Submitted to the Annals of Applied Statistics

PROBABILITY AGGREGATION IN TIME-SERIES: DYNAMIC HIERARCHICAL MODELING OF SPARSE EXPERT BELIEFS

(REVISION OF MANUSCRIPT AOAS1309-039)

POINT-BY-POINT RESPONSE

BY VILLE A. SATOPÄÄ, SHANE T. JENSEN, BARBARA A. MELLERS, PHILIP E. TETLOCK, AND LYLE H. UNGAR

Department of Statistics, The Wharton School of the University of Pennsylvania
E-mail: satopaa@wharton.upenn.edu*;* stjensen@wharton.upenn.edu
Department of Psychology, University of Pennsylvania
E-mail: mellers@wharton.upenn.edu*;* tetlock@wharton.upenn.edu
Department of Computer and Information Science, University of Pennsylvania
E-mail: ungar@cis.upenn.edu

We thank both the editor and reviewers for their helpful comments on our manuscript. We have addressed each of these comments in our revision, as well as providing a point-by-point response below. The comments are in boldface. Our comments are in normal typeface.

Editor. **Concerning the out-of-sampling applicability of your methods, the cross-validation technique used in Section 6 indeed does not yield time-forward predictions, as would be required in practice. If you could modify your technique so that it can be used for time-forward predictions, that would be most useful. At a minimum, you should be very clear about the limitations of your approach, and you ought to discuss in detail how these issues could be remedied in future work.**
We have added a brief paragraph on time-forward propagation. In practice, however, social scientists tend to be more interested in the single "best" estimate of the probability given all the information that is available at the moment. This is what our model addresses.

In my own reading, I also felt that your paper could be usefully shortened. For example, Figures 1 and 5 are redundant - so please remove Figure 1, and move Figure 5 forward to Section 1. Also, please move the Appendix to an Online Supplement. Throughout, please check for redundancies and remove them, unless the exposition requires otherwise.
The appendix has been moved to an Online Supplement. Figure 1 has been replaced with Figure 5. Some sections, including the discussion on Platt calibration and summary of findings, have been removed. We also have removed many redundancies. The synthetic data section has also been shortened and now discusses only the relevant results therein.

As regards Referee 3's second specific comment, I agree that the definitions on page 6 are unlikely to be appreciated, unless you introduce the (rather technical) prediction space setting that is implicit in Ranjan and Gneiting (2010) and explicit in Gneiting and Ranjan ("Combining predictive distributions", Electronic Journal of Statistics 7, 1747-1782, 2013). Therefore, let me suggest that you discuss these definitions informally (and within the text, rather than using displayed environments) and refer to the aforementioned references for technical details.
The definitions are now only discussed in the text with a reference to the aforementioned references for technical details.

In the references list, please unify the style used therein, capitalize all journal titles, expand acronyms such as AAAI, ECAI, and UAI, provide an update for Mellers et al. (2013), add the missing page ranges for Metropolis et al. (1953) and Niculescu-Mizil and Caruana (2005), the missing publisher for Pepe (2003), the missing authors, editors, and booktitle for Platt et al. (1999), and the multiply missing information for Wilson and Wilson (1994)

and Wright et al. (1994).

- All acronyms have been expanded. We have aimed to make the style more unified. All journal names have been capitalized.
- Mellers et al. (2013): This paper is currently under revision after receiving a friendly reject & revise.
- Metropolis et al. (1953): The missing page ranges have been added.
- Niculescu-Mizil and Caruana (2005): No more referenced in the paper.
- Pepe (2003): Thank you for noticing this. It had been listed as an article in the .bibtex file and hence skipped the publisher info.
- Platt et al. (1999): No more referenced in our paper.
- Wilson and Wilson (1994): This has been corrected. As this is a book, only the publisher was added.
- Wright et al. (1994): This reference seems to be complete.

Referee 1. **This is an interesting paper describing a model for probabilistic forecast aggregation over time. The model is novel and has the ability to handle sparse, irregular data. I have a few major comments and many smaller comments.**

Motivation: I think that more could be done to relate the proposed model to existing aggregation models. As the authors state, the usual time series models do not apply to these data. However, the proposed model can be viewed as an extension of a few psychometric models. In particular, the model presented in Section 3 is a hierarchical factor analysis model with the traditional roles of subjects and items reversed. Batchelder & Romney (1988) developed a non-time-series version of this model to aggregate judgments on problems where no outcome Z exists (under the name "Cultural Consensus Theory"). Karabatsos & Batchelder (2003) further developed Bayesian versions of this model. Thus, if we stop after Section 3, the proposed model is an extension of cultural consensus models to time series.

Thank you for pointing us to these papers. We were not aware of this work and it has definitely been an interesting read. There are so many static aggregation papers in the machine learning literature that missing some papers is unavoidable. To our understanding, the model described in Karabatsos & Batchelder (2003) is applied to binary responses to estimate the solution key (which is also considered binary). Therefore the setup is slightly different from the probability aggregation paradigm considered in our paper. However, a more recent paper by Batchelder et al. called *Cultural consensus theory: aggregating continuous responses in a finite interval* aims to use CCT to perform probability aggregation. We have cited this work in our updated introduction. This work, however, also considers static aggregation and is therefore very different in nature to our work. For this reason, and also for the sake of shortening the already too long paper, we have decided to not include an extensive literature review of the CCT.

Section 4 highlights the rotation issues inherent in factor analysis models. The authors fit the model under one constraint, then rotate the parameters to an equivalent set of parameters that minimize a scoring rule. So, following rotation, this model is no longer a strict consensus model because the outcomes Z impact the rotation.

In general, I believe the relationship to factor analysis and cultural consensus provides a nice motivation for the model and its rotation issues (i.e., the calibration step).

Staying with the rotation issue, there are alternative constraints one may adopt instead of fixing one element of b to 1. For example, I believe the model can be identified by restricting the sum of the elements of b to equal one (and possibly restricting the variance to 1). This restriction may avoid the issue described on p 12, first full paragraph.
It is enough to constrain the sum to gain identifiability, i.e. no additional constraint on the variance is needed. Constraining the mean of the elements of b to equal to one (or equivalently, their sum to equal to five) was considered at early stages of our implementation. Unfortunately, the problem discussed in synthetic data section is more due to the calibration step than the choice of restriction. In essence the problem occurs because the calibration cannot move any prediction from one side of zero to the other. To see how this happens even under the sum restriction, notice that by enforcing the bias terms to sum to, say five, implies that the groups are, on average, unbiased during the sampling step. If, however, the experts are on average very under-confident, then it is likely that the resulting estimates of the hidden states are on the wrong side of zero. This is again a problem that our one-dimensional calibration cannot fix.

This is a complex model, and there were a few issues that could be described in more detail. For example, I did not see details on how the parameters μ_0 and σ_0^2 were set. This seems to be a place where one might artificially improve model performance.
Thank you for pointing this out. In our studies we always set μ_0 and σ_0^2 to 0 and 1, respectively. This is now mentioned in the text. Based on our experience, as long as these values are reasonable, they do not influence the results significantly. These values are mere baseline initializations that are rapidly washed out by the data.

Additionally, there could be further discussion about the model's handling of data sparsity. Currently, the only details appear in the appendix (stating that an update is made for each observation at time t), even though data sparsity is offered as a motivation for this model.
We have added " The first step of our Gibbs sampler is to sample the hidden states via the *Forward-Filtering-Backward-Sampling* (FFBS) algorithm. FFBS first predicts the hidden states using a Kalman filter and then performs a backward sampling procedure that treats these predicted states as additional observations (see, e.g., Carter and Kohn (1994); Migon et al. (2005) for details on FFBS). Given that the Kalman filter can handle varying numbers or even no forecasts at different time points, it plays a very crucial role in our probability aggregation under sparse data."
Without describing the technical details of the Kalman Filter in the main text, it is difficult to explain how sparse data is handled. These details, however, are still kept

in the Online Supplement.

Implementation: I wonder how the models were implemented, along with the speed of the model estimation. I guess the models may take some time to run. It would be nice if model estimation code were shared.
We have added: "Our implementation of the sampling step is written in C++ and runs fast. For instance, to obtain a posterior sample of size 1,000 for 50 questions each with 100 time points and 50 experts takes about 215 seconds on the first author's computer (1.7 GHz Intel Core i5). " The calibration step is currently run separately in R once the samples have been obtained from the sampling step. Given that the calibration requires a univariate optimization, it can be performed very fast. We plan to post our implementation upon publication.

$BSAC_{log}$: I don't understand why we should expect this estimation method to be any different from the SAC_{log} method, especially because flat priors are used. In the Table 3 and Figure 3 results, it looks like there are some minor differences but not enough to be notable. Hence, I am not convinced that the discussion of $BSAC_{log}$'s overconfidence is meaningful.
The main difference between $BSAC_{log}$ and SAC_{log} is that $BSAC_{log}$ estimates the hidden states $\{X_{t,k}(1)\}$ and β simultaneously (alternating between the two until convergence), whereas SAC_{log} first estimates $\{X_{t,k}(1)\}$ and then calibrates them by finding β. Therefore $BSAC_{log}$ is slightly more flexible than SAC_{log}. Unfortunately, it tends to overfit the data. This results into the overconfidence that it seen in our analyses. We have added "Given that this model estimates $X_{t,k}(1)$ and β simultaneously, it is a little more flexible than SAC."

Finally, there were a variety of small comments and questions:

Top p 6: I do not understand the phrase "the entire b vector is shared". Does this mean that b has a single element?
This means that each of the elements of b are considered the same across the 166 questions. We have modified the text to make this more clear.

Bottom p 9, expressing logistic regression as linear regression: I don't understand the normal error in this equation. I would think that, if this has some equivalence to logistic regression, the error term should follow the logistic distribution.
The error term was not needed. This entire paragraph has been removed for the sake of brevity.

p. 11, third line: I believe should read "$166 - N_{train}$".
This has been corrected. Thanks for catching it!

p. 11: Please define the EWMA method before describing results associated with it.
The definition of EWMA has been moved.

Figure 2: Would be nice to have common y-axis limits for each row of the graph.
This is a good idea. For the sake of brevity, however, we have ended up removing most of these plots. We only kept the one that tells something non-obvious about our model, and then briefly mention the other results within the text. After all this section is a mere sanity check.

Appendix, p 26, top: I was initially confused because, e.g., e is missing an i subscript even though i appears in the equation. I guess this is related to the text immediately below, stating that a new update is repeated for each observation. But perhaps the notation could be clearer.
The Appendix has been moved to Online Supplement, where the update step is not described in an algorithmic way and with improved notation. We believe that it is clearer now.

Referee 2. **This paper attempts a difficult problem of combining sparse expert probabilities which evolve over time. Three characteristics of the approach are significant**

- **Experts are allowed to give multiple forecasts over time till the geopolitical event materializes**
- **Experts are can update forecasts at irregular intervals of their liking during the same period**
- **Experts are grouped into multiple categories based on their level of self-acclaimed expertise**

The extant literature doesnt address the issue of dynamic aggregation over time. The combination methods developed in the literature wont be able to do justice to evolving expert opinions over time as they will treat them as one combination problem at a given point of time. Also, the small no. of expert forecasts at any given time would make the combined forecast at a given time weak. The alternative is to borrow strength from past forecasts in a sensible way which is explored in the present paper.

The other contribution of the paper is study various groups of experts based on their self-assessed expertise. The parameters used in the framework directly lend themselves to the study of group specific over-confidence/under-confidence, disagreement between experts, question difficulty and evolving expert opinions. The combined forecast is shown to be empirically calibrated and sharp. The estimation involves two stages a sample step and a calibrate step. In the sample step the model parameters are sampled using a Gibbs sampling paradigm. In the calibrate step a one parameter optimization is used to calibrate the experts forecasts with the truth using a proper scoring rule.

The methodology presented in the paper for combination has been compared against simple exponentially weighted generalizations of existing methods like equally weighted probability forecasts and the beta-transformed probability forecast method proposed by Ranjan and Gneiting (2010). However, these generalizations have been restricted to single combination parameters only. One wonders why these methods were not generalized to include group specific combination parameters. This seems to me the natural thing to do otherwise, the comparison to other methods wont be fair.

Given that EWMLA already includes group specific combination parameters, only EWMA and EWMLA need to be generalized. In the revised manuscript EWMA and EWMLA make use the expertise information by incorporating a weighted mean (instead of the simple equally weighted mean used in the previous version) of the forecasts. The results, however, look very similar.

Another limitation of the present work is that it masks question specific and individual specific bias terms and only estimates group specific bias irrespective of question and individual. This wont allow one to understand any outcome specific bias that may exist in the groups or individuals. For example certain geo-political outcomes may be more desirable for people in general and individuals might be biased for that outcome. It would be interesting to see if authors can come up with an framework in which bias can be studied at multiple levels, be it question/ outcome specific, individual specific or group of experts specific or even time-specific. The present work only looks at group specific bias. Perhaps a hierarchical approach to bias decomposition may be more suitable.
We have added a sub-section called *Bias Structure* that discusses a hierarchical approach to bias decomposition. In this subsection we briefly explain how the bias can be estimated at question-specific and individual-specific. An extension to a time-specific bias structure is briefly mentioned in the conclusion.

Overall I am happy with the work as it explores a difficult and unexplored problem and offers some directions. The modeling of the problem exposes various parameters which have an intuitive interpretation for the researcher and helps understand the expert opinions better. However I would like to see the authors update Table-3 with group specific parameters used for other methods as well (in particular, EWMA, EWMBA, EWMLA). Also, I would like to see some discussion on how bias (under-forecasting and over-forecasting) can be estimated and studied more generally at multiple levels individuals, questions, groups and even time specific.

Referee 3. **This manuscript describes a model for aggregating probabilities judgements made by experts. It allows for the experts to update their beliefs at different points in time and it allows for different biases between different groups of experts. The general idea is very interesting, but the implementation is lacking. The model makes strong, simplistic assumptions, and the applicability in practice is questionable.**

General Comments

1. **The manuscript is generally well-written, but it is way too long. The writing is too leisurely throughout, and many issues and concepts are unnecessarily discussed two or more times.**
   The paper is now much shorter. We have moved material to the Online Supplement, removed some of the unnecessary sections and graphs, and cut many redundancies.

2. **In my opinion, the model is unnecessarily simplistic. The authors assume that experts are indistinguishable beyond (self-reported!) expertise. Bias and variance in their forecasts are assumed to be constant over experts, questions, and time. The authors fit a Bayesian hierarchical model that would allow for sophisticated shrinkage and borrowing of strength between experts and between questions, but this is basically not exploited in the model at all. In fact, the authors claim (incorrectly) that estimating a bias term separately for each expert is not possible. For example, why not specify a prior for the ith experts bias such as $\beta_i | b_{j(i)}, \sigma_2 \sim N(b_{j(i)}, \sigma^2)$, where $j(i)$ is the group membership of the ith expert?**
   Thank you for pointing this out. We have added a brief subsection that discusses question-specific and individual-specific bias structures. The choice made in this paper was partly driven by our real-world dataset and the preference of the social scientists working with it. One goal of this paper is to describe a basic framework that can be extended rather easily for any such analysis.

3. **I wonder how applicable this methodology actually is in practice. While I cannot comment on how interesting the fitted model parameters are to social scientists in terms interpreting things like question difficulty, the results to me offer little in terms of interpretation beyond what would have been clear from the outset or what can be learned from simple plots of the data.**
   What I consider the most interesting is to make predictions for future

events, which I am not sure is really possible in real-world situations using the model. Since the methodology is retrospective (i.e., you do not do on-line, filtering inference as I think would be more appropriate), it seems to be necessary to wait until a large study like the one described has been completed, with all outcomes having been observed. Then, you can fit the model. Only after this can the methodology be used to make forecasts for other, future events of interest, but it is not clear that a similar data situation could arise, and even if so, if the assumptions and parameters from your fitted model are still applicable to this new situation. For example, are the new experts comparable to the old ones? Since there seem to be large differences in question difficulties, what difficulty does a particular new question have?

The model is interesting for three different reasons:

(a) Model inference: This is inherently interesting to social scientist.

(b) Time-Forward Forecasting: Our methodology does not take long to fit in a retrospect fashion. Once the estimates have been obtained, the estimate of the final hidden state can be propagated forward. We have included a paragraph about this in the revised version of the paper.

(c) Aggregating at Current Time Point: The goal is to obtain the "best" probability estimate given all the information available now. This is the most interesting to social sciences as it reflects the current consensus about the event and its associated uncertainty. This is what our model was built to address. We evaluate this ability in our out-of-sample aggregation section.

The new experts tend to be very similar to new ones. They do not need to be the same individuals as long as they, as a group, look similar. In terms of determining the difficulty of a new problem, this is an active and very difficult research problem. One can use old questions to model the difficulty. With no information but the description of the problem, the best one can do is to assign the new problem an average difficulty level.

Specific Comments

1. **p. 2: The punctuation seems to be off in the second part of the second paragraph.**
   This has been corrected.

2. **p. 6: Can you explain why it is reasonable to assume that the forecast maximizes sharpness subject to calibration? Also, in Definition 2, it seems like p, q, and p0 are all fixed numbers on the unit interval, so**

I am not sure with respect to which random variables the expectations are meant to be taken.
$logit^{-1}(X_{t,k})$ represents the "best" forecast under all the relevant information available at time t. This is an idealized forecast that should be assumed to be well-calibrated. Furthermore, given that it is based on all the information available, it ought to be as close to 0 or 1 as possible; that is, as sharp as possible. Due to the editor's suggestions, the technical definitions have been removed.

3. **p. 7: Instead of writing drifts to inifity, I think things would be more easily understandable if you wrote drifts to positive or negative infinity."**
This has been corrected.

4. **p. 8: For completeness, you should briefly mention in the main text which parameters are assumed to be random and what their priors are. Also, even after reading the appendix, it is still not clear to me what you mean by parameter estimates. All you are obtaining with the Gibbs sampler are samples from the posterior distribution. A loss function would be needed to obtain (point) estimates of the parameters.**
We added "Therefore the parameters of the model are $\boldsymbol{b}$, σ_k^2, γ_k, and τ_k^2 for $k = 1, \ldots, K$. Their prior distributions are chosen to be non-informative, $p(\boldsymbol{b}, \sigma_k^2|\boldsymbol{X}_k) \propto \sigma_k^2$ and $p(\gamma_k, \tau_k^2|\boldsymbol{X}_k) \propto \tau_k^2$." and "These estimates are obtained by first computing a posterior sample via Gibbs sampling and then taking the average of the posterior sample."

5. **p. 10: While the simulated data are not generated exactly from your model, it does seem like the data-generating model is much closer to your model then to the model underlying EWMA.**
The underlying mechanism to generate the hidden states is very different from the one given in our dynamic linear model. The only similarity is in how the noise and bias are added to the hidden states. The noise in this case has mean zero and can be reduced by averaging. Therefore if the model can determine the bias terms accurately; that is, determine the extent to which each expertise group should be shifted for improved calibration, the model's performance ought to be very good.
Notice that this section has been updated to include an expertise-weighted version of EWMA. This procedure uses 6 parameters to calibrate the group-specific average probabilities. This is a much more flexible calibration function than the one in our SAC model. It is therefore expected that under an easy forecasting setup, such as the one in our paper, EWMA performs better

than SAC. It is not entirely clear whether this comparison is completely fair. It serves, however, only a sanity check. The more important comparison is done under the real-world data.

6. **p. 11: EWMA should be defined here, not much later in the manuscript.**
The definition of EWMA has been moved.

7. **p. 12: EWMA outperforms SAC roughly when $\beta \in [.95, 1.5]$, which in my opinion does not correspond to a very small bias. If I understand correctly, $\beta = 1$, for example, actually means that the bias varies from 0.5 to 2.**
That is correct: β is a multiplier for the vector $[1/2, 3/4, 1, 4/3, 2/1]$. When $\beta = 1$, the average bias among the experts is around 1.12. This bias is not very large.

8. **p. 14: Why is the SDLM model not hierarchical? The process model can clearly be written in a hierarchical way, and there are also parameters that (I believe) have prior distributions.**
We wanted to say that the model does not have any parameters that are shared across different questions. This has now been changed to "Given that this model does not share any parameters across different questions, estimates of the hidden process can be obtained directly for the questions in the testing set without fitting the model first on the training set."

9. **p. 14/15: For all models, how did you choose the number of iterations in the Gibbs sampler? Should the number of MCMC iterations be larger for the fully Bayesian model, since the hidden state is sampled using the Metropolis algorithm? By the way, if you assumed a probit-link instead of a logit-link, the full conditional distributions are available in closed form and no Metropolis steps are necessary.**
We have added: "Before introducing the competing models, note that all choices of thinning and burn-in made in this section are conservative and have been made based on pilot runs of the models. This was done to ensure a posterior sample that has low autocorrelation and arises from a converged chain." The number of iterations for the fully Bayesian model should be a bit larger. We first made choices for this model and then used the same choice for SAC.

References.

CARTER, C. K. and KOHN, R. (1994). On Gibbs sampling for state space models. *Biometrika* **81** 541–553.

MIGON, H. S., GAMERMAN, D., LOPES, H. F. and FERREIRA, M. A. (2005). Dynamic models. *Handbook of Statistics* **25** 553–588.

PHILADELPHIA, PA 19104- 6340, USA
E-MAIL: satopaa@wharton.upenn.edu