TO: Michael Alvarez, Editor
RE: PA Revision Memo
DT: October 13, 2011
FR: Michael D. Ward
Jacob Montgomery
Florian Hollenbach

In response to the reviews we received as well as your guidance, we have substantially revised our manuscript on Ensemble Bayesian Model Averaging. The reviewers were each helpful in pointing out where and how the manuscript could be strengthened, and we are grateful that they took the time to give us such thoughtful feedback. We believe that the final paper will be much stronger thanks to their efforts.

This revision has involved two major threads – reanalysis and rewriting. We wanted to be clear as to what changes we have made in each thread.

First, let us point out the *reanalyses* that we undertook for this revision.

- In order to make the exposition more straightforward and to more clearly advocate for a principled approach to forecasting, we redesigned the estimation process in our first example so that there are three sets of data (this same tri-partite partition was already implicit in our other examples, but we made that fact much clearer in the discussion as well). The first is a training data set on which model parameters for component forecasts are determined. The second is a validation set on which the weights for the EBMA model are calibrated and the third consists of data on which the component and EBMA predictions are compared and evaluated. This is more straightforward, is easier to explain, and more directly addresses concerns about over-fitting since model weights are themselves based on out-of-sample forecast accuracy.

- The GLM model for the insurgency example was criticized for being too simplistic, and as a result of this comment we made the model a bit more complex via additional covariates.

- The mixed effects model for the insurgency example was completely redone as well, in part owing to some convergence issues that emerged in the tripartite data division.

- For our insurgency example, we also calculated a simple unweighted average of point predictions to respond to R2 and show that EBMA dominates such an approach.

- In total, all of the component models, the EBMA forecasts, and all tables and figures associated with the first empirical application have been revised.

Second, we revised the presentation of the paper substantially.

- A number of suggestions were made to reduce the space allotted to the mathematics in Section 3 and to instead give greater attention to explaining the intuition of the EBMA method. Specific suggestions were to move much of this content to an appendix (R1 and R2), which we did. In doing so, we changed the notation and eliminated the triple subscript, following the suggestion of R3. More importantly, we revised Section 3 extensively to provide a more compelling explanation of the approach, backed up with a coherent appendix. Throughout Section 4, we also provide more discussion and interpretation of the results and offer practical direction for applied analysts.

- We significantly expand the framing and discussion of the Supreme Court example by way of illustrating the benefits of EBMA in combined forecasts that may be produced by subject-matter experts and other non-statistical models (e.g., classification trees). Subject-matter experts are widely used in the intelligence and warning communities, as well as the political/business risk communities. As such, the benefit of the approach is targeted to a wider swath of potentially interested analysts.

- In general, we worked very hard at rewriting the paper to make it more transparent and plausibly useful to those analysts who might actually be interested in adopting this approach.

In addition to these broad changes, we addressed each reviewer's specific concerns as follows.

R1
- R1 wanted an appendix, and a clear elucidation of the basic intuition (as well as the details of the approach). We now have the appendix. In addition, we have tried throughout to provide a better intuition and theoretical motivation for the EBMA method.
- We also cleaned up and sharpened the didactic discussion preceding the mathematical details (see especially Section 3.1).
- We have also introduced references to other ensemble work in computation modeling and machine learning (pp. 3-4 and Footnote 26).
- We've also addressed some of the basic practical discussions which R1 noted were missing, such as how to partition the data most effectively (Section 3.1, Footnote 4, and pg. 11). Unfortunately, there are not always definitive answers to some of these questions, but we do now raise and discuss, rather than ignore, the issues R1 brought to our attention.
- We discuss the issue of over-fitting more directly and point out that we are using a type of penalized regression since model weights are fit based on out-of-sample performance. We hope that the discussion of the three different data bins – training, validation, test – makes it easier to understand how EBMA deals with over-fitting issues (Section 3.1, Footnote 4, and pg. 11).
- R1 was concerned about the overly simplistic nature of the GLM model. We've made our models a little more complex, and have pointed out that other machine learning methods are also available (pp 3-4). But we haven't tried to make this into a general machine learning comparison. That literature is well developed and quite extensive. A different project might usefully compare the wide array of learning techniques to EBMA, but we decided that this was not the place for those comparisons. We also emphasize here that the SAE model is a simple GLM with 27 covariates. Although we would not want to call this model a "garbage can" or "kitchen sink" model, we feel that this addresses some of the fair competition concerns R1 raises.
- R1 questioned whether we needed three examples to make our point. We've rewritten the paper to make it more transparent why the three applications demonstrate different aspects that will be interesting to social scientists. The examples illustrate the technique in three different, interesting cases: dichotomous variables, continuous measurements, and the inclusion of non-statistical forecasts produced by subject-matter experts (c.f., pp 9-10 and Section 4.3).
- Concerning the correlation among forecasts, we have ramped up our discussion about what happens to the weights in the situation of highly correlated predictions throughout the paper when model weights are explained and discussed (see also pg. 19).
- R1 was concerned with the organization of the paper, which has now been reorganized more or less along the lines that R1 suggested. There is an appendix which captures much of the material in the former Section 3. We've also expanded the discussion of the meaning of the weights more extensively (e.g., pg. 6, pp 8-9, Section 4.2). We have also removed most of the technical concerns (e.g., convergence and the EM algorithm) to the appendix and focused the paper more sharply throughout on explaining the method clearly to readers.

R2
- We've followed the suggestion to explain the training, validation, and testing periods (e.g., Section 3.1, Footnote 4, and pg. 11). We also added Section 3.1 to provide concrete intuition early in the paper, and tried to more clearly relate the mathematics in Section 3.2 to a specific applied forecasting task (i.e., predicting insurgency in a set of countries).
- We have also followed the advice to provide a better explanation of the weighting and how it relates to model performance (e.g., pg. 6, pg. 8-9, Section 4.2), and how the EBMA parameters are estimated with separate (validation) data (e.g., footnote 4). In short, we think the three different data bins in example 1 and our clearer explanation of how the tri-partite division of the data works in examples 2 and 3 now address several of the suggestions R2 made.
- We discuss how EBMA penalizes over-fitting (pg. 6 and Footnote 4) and try to illustrate how that works (Section 4.1). We also show how much better EBMA is than the simple average, both conceptually and with an example (Footnote 13).

– As suggested, we moved much of the former Section 3 to an appendix and eliminated the former Equation 11. This also necessitates that we give less attention to the EM algorithm more generally, and instead offer intuition about the origin of model weights in the new Section 3.1. We also discuss the uncertainty issue associated with Bayesian estimation, but we are leaving a fully Bayesian implementation to future research.

R3 – As discussed above, we now explain how the w_k weights are determined and provide extended discussions when results are estimated (e.g., pg. 6, pp 8-9, Section 4.2). The origin of weights is also implicit in the extended explanation of the training, validation, and testing periods.

– We adopted the sub-scripting suggestion in the body of the text as well as the Appendix.

– We have a better discussion of the training, validation, and prediction sets and discuss a bit how to structure these sets in terms of the ratio of the signal to the noise in them (e.g., pg. 11).

– We try to unpack the identified sentence when discussing parameter estimation for a binary-outcome model and explain the intuition for the penalty term b in the Appendix.

– We have de-emphasized the RMSE or MAE criteria and follow Tilmann Gneiting's advice about the proper score created by the Brier index (pp 14-15 and Footnote 22).

– We have included a type of sharpness plot in Figure 3.

– We expanded the SCOTUS example rather than eliminating it because of the benefit it offers in terms of illustrating the use of EBMA for non-model forecasts. We think the new presentation better explains why the example is included (c.f., pp 9-10 and Section 4.3).