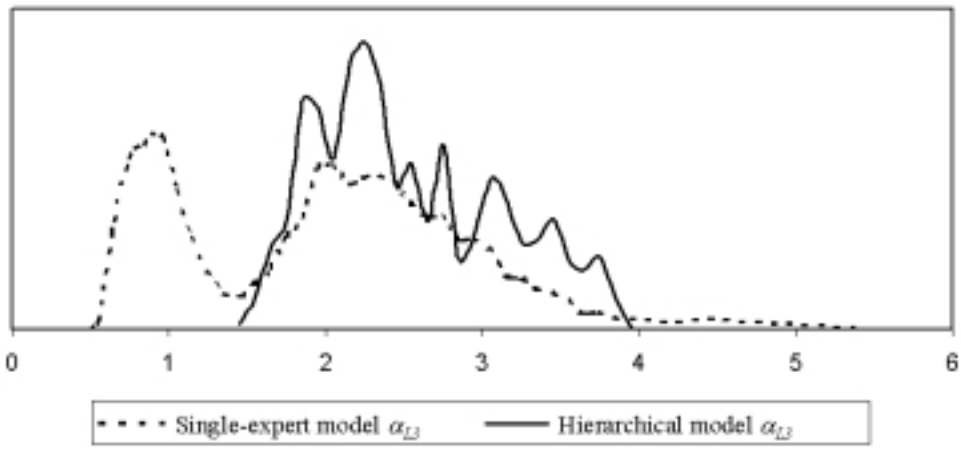

Figure 6: Posterior densities for α_{L3} for hierarchical and single-expert models.

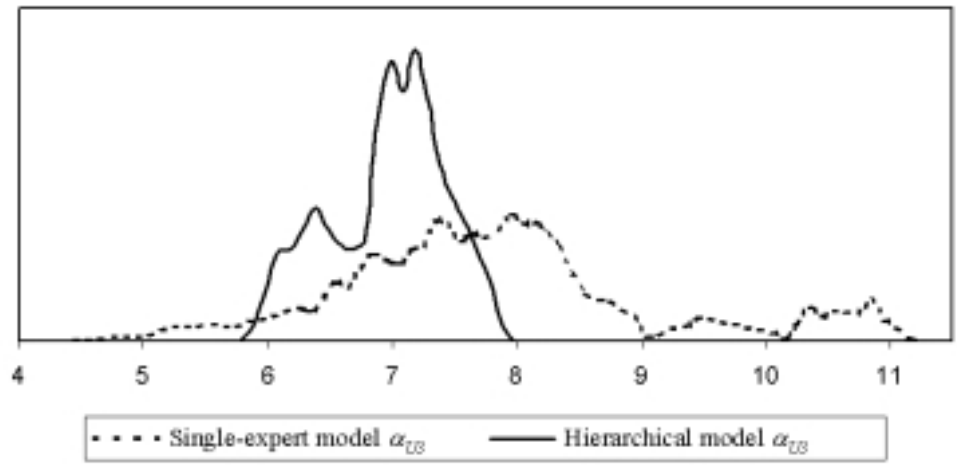

Figure 7: Posterior densities for α_{U3} for hierarchical and single-expert models.

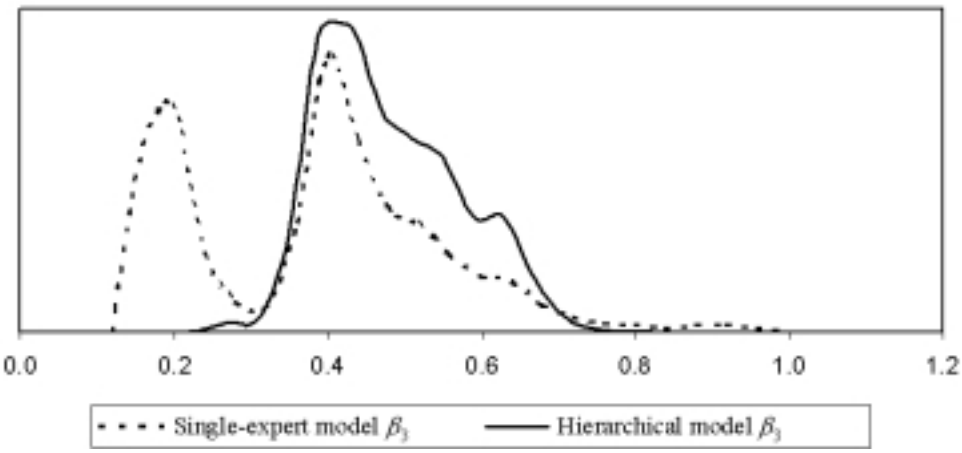

Figure 8: Posterior densities for β_3 for hierarchical and single-expert models.

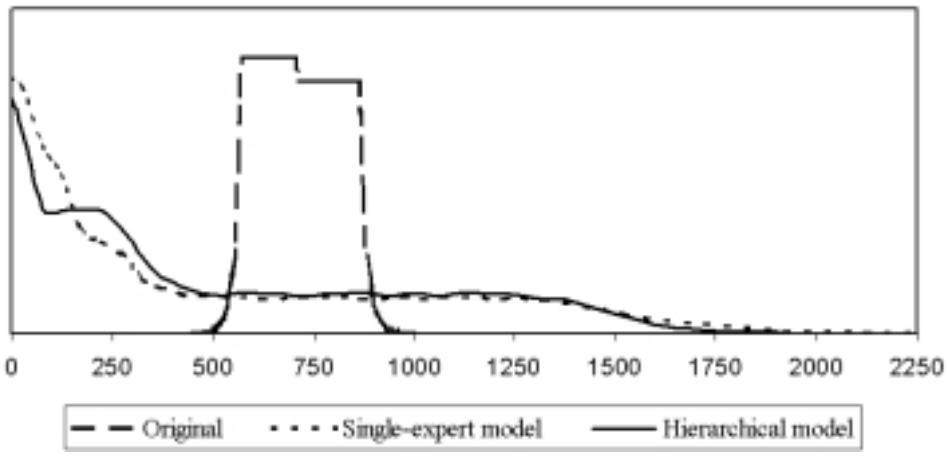

Figure 9: Expert 3's original and calibrated assessments for variable 27 using single-expert and hierarchical models.

models. Again, the hierarchical model gives different results, although in this case it is not radically different. The hierarchical model produces 5th, 50th, and 95th percentiles of 0, 374, and 1418, respectively, as compared to 2, 335, and 1490 for the single-expert model. The hierarchical model places slightly less mass near zero and in the lower portion of the domain, and likewise the the hierarchical model has slightly less mass in the upper tail. These observations reflect the fact that the hierarchical model has somewhat tighter densities for the parameters due to the incorporation of information from all experts.

4 Conclusion

Our Bayesian calibration model provides a way to debias expert probability assessments based on past performance data. Although the single-expert model is relatively straightforward to understand and implement, we prefer the hierarchical model. Modeling a population of experts provides important inferential advantages, which we have demonstrated: Any inferences about a single expert benefit from all the data, and the model enables the analyst to perform a preliminary calibration of a new expert before any specific performance data are available for that expert.

Our approach provides a way to adjust expert judgments after the fact. This is an "ex-post" approach to debiasing; an "ex-ante" approach would be to develop elicitation methods that counteract the expert's natural biases in the first place. Fischhoff (1982), Morgan & Henrion (1991), and McClelland & Bolger (1994) all discuss ex-ante debiasing techniques. For example, analysts can use counterfactual reasoning to push experts to consider extreme scenarios; doing

so can reduce overconfidence somewhat, but has not been shown to eliminate it (Koriat et al. 1980, Morgan & Henrion 1990). The most promising debiasing technique is training. While training can be effective in some circumstances (e.g., novel situations in which the assessor may learn what cues are diagnostic), the empirical results are not all positive. In particular, Morgan & Henrion (1991) discuss five studies that explicitly address overconfidence for continuous variables; although improvement was observed, in no case did training eliminate the bias. Benson & Onkal (1992) come to a similar conclusion in their review.

Another promising approach arises from the Brunswikian approach to probability assessment that has been recently promoted by Gigerenzer and others (see Gigerenzer 1991, Gigerenzer et al. 1991). This approach stresses the importance of asking an expert questions that are consistent with those typically encountered in his or her domain of expertise. Asking such questions is said to be "ecologically consistent" with the expert's experience and can improve calibration. In addition, framing assessment questions in terms of relative frequencies can improve calibration in comparison with the "degree of belief" framing typically used for subjective probability judgments. Neither of these approaches are a panacea, however. First, by their very nature risk assessments often involve asking experts questions that go beyond their day-to-day experience (e.g., the probability of a failure in a nuclear reactor containment vessel). Also, not all risk-assessment tasks are readily re-framed in frequency terms. Consider the space-debris example. If the problem is to assess the number of objects injected into orbit in 2010, how would one describe an equivalence class for which the expert could make a relative-frequency judgment?

This discussion suggests that ex-ante debiasing is a difficult and potentially unattainable goal. For that reason, we believe ex-post calibration of the type we describe here to be of value in risk assessment. The idea of ex-post calibration has been eschewed, however, by behavioral decision theorists and others for many years. Savage (1971) argues eloquently that "You might discover with experience that your expert is optimistic or pessimistic in some respect and therefore temper his judgments. Should he suspect you of this, however, you and he may well be on the escalator to perdition" (p. 796). Lichtenstein et al. (1982) and von Winterfeldt & Edwards (1986) add that ex-post calibration can require a substantial quantity of data.

Our model and example demonstrate that calibration can be performed with a data set of reasonable size. Although we do not have a strong counterargument to Savage, we believe that most experts would prefer to be calibrated and that a suitable system can be developed to accomplish calibration in an ex-post fashion. Finally, we believe that our Bayesian approach may prove valuable for analyzing experimental data from studies of ex-ante debiasing methods. In particular, where most studies to date have focused on aggregate characteristics of the population of experts, our model would permit researchers to study individual differences in calibration.

Although we have developed and demonstrated our model on the basis of three assessments (L_i, R_i, and U_i), the model readily extends beyond these three. For example, an expert might provide quartiles as well as the 5th, 50th,

and 95th percentiles. Also, by incorporating additional parameters, one could extend the model to handle experts who provide different quantiles. For example, one expert might provide 5th and 95th percentiles, and another 10th and 90th percentiles.

The model does include ad hoc assumptions for the purpose of translating an expert's judgments into a density function; we assume a piecewise-uniform density between the assessed quantiles and exponential tails. These specifications give us modeling convenience and tractability in the MCMC implementation. Other specifications are also possible, although we believe that the qualitative results—the calibrated expert density—will be robust to alternate specifications.

Our examples have demonstrated the feasibility of a Bayesian calibration process. The hierarchical model in particular has potential for further studies. For example, we may apply it to the same expert making assessments in different domains. If an expert makes judgments on almanac-type questions, those data may be useful for calibrating the expert in a specialty domain such as space debris or nuclear waste. Further, this model may extend to multiple experts making judgments in multiple domains and to the problem of combining judgments from multiple experts. In addition, scoring rules (Savage 1971, Winkler 1967, Winkler & Matheson 1976) can be used to analyze the performance of the calibrated judgments. Of particular interest are the relative performance of the single-expert and hierarchical models as well as the performance of calibrated judgments in a variety of domains. Research is under way on these and other issues.

References

[1] P. G. Benson and D. Onkal. The effects of feedback and training on the performance of probability forecasters. *International Journal of Forecasting*, 8:559—573, 1992.

[2] D. V. Budescu, I. Erev, T. S. Wallsten, and J. F. Yates. Stochastic and cognitive models of confidence [special issue]. *Journal of Behavioral Decision Making*, 10, 1997.

[3] R. M. Cooke. *Experts in Uncertainty: Opinion and Subjective Uncertainty in Science*. Oxford University Press, New York, 1991.

[4] D. R. Cox. Two further applications of a model for binary regression. *Biometrica*, 45:562—565, 1958.

[5] B. de Finetti. La prévision: Ses lois logiques, ses sources subjectives. *Annales de l'Institut Henri Poincaré*, 7:1—68, 1937.

[6] B. Fischhoff. *Debiasing*, pages 422—444. In Kahneman et al. [11], 1982.

[7] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis*. Chapman & Hall, London, 1995.

[8] G. Gigerenzer. How to make cognitive illusions disappear: Beyond heuristics and biases. *European Review of Social Psychology*, 2, 1991.

[9] G. Gigerenzer, U. Hoffrage, and H. Kleinbolting. Probabilistic mental models: A Brunswikian theory of confidence. *Psychological Review*, 98:506—528, 1991.

[10] D. W. Griffin and C. A. Varey. Towards a consensus on overconfidence. *Organizational Behavior and Human Decision Processes*, 65:227—231, 1996.

[11] D. Kahneman, P. Slovic, and A. Tversky, editors. *Judgment Under Uncertainty: Heuristics and Biases*. Cambridge University Press, Cambridge, 1982.

[12] G. Keren. Calibration and probability judgments: Conceptual and methodological issues. *Acta Psychologica*, 77:217—273, 1991.

[13] J. Klayman, J. B. Soll, C. González-Vallejo, and S. Barlas. Overconfidence: It depends on how, what, and whom you ask. *Organizational Behavior and Human Decision Processes*, 79:216—247, 1999.

[14] A. Koriat, S. Lichtenstein, and B. Fischhoff. Reasons for confidence. *Journal of Experimental Psychology: Human Learning and Memory*, 6:107—118, 1980.

[15] S. Lichtenstein and B. Fischhoff. Do those who know more also know more about how much they know? the calibration of probability judgments. *Organizational Behavior and Human Performance*, 20:159—183, 1977.

[16] S. Lichtenstein, B. Fischhoff, and L. D. Phillips. *Calibration of probabilities: The state of the art to 1980*, pages 306–334. In Kahneman et al. [11], 1982.

[17] A. G. R. McClelland and F. Bolger. *The calibration of subjective probabilities: Theories and models 1980-94*, pages 453—482. In Wright and Ayton [27], 1994.

[18] M. W. Merkhofer. Quantifying judgmental uncertainty: Methodology, experiences, and insights. *IEEE Transactions on Systems, Man, and Cybernetics*, 17:741—752, 1987.

[19] M. G. Morgan and M. Henrion. *Uncertainty: A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis*. Cambridge University Press, Cambridge, 1990.

[20] P. A. Morris. Decision analysis expert use. *Management Science*, 20:1233—1241, 1974.

[21] L. J. Savage. The elicitation of personal probabilities and expectations. *Journal of the American Statistical Association*, 66:783—801, 1971.

[22] A. I. Shlyakhter. Improved framework for uncertainty analysis: Accounting for unsuspected errors. *Risk Analysis*, 14:441—447, 1994.

[23] A. I. Shlyakhter, D. M. Kammen, C. L. Brodio, and R. Wilson. Quantifying the credibility of energy projections from trends in past data: The u.s. energy sector. *Energy Policy*, 22:119—130, 1994.

[24] D. von Winterfeldt and W. Edwards, editors. *Decision Analysis and Behavioral Research*. Cambridge University Press, Cambridge, 1986.

[25] R. L. Winkler. The quantification of judgment: Some methodological suggestions. *Journal of the American Statistical Association*, 62:1105—1120, 1967.

[26] R. L. Winkler and J. E. Matheson. Scoring rules for continuous probability distributions. *Management Science*, 22:1087—1096, 1976.

[27] G. Wright and P. Ayton, editors. *Subjective Probability*. John Wiley, New York, 1994.