Combining Experts: Decomposition and Aggregation Order

Erin Baker*, Olaitan Olaleye†

September 14, 2012

*Correspondence Address: Erin Baker, 220 Elab, University of Massachusetts, Amherst, MA 01003; edbaker@ecs.umass.edu; 413-545-0670

†Olaitan Olaleye, University of Massachusetts, Amherst, MA 01003, USA , oolaleye@engin.umass.edu

Abstract

In this paper we ask, if quantities in an elicitation have been decomposed, is it better to combine experts before or after recomposing the quantities? We find that combining experts earlier, before re-composition of the quantities, leads to smaller errors with less variance. A simulation shows that these differences may be quite small on average; while an application to actual data shows that the differences can be significant in individual decision problems.

KEY WORDS: Expert judgment; decomposition; aggregation; correlation;

1 INTRODUCTION

This paper addresses a simple question associated with combining expert opinions. If quantities in an elicitation have been decomposed, is it better to combine experts before or after recomposing the quantities? This question boils down to whether expert opinions should be combined early or late. It has been noted that results will generally differ depending on this aggregation order [3][26].

Mosleh and Bier [26] addressed this question in their paper on decomposition and aggregation error. They show that what they term "aggregation order" will almost always have an effect, but they concluded that when using mathematical averaging... "[u]nfortunately, it is not immediately apparent which aggregation order is more accurate, nor is it clear which approach (if any) is more appropriate on axiomatic grounds". In this paper we show that when using the commonly applied arithmetic average, it is more accurate to combine experts early rather than late. Mosleh and Bier [26] go on to analyze a similar problem involving Bayesian updating. They show that it is generally better to perform Bayesian updating on each individual component first, and then combine the components to get the quantity of interest. This is parallel with the result in this paper.

Several authors have shown that decomposing probabilities to elicit them is often superior [1][2][10][28]. If there are multiple experts with differing opinions, and if one wants to use the probabilities in a decision model, then one is faced with the decision of how to combine the probabilities across experts, and specifically at what level. There are a number of issues related to this question, including incoherent or abstaining experts [27], experts who partition events in different ways [8], and correlation between experts [23]. While each of these papers presents solutions to these problems, none of them address the question of aggregation order. In this paper, we pay special attention to one particular issue – what we call the level of experts' self-correlation – and on the question of aggregation order. We show that combining experts earlier is generally better, but if it is likely that experts are positively or negatively self-correlated regarding compound events then aggregating earlier is even better.

It is well understood that the judgements of multiple experts can be correlated ([9][15][16][19][21] [22][25][33]). It also appears that experts can be self-correlated. By self-correlated, we mean that probability judgements from a single expert about two otherwise independent events may be correlated to each other. The most obvious example of positive correlation is that many experts seem to be optimists or pessimists. Cooke [12] provides some evidence for this from a particular study. Another example of self-correlation may be an expert who is biased toward a particular technology, and thus gives it higher probabilities all around and lower probabilities to its competitors: this expert would have positive correlation within the preferred technology, but a negative correlation between technologies. Harrison [18] argues more generally that experts will tend to have self-correlation due to uncertainty about their own calibration.

There is some controversy about whether experts should be aggregated at all. Morgan and Henrion [24] argue that "faced with public decision making environments in which for a variety of reasons the decision makers cannot make formal quantitative evaluations of the alternative expert views or their consequences, the analyst can help by performing parametric analysis that allows the decision maker to work the problem backward, that is, to see the consequences of a range of possible alternative combinations or weightings." Additionally, they comment that it may not be best to combine opinions at the input level, implying it may be better at the output level. Similarly Keith [20] argues that "it is rarely appropriate to combine divergent expert judgments." These arguments imply that expert opinion should be combined later (and implicitly), rather than earlier (and explicitly). Our results in this paper, however, suggest that combining expert opinions later in the process will tend to lead to systematic errors.

We consider only mathematical combination, rather than behavioral approaches. Clemen and Winkler [10] conclude in their review of the literature that mathematical and behavioral aggregation perform about equally, with mathematical aggregation having a slight edge. Additionally, while there exists several forms of mathematical aggregation, including Cooke's Classical Method and other weighting schemes [13][14][29][30][31][34] and Bayesian aggregation, they conclude that simple combination rules, such as a simple average, perform quite well. For example, Clements and Harvey [11] consider different methods of combining probabilistic forecasts, and find that while the simple average is not theoretically the best, it performs the best in a simulation when the number of experts is small. Given that a simple average is also a great deal less resource-intensive than behavioral approaches or more complex mathematical approaches, this is the combination method we focus on.

A related question to ours is whether, and to what degree, to decompose quantities in elicitations. For example, if we suspect a high likelihood of modeling error, then decomposition may not be a good idea [26]. In general, decomposition is not warranted if the analyst is not sure about the model, not confident about the quality of the decomposed probabilities, or does not have the resources to decompose the problem. Azaiez and Bier [3] discuss this and conclude that, ."...determining the appropriate level of decomposition is likely to involve trade-offs among several competing sources of error, among which aggregation error is one." This paper does not address this question directly, but in providing a way to minimize aggregation error may increase the number of cases in which decomposition is warranted.

The rest of the paper is organized as follows. In Section 2 we present a simple model of elicited probabilities, and show that the mean and variance of the errors are larger when opinions are combined later. Our simple model is only an approximation, however, as it may lead to probabilities that are less than zero or greater than 1. So, in Section 3 we perform a simulation that allows us

to relax some of our assumptions and avoid this problem. In Section 4 we analyze some real data from previous elicitations, to give a sense of the impact of the different aggregation methods in a real problem. We conclude in Section 5.

2 MODEL

We present a very simple model of elicited probabilities. We consider independent (or conditionally independent) events indexed by i. We are interested specifically in multiplicative decomposition or the intersection of multiple events. For an example of decomposition, event i may be "achieving a solar cell with 15% efficiency" with probability p_i and event i' may be "achieving a cost of \$50/m2, given an efficiency of 15%" with probability $p_{i'}$. The joint event of interest is "achieving a solar cell with 15% efficiency and a cost of \$50/m2" with a probability of $p_i p_{i'}$. For an example of the intersection of multiple events, consider event i (i') is success in technology i (i'). If we are interested in a portfolio of technologies, then we are interested in the event that multiple technologies are successful.

We assume that each event i has a "true" probability, p_i, that represents the degree of belief of the entire community if all biases and errors could be avoided. However, the individual probabilities that are elicited from expert j , q_{ij} are subject to a number of deviations from the true probability (which we will call errors), including errors unique to the particular event i that are correlated across experts; errors unique to the expert j that are correlated across events ("self-correlation"); and idiosyncratic errors δ_{ij}. We define an extremely simple, additive model, as follows:

$$q_{ij} = p_i + \varepsilon_{ij} + \mu_{ij} + \delta_{ij} \tag{1}$$

where each of the different types of errors are independent of each other (i.e. ε and μ, ε and δ, etc.). Each individual error has an *a priori* mean of zero – that is, we have no way to predict the direction of any of the errors. The ε_{ij} are correlated across experts but independent within experts; the μ_{ij} are correlated within experts but not across; the δ_{ij} are independent. That is, ε_{ij} is correlated to $\varepsilon_{ij'}$, but is independent of $\varepsilon_{i'j}$. μ_{ij} is correlated with $\mu_{i'j}$ but not $\mu_{ij'}$. Note that this model is similar to the psychometric model proposed by Wallsten et al. [32], and used in Ravinder et al. [28], except that we decompose their variable error e into three parts. They used this model to consider the reliability of total probabilities calculated from elicited marginal probabilities, while we focus on joint probabilities. Note that our formulation may lead to elicited probabilities q_{ij} that are not between 0 and 1. We address this issue in the next section with a simulation.

We are interested in the probability of the joint event

$$p_{ii'} = p_i p_{i'} \tag{2}$$

We consider the case of simple, equal-weighted averaging over experts. Given this, we have two choices for mathematical combination. We can combine the individual probabilities for each event and then calculate the probability of the joint event (we will call this Method I); or we can calculate the probability of the joint event for each expert and then combine those (we will call this Method II). We will compare these two aggregation orders.

We start by calculating the expected error for the two methods. We find there is a systematic error due to self-correlation. That is, the expected value of the error term is non-zero, and it is larger in Method II than in Method I. We then calculate the variance of the error terms, and find that it is larger in Method II than Method I. We show that this is true for all three errors – independent, correlated across experts, and correlated within experts.

2.1 Method I: Average Experts First

If we combine experts before we recompose, we get an estimated probability for each event i, as follows:

$$q_i^{(n)} \equiv \frac{1}{n} \sum_{j=1}^{n} q_{ij} \tag{3}$$

where n is the number of experts. Similarly define $\varepsilon_i^{(n)}, \mu_i^{(n)}, \delta_i^{(n)}$. Then the estimated probability of the intersection is

$$\begin{aligned} q_{ii'}^{(n)} &\equiv \frac{1}{n^2} \sum_{j=1}^{n} q_{ij} \sum_{j=1}^{n} q_{i'j} & (4) \\ &= \left(p_i + \varepsilon_i^{(n)} + \mu_i^{(n)} + \delta_i^{(n)}\right) \left(p_{i'} + \varepsilon_{i'}^{(n)} + \mu_{i'}^{(n)} + \delta_{i'}^{(n)}\right) & (5) \end{aligned}$$

Let ρ represent the correlation between each μ_{ij} and $\mu_{i'j}$, and σ^2 represent the variance of μ_{ij}. Using these definitions we show in Appendix A.1 that

$$E\left[q_{ii'}^{(n)}\right] = p_i p_{i'} + \frac{\rho\sigma^2}{n} \tag{6}$$

Thus, self-correlation leads to a systematic error – the estimator $q_{ii'}^{(n)}$ is biased.

2.2 Method II: Average experts second

If we calculate the probability of the combined event first we get $q_{ij}q_{i'j}$ for each expert, then average across that. In this case the estimated joint probability for n experts will be

$$\tilde{q}_{ii'}^{(n)} \equiv \frac{1}{n}\sum_{j=1}^{n} q_{ij}q_{i'j} \tag{7}$$

Following the same method as above (see Appendix A.2 for details) we show that

$$E\left[\tilde{q}_{ii'}^{(n)}\right] = p_i p_{i'} + \rho\sigma^2 \tag{8}$$

Thus, the absolute value of the expected error in this case is systematically larger than the absolute value of the expected error in the first case. Moreover, it is independent of the number of experts! That is, Method II does not correct for the problem of experts' self correlation at all. Specifically, the expected value of the error in Method II is n times the expected error in Method I.

2.3 Variance of errors

From the above, we see that the expected value of the error is larger in Method II. But, it would be nice to have a sense of the distribution of the errors. This is difficult, however, since there is no closed form representation of the distribution of the product of two correlated variables, even under the assumption that they are normal. In this section we focus on the distribution of only the error terms, rather than of the entire elicited probability. In Method I, we define the error term as:

$$\mu_{ii'} \equiv \mu_i^{(n)}\mu_{i'}^{(n)} \tag{9}$$

and for Method 2:

$$\tilde{\mu}_{ii'} \equiv \frac{1}{n}\sum_{j=1}^{n} \mu_{ij}\mu_{i'j} \tag{10}$$

In order to analyze this generally we need to make an assumption about how the variables are correlated. Let $\mu_{i'j} = \rho\mu_{ij} + \sqrt{1-\rho^2}x_{i'j}$, where μ_{ij} and $x_{i'j}$ are independent with mean 0 and standard deviation σ. Then μ_{ij} and $\mu_{i'j}$ have correlation ρ and the same means and variances.

In order to find the variances of the error terms, we need to find their second moments. For Method I, we show in Appendix A.3 that, under the assumption that each μ_{ij} is normally distrib-

uted:

$$E\left[\mu_{ii'}^2\right] = \frac{1}{n^4}\left[\rho^2 3n^2\sigma^4 + \left(1-\rho^2\right)n^2\sigma^4\right] = \frac{\sigma^4}{n^2}\left[1+2\rho^2\right] \tag{11}$$

Subtracting off the square of the mean we get the variance:

$$var\left[\mu_{ii'}\right] = \frac{\sigma^4}{n^2}\left[1+2\rho^2\right] - \frac{\rho^2\sigma^4}{n^2} = \frac{\sigma^4}{n^2}\left(1+\rho^2\right) \tag{12}$$

Method II follows a similar argument to get:

$$var\left[\tilde{\mu}_{ii'}\right] = \frac{\sigma^4}{n}\left(1+\rho^2\right) \tag{13}$$

This result is similar to the result for the means, in that the variance of the error in Method II is n times the variance of the error in Method I. Note that in the case where the errors are independent, these equations hold without the assumption of normality. In particular, we see that the variance in Method II is still higher than the variance in Method I, even in the absence of correlation. This is interesting because it means that even if there is no self-correlation, if we only expect each expert to have independent, idiosyncratic errors, then Method I is still preferable to Method II. That is, this analysis equally applies to the uncorrelated errors δ_{ij}.

In Figure 1 we use these results to illustrate how the errors converge as we add more experts for the two methods. We show the mean error and one standard deviation above and below, under the assumptions of $\sigma = 0.2$ and $\rho = 0.8$. We see that Method I converges to zero, with relatively tight error bands as the number of experts grows. Method II on the other hand, does not converge to zero. The mean error is fixed regardless of the number of experts. The error bands decrease in the number of experts, but are quite large compared to Method I.

2.3.1 Error correlated across experts (but not within)

We can also look at how the methods compare on the variance of the error that is correlated across experts, but not within experts: ε_{ij}. We focus only on the error term that is not shared by the two methods. From equation (6) we see that the expected value of this error is 0, thus we focus only on the variance. Define the correlation across experts to be ρ_ε. In Appendix A.4, we show that for Method I:

$$var\left[\varepsilon_i^{(n)}\varepsilon_{i'}^{(n)}\right] = \frac{1}{n^2}\left(1+(n-1)\rho_\varepsilon\right)^2\sigma_\varepsilon^4 \tag{14}$$

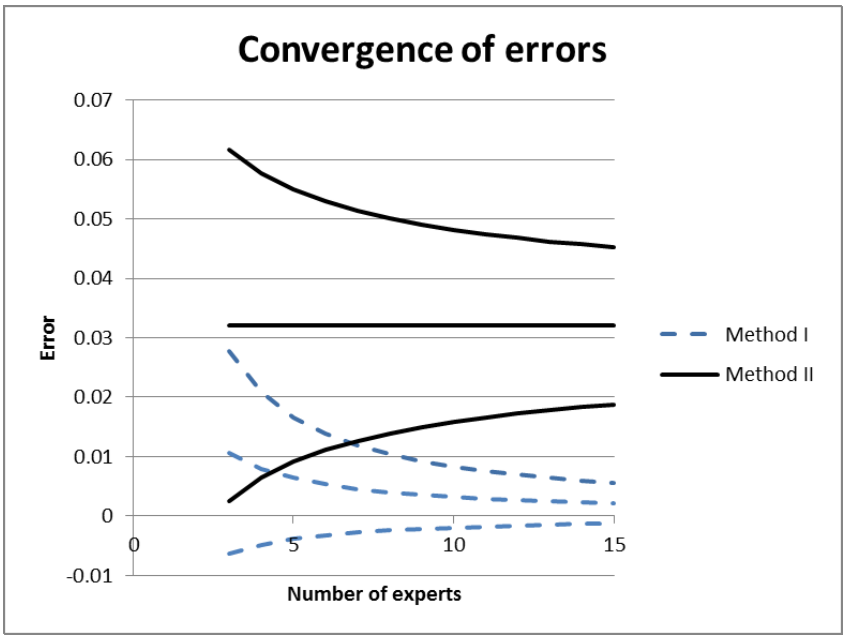

Figure 1: Convergence of errors

While for Method II:

$$var\left[\frac{1}{n}\sum_{j}^{n}\varepsilon_{ij}\varepsilon_{i'j}\right] = \frac{1}{n}\left(1+(n-1)\,\rho_{\varepsilon}^{2}\right)\sigma_{\varepsilon}^{4} \tag{15}$$

When the correlation among experts $\rho_{\varepsilon} = 0$, then the variances of the two methods are $\frac{1}{n^2}\sigma_{\varepsilon}^4$ and $\frac{1}{n}\sigma_{\varepsilon}^4$ respectively, consistent with Section 2.3 above. When $\rho_{\varepsilon} = 1$ both variances are the same: σ_{ε}^2. This second result is because the error terms are the same ($\varepsilon_i^{(n)}\varepsilon_{i'}^{(n)} = \frac{1}{n}\sum_j^n \varepsilon_{ij}\varepsilon_{i'j}$) for the two methods when there is perfect correlation. Otherwise, Method II has a larger variance than Method I.

2.4 Unions of events

Here we briefly discuss what happens if the decision tree includes structures other than joint events. First consider the union of events. If the events are independent and the elicitation is decomposed into the two individual events p_1 and p_2 then the probability of event 1 or event 2 is $p_1 + p_2 - p_1p_2$. Thus, the analysis is the same as above, except that the sign of the error would be opposite (See Appendix A.5). If the events are dependent, then it depends on how the elicitation proceeds. If,

for example, the analyst elicits the probability of event 1 and the probability of event 2 given that event 1 does not happen, then the analysis is the same. If, on the other hand, the analyst explicitly elicits the probability of the joint event p_{12}, then the expected error is zero and the distribution of the error is the same in both methods. All other structures are combination of intersections and unions of events.

Table 1: Theory Summary.

	Method I	Method II
Estimated Probability of Joint Event	$q_{ii'}^{(n)} \equiv \frac{1}{n}\sum_{j=1}^{n} q_{ij} \frac{1}{n}\sum_{j=1}^{n} q_{i'j}$	$\tilde{q}_{ii'}^{(n)} \equiv \frac{1}{n}\sum_{j=1}^{n} q_{ij} q_{i'j}$
Expected Value of Error	$\frac{\rho\sigma^2}{n}$	$\rho\sigma^2$
Variance of Error Term	$\mathrm{var}[\mu_{ii'}] = \frac{\sigma^4}{n^2}(1+\rho^2)$	$\mathrm{var}[\tilde{\mu}_{ii'}] = \frac{\sigma^4}{n}(1+\rho^2)$

2.5 Summary

Table 1 summarizes the results. We have shown that the absolute value of the expected error of Method II is larger than that of Method I as long as experts exhibit some self-correlation, either positive or negative. Moreover, the variance of the error terms from all three errors is larger for Method II than for Method I.

3 SIMULATION

In this section we run some simulations to better understand how the distribution of errors is impacted by the two methods. We first look at a simulation based directly on our theoretical model, but consider a wider variety of metrics to evaluate the two errors. We then use a different theoretical model, based on log odds, that removes the problem of probabilities below zero or above one.

Above we have calculated the expected error of the combined probability and the variance

of the error term. Here we look at some other metrics of interest. First, we consider the mean absolute error: $E\left[|Q-p_{ii}|\right]$ where Q represents the estimated joint probability $q_{ii}^{(n)}$ for Method I and $\tilde{q}_{ii}^{(n)}$ for Method II. This tells us the absolute size of our error on average. Second, we consider $var\left[Q-p_{ii}\right]$. This differs from the variance of the error term, as there are cross products. In general, these variances will be higher than the variance of the error terms, and there will be less difference between the two methods. Finally, we look at the chance that Method II turns out to be better than Method I: $\Pr\left(\left|q_{ii}^{(n)}-p_{ii}\right|>\left|\tilde{q}_{ii}^{(n)}-p_{ii}\right|\right)$.

3.1 Simulation Based on Theoretical Model

We assume 5 experts. In each case we first randomly generated p_i and $p_{i'}$ as uniform between 0 and 1. These are fixed across the 5 experts. We then generated ε_{ij} for each expert. These were distributed normally with mean 0 and various assigned standard deviations σ_ε as reported in the tables below. Each of the ε_{ij} were generated to be correlated with correlation $\sqrt{\rho_\varepsilon}$ to an unused 'support' expert, implying that the experts are correlated to each other with correlation ρ_ε. We generated μ_{ij} for each expert as well, again with mean 0. The μ_{ij} were correlated with the $\mu_{i'j}$. We also generated an independent random error δ_{ij}. We ran a total of 1,000,000 simulations for each combination of parameters.

Table 2 shows the results for various values of σ and ρ, when the ε and the δ are set to zero. The first two columns report the self-correlation ρ and the standard deviation σ of the errors μ_{ij}. The 3rd and 4th columns report the mean error $E\left[Q-p_{ii}\right]$ (averaged over the 1,000,000 simulations). These match the theoretically predicted values. The 5th and 6th columns report the $var\left[Q-p_{ii}\right]$. We note that these are much higher than the variance of the error terms, and the difference between them is smaller. Nevertheless, Method I always has a smaller variance than Method II. The 7th and 8th columns report on the mean of the absolute error $E\left[|Q-p_{ii}|\right]$. These errors are larger than the the mean errors, and again the difference between them is not as large. Finally, the last column reports the probability that the absolute error in Method I is higher than the absolute error in Method II. In all cases, Method I is superior to Method II, although the difference is not always large: the probability that Method II is better is larger than 0.45 in more than half the cases.

Figure 2 illustrates the percentage by which the absolute mean error in Method II is larger than in Method I. That is, this figure shows

$$\frac{E\left[\left|\tilde{q}_{ii}^{(n)}-p_{ii}\right|\right]}{E\left[\left|q_{ii}^{(n)}-p_{ii}\right|\right]}-1 \tag{16}$$

If this value is greater than 0 then Method I has a smaller error than Method II. The larger this

Table 2: Simulation Outputs.

μ_{ij}		Mean(Q-P)		Var(Q-P)		Mean|Q-P|		$\Pr\{\|Q_{II}\text{-P}\|<\|Q_I\text{-P}\|\}$
ρ Corr	σ sd	Mt I	Mt II	Mt I	Mt II	Mt I	Mt II	
0	0.05	0	0	0.0003	0.0003	0.0136	0.0137	0.4876
0	0.1	0	0	0.0013	0.0014	0.0274	0.0276	0.478
0	0.2	-0.0001	-0.0001	0.0054	0.0057	0.0549	0.0566	0.4559
0	0.5	0	0.0001	0.0358	0.0459	0.1397	0.163	0.4012
0.2	0.05	0.0001	0.0005	0.0004	0.0004	0.0146	0.0146	0.4884
0.2	0.1	0.0004	0.002	0.0015	0.0016	0.0293	0.0295	0.4781
0.2	0.2	0.0016	0.008	0.0062	0.0065	0.0585	0.0604	0.4579
0.2	0.5	0.0101	0.05	0.0409	0.0512	0.1481	0.1733	0.4023
0.8	0.05	0.0004	0.002	0.0005	0.0005	0.0171	0.0172	0.4801
0.8	0.1	0.0016	0.008	0.0021	0.0022	0.0342	0.035	0.4606
0.8	0.2	0.0065	0.0321	0.0087	0.0091	0.0684	0.0745	0.4223
0.8	0.5	0.0402	0.2002	0.0574	0.0739	0.172	0.2512	0.3136
1	0.05	0.0005	0.0025	0.0006	0.0006	0.0178	0.018	0.4752
1	0.1	0.002	0.01	0.0023	0.0024	0.0357	0.0368	0.452
1	0.2	0.008	0.04	0.0095	0.01	0.0715	0.0799	0.4032
1	0.5	0.05	0.2502	0.0635	0.0835	0.1797	0.2884	0.2718
-0.8	0.05	-0.0004	-0.002	0.0001	0.0001	0.0086	0.0088	0.4508
-0.8	0.1	-0.0016	-0.008	0.0005	0.0006	0.0172	0.0189	0.4057
-0.8	0.2	-0.0064	-0.032	0.0022	0.0027	0.035	0.0464	0.3293
-0.8	0.5	-0.0399	-0.1999	0.0174	0.0338	0.0967	0.2138	0.1831

value is, the better is Method I in comparison to Method II. The figure shows that Method I always has a smaller expected error, and that the error is impacted more strongly by the variance of the error than by the correlation.

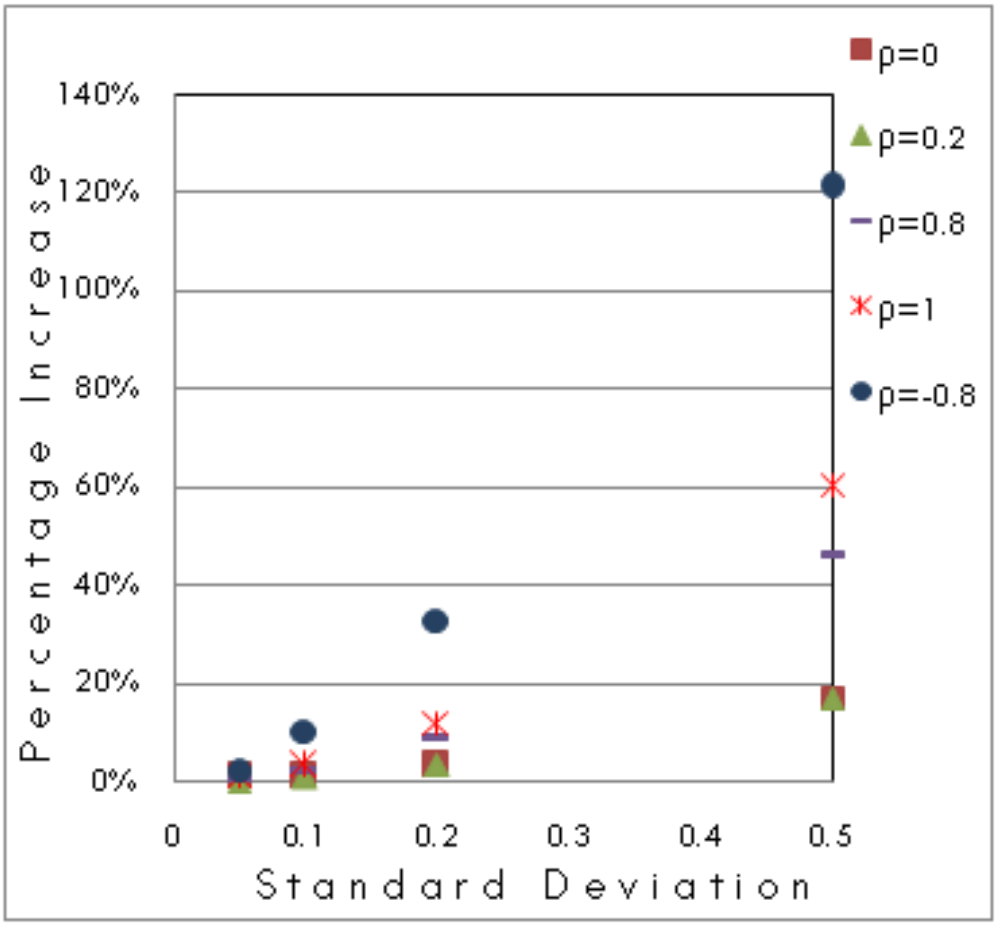

Figure 2: Percentage difference of mean absolute error between Method I and II. Correlation ■ refers to self-correlation between experts.

We ran this experiment with a number of different combinations of parameters, including all permutations of the 5 values of $\rho_\mu = \{-.08, 0, 0.2, 0.8, 1\}$ and the 4 values of $\sigma_\mu = \{0.05, 0.1, 0.2, 0.5\}$ with the 18 specific combinations of the other parameters shown in Table 3, for a total of 360 cases. In every case the mean error, the mean absolute error, and the variance are larger under Method II than Method I.

Table 3: Actual parameters simulated.

ρ_ε	0	0.05	0.05	0.05	0.05	0.05	0.1	0.1	0.1	0.1	0.1	0.2	0.2	0.2	0.2	0.2	0.5	0.5
σ_ε	0	0.05	0.1	0.1	0.2	0.2	0.05	0.1	0.1	0.2	0.2	0.1	0.1	0.2	0.2	0.2	0.2	0.5
σ_δ	0	0.05	0.05	0.1	0.05	0.2	0.05	0.05	0.1	0.05	0.2	0.05	0.1	0.05	0.1	0.2	0.2	0.2

3.2 Log odds model

We perform a second set of simulations, using a log-odds model. First, the log-odds of the true probabilities are calculated:

$$r_i \equiv \ln \frac{p_i}{1 - p_i} \tag{17}$$

Second, the log-odds estimate is derived by adding normal error terms to this value:

$$y_{ij} = r_i + \varepsilon_{ij} + \mu_{ij} + \delta_{ij} \tag{18}$$

Finally, the elicited value is calculated by reversing the process:

$$q_{ij} = \frac{e^{y_{ij}}}{1 + e^{y_{ij}}} \tag{19}$$

This assures that the elicited probabilities, q_{ij}, will be between 0 and 1. With the exception of the q_{ij} all values are generated in the same manner as above.

Table 4 shows the results when ε and the δ are simply set to zero. The results show a similar pattern to the previous section. Additionally, we perform simulations using combinations of the parameters from Table 5. We considered all combinations of ρ_μ and σ_μ in combination with 14 combinations of the other parameters, for a total of 210 cases. Figure 3, similar to Figure 2, illustrates the percentage by which the absolute mean error in Method II is larger than in Method I, for a number of cases. The legend refers to the values of $\rho_\varepsilon, \sigma_\varepsilon, \sigma_\delta$ respectively. For example, the middle entry in the legend, "med, high, high," refers to the simulation of the medium non-zero value of $\rho_\varepsilon = 0.5$, along with the highest values for $\sigma_\varepsilon = 1$, and $\sigma_\delta = 1$. Each of these cases is shown for the three different values of σ_μ, and graphed against ρ_μ. The results obtained here are similar to the base model simulation results with Method I outperforming Method II.

3.3 Summary

In this section we have simulated the aggregation errors caused by Method I and Method II using two different models of error generation. In all cases we find that Method I performs better than Method II, in terms of the mean error, the mean absolute error, and the variance of the errors.

Table 4: Log Odds Simulation Outputs.

μ_{ij}		Mean(Q-P)		Var(Q-P)		Mean|Q-P|		Pr{|Q_{II}-P|<|Q_I-P|}
ρ Corr	σ sd	Mt I	Mt II	Mt I	Mt II	Mt I	Mt II	
0	0.2	0	0	0.0002	0.0002	0.0093	0.0093	0.4928
0	0.5	0	0	0.0011	0.0011	0.0233	0.0234	0.4816
0	1	0	0.0001	0.0042	0.0043	0.0463	0.0471	0.4698
0.2	0.2	0	0.0002	0.0002	0.0002	0.0099	0.0099	0.4878
0.2	0.5	0.0002	0.0013	0.0012	0.0012	0.0247	0.0249	0.4735
0.2	1	0.001	0.0047	0.0047	0.0048	0.0488	0.05	0.4549
0.8	0.2	0.0002	0.0009	0.0003	0.0003	0.0114	0.0115	0.4748
0.8	0.5	0.0011	0.0053	0.0016	0.0016	0.0284	0.0291	0.4436
0.8	1	0.0037	0.0186	0.0061	0.0062	0.0558	0.0593	0.4065
1	0.2	0.0002	0.0011	0.0003	0.0003	0.0119	0.0119	0.4729
1	0.5	0.0014	0.0067	0.0018	0.0018	0.0295	0.0304	0.4367
1	1	0.0047	0.0234	0.0065	0.0067	0.058	0.0625	0.3953
-0.8	0.2	-0.0002	-0.0009	0.0001	0.0001	0.0063	0.0063	0.4907
-0.8	0.5	-0.0011	-0.0053	0.0006	0.0006	0.0161	0.0166	0.4797
-0.8	1	-0.0037	-0.0186	0.0023	0.0025	0.0337	0.0359	0.4683

Table 5: Parameters for simulation for Log Odds model.

ρ_μ				
-0.8	0	0.2	0.8	1

σ_μ		
0.2	0.5	1

ρ_ε		
0.2	0.5	1

σ_ε		
0.2	0.5	1

σ_δ		
0.2	0.5	1

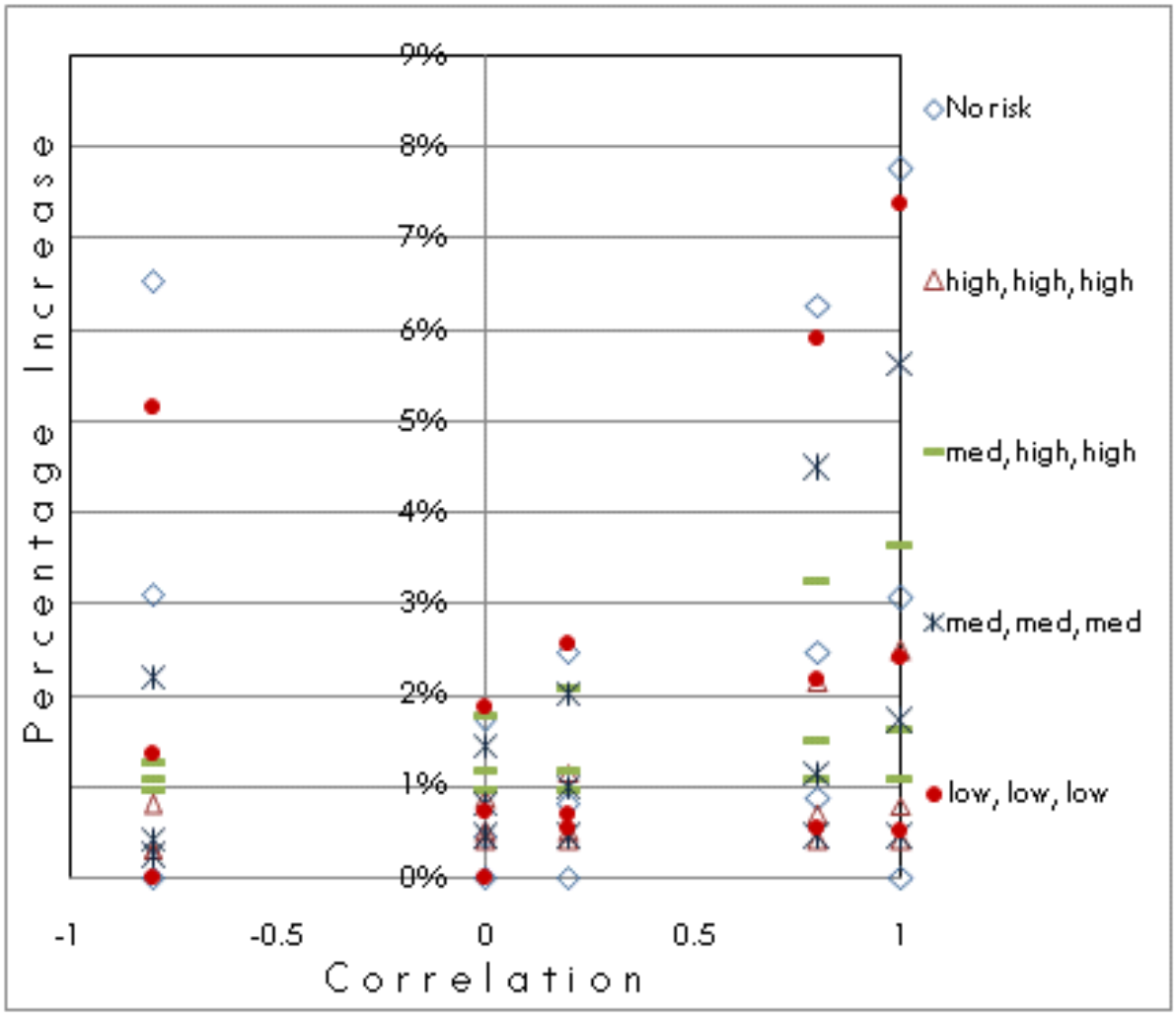

Figure 3: Log Odds Model: Percentage difference between Method I and II. The legend refers to values of $\rho_\varepsilon, \sigma_\varepsilon, \sigma_\delta$ respectively. No Risk refers to the case where each of the parameters $\rho_\varepsilon, \sigma_\varepsilon, \sigma_\delta$ =0. Correlation refers to self-corelation.

Nevertheless, Method II can outperform Method I in individual cases given the randomness of the responses. We found that the probability that Method II gives a better estimate ranges from about 40% to 49% in the log-odds model. In fact, in the single experiment we are aware of, Method II did outperform Method I [17]. However, it would require a very large data set of multiple elicitations to empirically test our results.

4 EMPIRICAL EXAMPLE

In this section we consider the impact that the two different methods have on a real example. We apply the two methods of aggregation to elicitation data on climate change energy technology R&D, and then use both sets of aggregated probabilities in an R&D portfolio optimization problem. We find that the two methods give different optimal portfolios in 4 of the 13 cases that we considered, and that the welfare loss from using Method II (assuming that Method I is better) ranges between $1.4 and $13.3 billion, in a problem on the order of $13 trillion.

4.1 Data

We use data from three sets of expert elicitations that gathered information about the probability of success of technology R&D into solar photovoltaics, carbon capture and storage (CCS), and nuclear [4][5][6]. We have applied both Method I and Method II to each of these data sets. Note that in some cases, not every expert answered every question[1]. In these cases we dropped the expert when using Method II; whereas all experts were used in Method I.[2] Also, while most of the elicitations were decomposed into hurdles that all had to be satisfied for success (i.e. the kinds of joint events we have been focusing on), one technology, post-combustion CCS, was decomposed into four events, any of which lead to success (i.e. a union).

The resulting aggregated probabilities are in Table 6. Note that the aggregated probabilities are significantly different in many of the cases. From looking at the original data, we observed that there was a tendency among the CCS and solar experts toward being optimists or pessimists; that is, they showed somewhat strong positive self-correlation. In the presence of positive self correlation, our theoretical results (summarized in Table 1) indicate that probabilities calculated with Method II would be higher on average than probabilities calculated with Method I. The results we see in Table 6 are consistent with this observation, with most Method II probabilities being higher than Method I probabilities (except in Post-combustion where the reverse is expected since it is a

[1] Specifically, the Medium funding level in Chemical Looping CCS and all funding levels (High and Low Success) in High Temperature Nuclear each had one expert who did not answer all the questions.

[2] The qualitative results are not different even if we drop incomplete experts for Method I as well, and the optimal portfolios remain the same.

union rather than an intersection). On the other hand, the nuclear experts did not exhibit clear tendencies toward self-correlation, and we see less of a pattern there.

Table 6: Aggregated Probabilities from Method I and Method II. The larger aggregated probability between the two methods is highlighted in bold font.

Fund. Level	Carbon Capture and Storage						Nuclear High Success		Nuclear Low Success		Solar High Success						Solar Low Success	
	Pre-Comb		Chem Loop		Post-Comb		HTR		HTR		Organic		In-Organic		3rd Gen		Organic	
	Mt I	Mt II	Mt I	Mt II	Mt I	Mt II	Mt I	Mt II	Mt I	Mt II	Mt I	Mt II	Mt I	Mt II	Mt I	Mt II	Mt I	Mt II
High	**0.23**	0.22	0.16	**0.42**	**0.93**	0.79	0.25	**0.30**	**0.17**	0.10								
Med	0.11	0.11	0.14	**0.30**	**0.86**	0.70	0.14	**0.17**	**0.11**	0.09	0.03	**0.04**	0.29	**0.43**			0.13	**0.25**
Low	**0.05**	0.03	0.02	**0.08**	**0.68**	0.59	**0.01**	0.00	**0.02**	0.01			0.15	**0.27**	**0.09**	0.02	0.09	**0.13**

4.2 Portfolio

We use both sets of probabilities to run a climate change energy technology R&D portfolio model, originally described in Baker and Peng [7]. The model is implemented as a stochastic Mixed Integer Program, with the key decision variables measuring whether to invest in a particular technology at a particular funding level. There is also a second stage decision about how much to reduce the greenhouse gas emissions that cause climate change. The objective is to minimize the cost of reducing the emissions *plus* the cost of damages from climate change, subject to a budget level for R&D. The new technologies, when they are successful, reduce the cost of reducing greenhouse gas emissions. We ran this model for 13 budget levels (measured in terms of Net Present Value), ranging from $200 million to $20,000 million. We found that the optimal portfolio stayed the same for both sets of probabilities for 9 of the 13 budget levels.

Table 7 shows how the optimal portfolio changed in the four cases. It shows how much more or less is invested in a particular technology, when Method II is used compared to Method I. So, for example, we see that at a budget of $200 million, Method II will lead to an investment of $39 million less into pre-combustion CCS than Method I. (We note that the change in the portfolios do not sum to zero, as the total budget spent is not always equal to the budget available in this kind of problem.) In the last column, we also show the increase in total societal cost due to using Method II. This evaluates the difference in total societal cost for the two portfolios, using method I probabilities to evaluate. The bottom row shows the maximum possible investment in each technology, for comparison purposes. (Note that the investment in nuclear was always the

Table 7: Change in Optimal Portfolio [Method II - Method I].

CHANGE IN OPTIMAL PORTFOLIO [METHOD II - METHOD I]									
	CCS				SOLAR				Increase In Total Societal Cost
Budget (mill $)	Pre Comb	Chem Loop	Post Comb	Total Change	Organic	In Organic	3rd Gen	Total Change	
200	-39			-39		38		38	1,400
600	39	37	-172	-96		77		77	13,300
4000					714		-386	328	1,900
10000					714		-386	328	1,900
Max Investment	386	56	519	961	830	77	386	907	

same under the two methods.) This table shows that Method II (combining later) seems to favor solar at the expense of CCS. Method I leads to higher investments in CCS and lower investments in solar than does Method II. If we assume that Method I is better, in the sense of having lower systematic errors and lower variance, then the lost value from using Method II ranges from $1.4 billion (out of total cost of $13.4 trillion) for a $200 million budget, to $13.3 billion (out of total cost of $13.1 trillion) for a budget of $600 million. On the one hand, these are very large numbers (due to the scale of the climate change problem). On the other hand, the percentage loss is small, ranging between 0.1% to 0.01%.

This example shows that the choice of method may not always be important – as is the case with Nuclear, and in 9 out of 13 budget levels; but that it can sometimes make a significant difference in the optimal choice.

5 CONCLUSION

We used a very simple additive model and focussed on the probability of joint events and combining experts through linear averaging. We show that combining experts before recomposing has a smaller expected error and smaller variance than recomposing first and then combining experts. Our simulation shows that the difference between the two methods may not be very large on average, and that combining after recomposing may have smaller errors, due to randomness, a reasonably large amount of the time. This implies that (1) it may be very hard to test this finding experimentally; and (2) if there are other factors that lead an analyst to prefer combining later, these factors may very well out weigh the expected errors. On the other hand, our empirical

example shows that the two methods can have significantly different results and lead to different optimal choices. Given this, and that all other things being equal averaging early has smaller errors, we suggest that averaging early should be adopted unless an analyst has strong reasons to do otherwise.

A APPENDIX

A.1 Method I

Consider the expected value of the estimated probability of the intersection for Method I found in equations (4) and (5):

$$E\left[q_{ii'}^{(n)}\right] = E\left[\left(p_i + \varepsilon_i^{(n)} + \mu_i^{(n)} + \delta_i^{(n)}\right)\left(p_{i'} + \varepsilon_{i'}^{(n)} + \mu_{i'}^{(n)} + \delta_{i'}^{(n)}\right)\right] \tag{20}$$

When we expand this, a number of the cross terms will be equal to zero since they have zero mean and are independent. We are left with

$$E\left[q_{ii'}^{(n)}\right] = p_i p_{i'} + E\left[\mu_i^{(n)}\mu_{i'}^{(n)}\right] \tag{21}$$

Expanding the error term we get

$$E\left[\mu_i^{(n)}\mu_{i'}^{(n)}\right] = E\left[\frac{1}{n}\sum_{j=1}^{n}\mu_{ij}\frac{1}{n}\sum_{j=1}^{n}\mu_{i'j}\right] = E\left[\frac{1}{n^2}\left(\mu_{i1}\mu_{i'1} + \mu_{i1}\mu_{i'2} + ... + \mu_{in}\mu_{i'n}\right)\right] \tag{22}$$

Again, all the cross terms in which $j \neq j'$ have expected value zero, so this reduces to:

$$E\left[\mu_i^{(n)}\mu_{i'}^{(n)}\right] = E\left[\frac{1}{n^2}\sum_{j=1}^{n}\mu_{ij}\mu_{i'j}\right] \tag{23}$$

Since each μ_{ij} has mean zero, the correlation as defined in the text in Section 2.1 is (by definition):

$$\rho = \frac{E\left[\mu_{ij}\mu_{i'j}\right]}{\sigma^2} \tag{24}$$

Thus the expected error can be simplified to

$$E\left[\mu_i^{(n)}\mu_{i'}^{(n)}\right] = \frac{\rho\sigma^2}{n} \tag{25}$$

A.2 Method II

Similarly, from the estimated probability of the intersection for Method II found in equation (7), we expand to get

$$\tilde{q}_{ii'}^{(n)} \equiv \frac{1}{n}\sum_{j=1}^{n}\left(p_i + \varepsilon_{ij} + \mu_{ij} + \delta_{ij}\right)\left(p_{i'} + \varepsilon_{i'j} + \mu_{i'j} + \delta_{i'j}\right) \tag{26}$$

When we take the expected value, again many of the cross terms fall out leaving:

$$\begin{aligned} E\left[\tilde{q}_{ii'}^{(n)}\right] &= p_i p_{i'} + E\left[\frac{1}{n}\sum_{j=1}^{n}\mu_{ij}\mu_{i'j}\right] && (27)\\ &= p_i p_{i'} + \rho\sigma^2 && (28)\end{aligned}$$

A.3 Variance of Errors

For Method I:

$$\begin{aligned} E\left[\mu_{ii'}^2\right] &= E\left[\left(\frac{1}{n}\sum_{j=1}^{n}\mu_{ij}\left\{\frac{1}{n}\sum_{j=1}^{n}\left(\rho\mu_{ij} + \sqrt{1-\rho^2}x_{i'j}\right)\right\}\right)^2\right] && (29)\\ &= \frac{1}{n^4}E\left[\rho^2\left(\sum_{j=1}^{n}\mu_{ij}\right)^4 + \left(1-\rho^2\right)\left(\sum_{j=1}^{n}\mu_{ij}\right)^2\left(\sum_{j=1}^{n}x_{i'j}\right)^2\right] && (30)\end{aligned}$$

since x and μ are independent.

$$= \frac{1}{n^4}\left[\rho^2 E\left[\left(\sum_{j=1}^{n}\mu_{ij}\right)^4\right] + \left(1-\rho^2\right)E\left[\left(\sum_{j=1}^{n}\mu_{ij}\right)^2\right]E\left[\left(\sum_{j=1}^{n}x_{i'j}\right)^2\right]\right] \tag{31}$$

Note that all the terms within each expectation in which $j \neq j'$ have expectation of zero. So

$$E\left[\left(\sum_{j=1}^{n}\mu_{ij}\right)^2\right] = E\left[\left(\sum_{j=1}^{n}x_{i'j}\right)^2\right] = E\left[\sum_{j=1}^{n}\mu_{ij}^2\right] = n\sigma^2 \tag{32}$$

In order to evaluate the first term in (31), we assume that each μ_{ij} is normally distributed. This implies that $\sum_{j=1}^{n}\mu_{ij}$ is normally distributed with variance $n\sigma^2$. The fourth moment of this random variable is $3n^2\sigma^4$. Adding the two terms in (31) we get equation (11) in Section 2 above.

A.4 Correlation Across Experts

When considering the correlation across experts using Method I, the variance is just the second moment: $E\left[\varepsilon_i^{(n)2}\varepsilon_{i'}^{(n)2}\right] = E\left[\varepsilon_i^{(n)2}\right]E\left[\varepsilon_{i'}^{(n)2}\right]$,where

$$\begin{aligned} E\left[\varepsilon_i^{(n)2}\right] &= E\left[\left(\frac{1}{n}\sum_{j=1}^{n}\varepsilon_{ij}\right)^2\right] && (33)\\ &= \frac{1}{n^2}E\left[\sum_{j=1}^{n}\varepsilon_{jj}^2 + \sum_{j,j'\neq j}^{n}\varepsilon_{ij}\varepsilon_{ij'}\right] && (34)\\ &= \frac{1}{n^2}\left(n\sigma_\varepsilon^2 + n\left(n-1\right)\rho_\varepsilon\sigma_\varepsilon^2\right) = \frac{1}{n}\left(1+\left(n-1\right)\rho_\varepsilon\right)\sigma_\varepsilon^2 && (35)\end{aligned}$$

And so the variance of the error term is

$$\frac{1}{n^2}\left(1+\left(n-1\right)\rho_\varepsilon\right)^2\sigma_\varepsilon^4 \tag{36}$$

For Method II we need to calculate the variance of $\frac{1}{n}\sum_j^n \varepsilon_{ij}\varepsilon_{i'j}$ (which is again equal to the second moment).

$$\begin{aligned} E\left[\frac{1}{n^n}\left(\sum_{j}^{n}\varepsilon_{ij}\varepsilon_{i'j}\right)^2\right] &= \frac{1}{n^2}\left(n\sigma_\varepsilon^4 + n\left(n-1\right)\rho_\varepsilon^2\sigma_\varepsilon^4\right) && (37)\\ &= \frac{1}{n}\left(1+\left(n-1\right)\rho_\varepsilon^2\right)\sigma_\varepsilon^4 && (38)\end{aligned}$$

A.5 Union of events

Considering the union of two independent events $p_1$and p_2. The probability of the union of the events, defined as $\tilde{p}_{ii'}$, is

$$\tilde{p}_{ii'} = p_i + p_{i'} - p_i p_{i'} \tag{39}$$

The estimated probability is now

$$\tilde{q}_{ii'} = q_i + q_{i'} - q_i q_{i'} \tag{40}$$

For Method I, evaluating the expected value gives

$$E[\tilde{q}_{ii'}] = E[q_i] + E[q_{i'}] - E[q_i q_{i'}] = p_i + p_{i'} - p_i p_{i'} - \frac{\rho\sigma^2}{n} \tag{41}$$

From (39), this results in

$$E[\tilde{q}_{ii'}] = \tilde{p}_{ii'} - \frac{\rho\sigma^2}{n} \tag{42}$$

For Method II, this results in

$$E[\tilde{q}_{ii'}] = \tilde{p}_{ii'} - \rho\sigma^2 \tag{43}$$

Hence, the resulting error from the union of events is just opposite in sign to that from the intersection of events (6).

References

[1] Sigrun Andradottir and Vicki M. Bier. Choosing the number of conditioning events in judgemental forecasting. *Journal of Forecasting*, 16:255–286, 1997.

[2] Sigrun Andradottir and Vicki M. Bier. An analysis of decomposition for subjective estimation in decision analysis. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 28(4):443–453, 1998.

[3] M. Naceur Azaiez and Vicki M. Bier. Perfect aggregation for a class of general reliability models with bayesian updating. *Applied Mathematics and Computation*, 73:281–302, 1995.

[4] Erin Baker, Haewon Chon, and Jeffrey Keisler. Advanced Nuclear Power: Combining economic analysis with expert elicitations to inform climate policy. Technical report, 2008. http://dx.doi.org/10.2139/ssrn.1407048.

[5] Erin Baker, Haewon Chon, and Jeffrey Keisler. Advanced Solar R&D: Applying expert elicitations to inform climate policy. *Energy Economics*, 31:37–49, 2009.

[6] Erin Baker, Haewon Chon, and Jeffrey Keisler. Carbon Capture and Storage: Combining expert elicitations to inform climate policy. *Climatic Change*, 96(3):379, 2009.

[7] Erin Baker and Yiming Peng. The value of better information on technology R&D programs response to climate change. *Environmental Modeling and Assessment*, 17:107(15), 2010.

[8] Robert F. Bordley. Combining the opinions of experts who partition events differently. *Decision Analysis*, 6:38–46, 2009.

[9] Robert T. Clemen. Combining overlapping information. *Management Science*, 33:373–380, 1987.

[10] R.T. Clemen and R.L. Winkler. Combining probability distributions from experts in risk analysis. *Risk Analysis*, 19:187–203, 1999.

[11] Michael P. Clements and David I. Harvey. Combining probability forecasts. *International Journal of Forecasting*, 27:208–223, 2009.

[12] Roger M. Cooke. Problems with empirical bayes. *Risk Analysis*, 6:269–74, 1986.

[13] Roger M. Cooke, Susie Elsaadany, and Xinzheng Huang. On the performance of social network and likelihood based expert weighting schemes. *Reliability Engineering and System Safety*, 93:745–756, 2008.

[14] Roger M. Cooke and Louis L.H.J. Goosens. Tu delft expert judgment data base. *Reliability Engineering and System Safety*, 93:657–674, 2008.

[15] Simon French. Updating of belief in the light of someone else's opinion. *Journal of the Royal Statistical Society. Series A (General)*, 143(1):43–48, 1980.

[16] Simon French. Consensus of opinion. *European Journal of Operations Research*, 7:332–340, 1981.

[17] S. C. Hora, N. G. Dodd, and J. A. Hora. The use of decomposition in probability assessments of continuous variables. *Journal of behavioral decision making*, 6:133–147, 1993.

[18] Michael Harrison J. Independence and calibration in decision analysis. *Management Science*, 24(3):320–328, 1977.

[19] Mohamed N. Jouini and Robert T. Clemen. Copula models for aggregating expert opinions. *Operations Research*, 44(3):444–457, 1996.

[20] David W. Keith. When is it appropriate to combine expert judgments. *Climatic Change*, 33:139–143, 1996.

[21] Dennis Lindley. Reconciliation of probability distributions. *Operations Research*, 31(5):866–880, 1983.

[22] Dennis Lindley. Reconciliation of discrete probability distributions. *Bayesian Statistics 2*, 2:375–390, 1985.

[23] J. R. W. Merrick, J. R. van Dorp, and A. Singh. Analysis of correlated expert judgments from pairwise comparisons. *Decision Analysis*, 2(1):17–29, 2005.

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

[25] A Mosleh, V.M Bier, and G Apostolakis. Methods for the elicitation and use of expert opinion in risk assessment: Phase 1, a critical evaluation and directions for future research. Technical report, 1987.

[26] Ali Mosleh and Vicki Bier. On Decomposition and Aggregation Error in Estimation: Some basic principles and examples. *Risk Analysis*, 12(2):203–214, 1991.

[27] J. B. Predd, D. Osherson, S. R. Kulkarni, and H. V. Poor. Aggregating forecasts of chance from incoherent and abstaining experts. *Decision Analysis*, 5(4):177–189, 2008.

[28] H. Ravinder, D. Kleinmuntz, and J. Dyer. Reliability of subjective probabilities obtained through decomposition. *Journal of Behavioral Decision Making*, 34:186–199, 1986.

[29] Shi-Woei and Vicki M. Bier. A study of expert overconfidence. *Reliability Engineering and System Safety*, 93(5):711–721, 2008.

[30] Shi-Woei and Chih-Hsing Cheng. The reliability of aggregated probability judgments obtained through cooke's classical model. *Journal of Modelling in Management*, 4(2):149–161, 2009.

[31] Jouni T. Tuomisto, Andrew Wilson, John S. Evans, and Marko Tainio. Uncertainity in mortality response to airborne fine particulate matter: Combining european air pollution experts. *Reliability Engineering and System Safety*, 93(5):732–744, 2008.

[32] Thomas S. Wallsten and David V. Budescu. Encoding subjective probabilities: A psychological and psychometric review. *Management Science*, 29(2):151–173, 1983.

[33] Robert Winkler. Combining probability distributions from dependent information sources. *Management Science*, 27(4):479–488, 1981.

[34] Bram Wisse, Tim Bedford, and John Quigley. Expert judgement combination using moment methods. *Reliability Engineering and System Safety*, 93(5):675–686, 2008.