TO: Jacob & Flo (for now)
RE: PA Revision Memo
DT: September 8, 2011
FR: Michael D. Ward
tel: 919.660.4373
michael.d.ward@duke.edu

1 Reviewer Comments

1.1 Reviewer 1

The idea of making many predictions from many different prediction models and combining these predictions is a powerful one: for example, in formal theory Scott Page and collaborators have shown that ensembles of diverse agents solve problems more effectively than ensembles of homogeneous agents, and the development of techniques such as bootstrap averaging (bagging), boosting, or the recent work on "targeted learning" (among many other such ensemble and penalized methods) has shown the promise of the general idea that many models/predictors/agents are better than one. I am glad to see that more techniques from the statistical learning/machine learning literature are making their way into political science.

Ensemble Bayesian Model Averaging (EBMA) in particular seems potentially quite fruitful for the application areas analyzed in this article. Another benefit of focusing on prediction is that we may obsess less over the precise nature of our data generating processes: we interpret combinations (often linear) of data and coefficients which may have nicer properties than single coefficients. That said, we still tend to pose questions about theoretically interesting comparisons in political science academia more so that we compete over who can tell the future. I think there is a very close link between theory assessment and future prediction (Rubin in particular makes this case when he talks about the Bayesian take on causal inference). Yet, I do not see these authors engaging with this issue — and it probably is an important one given the preoccupations of our profession.

I think that this paper could be revised relatively easily and would be a contribution to our discipline. My overall concern about this paper as it stands is that (1) it avoids discussion of theory and (2) it is confusing as pedagogy. I already mentioned the first concern. The second concern involves a number of minor points which I list below. I suspect that if the authors dumped one (or two) of their applications, put formulas that are not returned too or explained in depth in an appendix, and spent a bit more time providing intuition, that this article would compel more readers to spend the time necessary to learn more about EBMA and try to use it in their own work.

MDW Comments:

- *We could add a great deal of information about theories of electoral politics, theories of revolutions, and theories of Supreme court voting. I guess we need to highlight these a bit more. But maybe a paragraph or two each. Flo, can you do revolutions, Jacob voting, and I'll do courts.*
- *However we need more to emphasize that prediction itself is a good theoretical heuristic, and I'll do this.*

FH Comments:

- *I will do the revolutions*
- *seems to me the reviewer may be talking more about theoretical underpinning of forecasting and predictions in general, but I'm not sure*

JMM Comments: I don't think the reviewer means theory in the sense of substantive theory, but in a more abstract "how to learn from data" way. In particular, I think he wants us to:

1. *Ramp up the discussion of Page's book and its near neighbors in the citation network (which I will do .. book already ordered).*

2. *A discussion of the relationship between theory assessment and out-of-sample forecasting. I can also do this, but if you (Mike) know any strong citations in this area to get me started that would be helpful.*

Minor points:

Teaching about a new statistical learner of this kinds probably requires more engagement with some of the central issues in statistical learning as well as with some of the seemingly throw-away points made in the paper, and perhaps less marching through formulae and application. Here are some of the questions I had:

(1) How to choose the training set? It seems like a whole set of new research design questions arises here: how long should the training series be compared to the testing series? When should it be chosen [to be at the beginning of the observed data? Or some trimmed version of the early data? Or some data in the middle?] On what basis should such choices be made? [Power? Anticipated effect size? Something else?]

MDW Comments:

- *This is a good question, and we don't know the answer to it. Jacob could you find out if there is any lit that addresses this. I propose we add this to the research program as something to be expored. Jacob add this to the discussion at the end of the paper.*

JMM:

- *Mike I have your email on this, and will incorporate this into the paper.*

(2) What about overfitting? These authors mention it in passing, but it is my sense that overfitting is a big problem for statistical/machine learners of this type. [Perhaps the use of priors can penalize the predictions an help manage this problem. Yet, it seems like if we are to begin to take the statistical learning literature serious in political science, we also need to take overfitting as a problem seriously.] Perhaps the authors can mention that they get around the overfitting problem by (a) using priors to penalize the fitting and (b) focusing on prediction.

MDW Comments:

- *This is a type of penalized regression. But our main approach is to use the prediction to winnow out over prediction.*

JMM Comments:

- *I think the reviewer is missing the point that we are using past out-of-sample predictive performance to calibrate the model weights, which significantly reduces over-fitting. I will add some language emphasizing this point into section 3. My feeling is that I should just add a "discussion" section at the end of section 3 to deal with this reviewers comments and this will be one point. However, I imagine this being a more fluffy discussion (with some sort of running example) for how the process works, and what it means, and why it is wonderful.*

(3) Fair competition: Perhaps in political science the competition for EMBA is simple glm or kitchen-sink glm or multilevel model. Elsewhere, of course, the competition are penalized models of varying kinds and other learners (simple bagging, for example, or adaptive lasso — neither of these selects among models, but, from the perspective of the examples in section 4, they are more or less equivalent to the comparison made — the comparison basically is a variable selection competition (plus, of course, the random intercepts in the multilevel models). Anyway, again, it is good to see that the simple component models do what we expect — provide not very accurate predictions in the testing data — but it is not clear that those are the models we'd necessarily require.

MDW: Don' really know what reviewer is getting at. Anybody?

FH: As I understand him, he would like us to include some more sophisticated models in the examples? I guess we could run some variable selection model in the ICEWS example and include it in the example and show that EBMA is still better? Or we could just ignore this point.

JMM: My reading is that he is confusing EBMA with BMA. The other methods he is discussing are methods that involve throwing a bunch of covariates (as opposed to models) at an outcome of interest. I will add a footnote to section 4.1 addressing (or rather re-emphasizing) this important distinction.

(4) Why are we seeing the coverage rate comparison only for one of the examples? Shouldn't all of the examples be judged using the same criteria?

MDW: Can we do this, Jacob?

FH: this is only possible for the continuous example, no?

JMM: Yes. How would we "cover" a dichotomous outcome?

(5) Do we really need three applications? Do they serve some important pedagogical purpose?

they show that it works in different contexts, relevant to the political science prediction realm. Add sentence or two in intro to justify three. Come back in conclusion to the fact that it works in three domains as well.

-Jmm will do.

(6) Correlation among forecasts is a bigger deal too (and also one that plagues the regularized regression field [i.e. lasso and its ilk]. Again, the authors just mention this issue, but it is a fairly big issue — as the number of models increases and the number of observations remains the same, I speculate that the median pairwise correlation between models will increase. [I say this based on papers in the regularized regression literature showing that as sample size remains fixed but the number of variables increases, a similar relationship holds — and this relationship causes trouble when people try to use the lasso to do variable and model selection.]

MW: Don't know what to do with this comment. Ideas?

FH: I'm not sure his argument actually matters for the forecasting example, since the number of models is unlikely to be large enough to lead to this effect? but we should probably add one or two sentences to the part about correlation between the models.

JMM: We can ramp up the discussion for this in section 4.2. Highly correlated predictions will each individually receive a lower weight, but their joint weight may not necessarily be affected (although I don't know for sure). Quite frankly, it is difficult to address since all he says is that it may cause "trouble." But I could add additional language so we can at least have something to say in our response memo.

(7) Paper organization: I felt like much of section 3 consisted of announcements rather than pedagogy. I suspect that the authors were trying to provide the most important formulas to the readers, but the formulas are sort of decontextualized and listed rather than explained. Would it be better to have a longer and more explanatory appendix in which the derivations are provided and a shorter section 3 aiming to inspire and provide intuition? For example, section 3 ends with "we can see that: [formula for conditional probability of an out-of-sample event as a weight sum of logit link transformed linear models]" Yet, I wonder how many people will find this conclusion that useful: it sort of follows from the authors simply saying that they plan to produce weighted averages of predictive models, and these models would use logistic regression. So, it is either redundant to list the math in this way, or perhaps there is more about the math that the authors need the casual reader to understand (or perhaps the serious reader to understand, too).

MDW: Jacob, what do you think. This might make sense. For PA the math is helpful up front I think. But let's consider the alternative.

JMM: I prefer it the way we wrote it of course. And I really don't think this is an excessive amount of math for a PA article. His suggestion would be like sending in a formal modeling piece where the formal model is in the appendix and the main text is just discussion. I will think more on this. Abstract verbal discussion (without reference to any actual results) of math formulas sounds like a recipe for confusion and misstatement.

However, I think moving section 3.4 to an Appendix (per R2) and expanding a bit on meaning of model weights (per R2) and perhaps an extra sentence or two interpreting the meaning of the different formulas would be an improvement. After Flo gives me his interpretive poem in the EM algorithm, I'll revise the entire section to have more pedagogy with (potentially) some running example.

(8) I liked the honesty of the authors when they spoke about trying to improve their work on this topic in the future [say, in fn 6 or or page 9]. Yet, these statements make me feel like this is more a technical report or working paper rather than a proposal for methodological change that researchers can use immediately. One way around this problem might be for the authors to tell us why they would advise those reading this article to not use the EM algorithm to estimate w or how readers inspired to pursue these ideas ought to think about convergence.

MDW: jeez this guy really read the paper, footnotes and all. Jacob, this is in your court.

JMM: Will do. That will be a long footnote.

1.2 Reviewer 2

This paper introduces to political scientists the ensemble Bayesian model averaging average method for forecasting. The authors relate it to BMA, then show how that (now familiar framework, thanks to recent work) can be related to dynamic forecasting. The authors then provide three well constructed and interesting examples. The conclusion is that the EBMA model outperforms these other forecasting method.

There is a lot to like about this article and certainly makes an important contribution worthy of political analysis. But for it to be published I think the authors should substantially revise this draft (I'll offer a summary of a plan to do this, below). **Missing form the current analysis is much intuition. Further, the current plan for explanation in Sections 1-3 are often confusing and I have serious questions about the results from examples. Together, I think this paper could reasonably appear with revisions that can be done.**

JMM: I love this so much. An excellent summary! With bold font. And totally incoherent.

First, this is an article about forecasting in political science (and then ensembles). So I think it would be useful to emphasize that you have two training periods–and to explain in depth the training period where you construct forecasts. At the moment, the authors assume the existence of a set of forecasts. I think this articles impact would be raised substantially if the authors could (early in the paper) provide concrete examples of forecasting models and how they would be used in the EBMA model. *MDW: Huh? FH: I thought that's what we were doing?*

Jmm: Ummm .... My best guess is that he wants story time about forecasting. We explain that this method takes place in the context of ongoing forecasting efforts by multiple teams or models. We then use these forecasts to construct new forecasts in some future period. Again ... i think this could go into a discussion section at the end of Section 3.

The lack of discussion about forecasts leads me to my only major statistical reservation: the possibility of overfit. At the moment, it appears that the forecasts + weights are constructed on the same data set. This then is an *in sample* fit and therefore will prioritize models that provide too much fit to the data. Pushed to the limit, wouldn't the EBMA model place all the weight on a model that included a fixed effect for each observation in the training period. Illuminating how the EBMA is able to avoid this would be a useful contribution (otherwise, the model would seriously over fit in many applications).

MW : We need to have a subsection on how forecasting guards against overfitting. Jacob, that'd be you.

JMM: Yes. Omnimbus discussion at end of section 3.

This point aside, the authors should spend more time conveying the intuition behind the weights once the forecasts are obtained. This is exactly one example where it would be really useful to walk the reader through the intuition of the EM algorithm. If you were to do this, you would tell the reader that you're measuring the relative ability of each model to predict each observation (to produce $z's$). You can convey what the z's mean. If two models tie for "best fit", then the z will divide its weight over those observations. If one model dominates, then we'd expect z to allocate all its weight to that forecast. Based upon these z's then, we're able to summarize the weights, which essentially capture the proportion of observations that are "best" described by the forecast. *Okay, let's work on an example that works this through. Florian, can you write this up and pass it to Jacob.*

FH: Will do my best.

JMM: Excellent. I suggest starting with the elegant prose R2 provides above.

Providing some intuition about how EBMA generates weights, then, is essential for the reader to understand how the model works (and where it would likely fail) [Regardless of who you do it].

It will also be incredibly useful for you to compare the way EBMA allocates weights to the results. Eyeballing Table 1, it seems that models with very different weights perform similarly and that models with very similar weights perform differently. How does EBMA fit compare to the fit you care about?

incredibly usefully, eh? Okay, let' tackle this one, JM.

JMM: My interpretation is that he wants us to discuss the relationship between model weights and model performance – with something in there about correlated model predictions. I propose a two pronged solution. (1) Add some to the omnibus discusion at the end of Sect. 3 to serve as a guidepost for what is coming. (2) Expand upon the role of correlation and model weights in Section 4.2 per Reviewer 1. I will do these things.

Also, how much of this performance can we attribute to EBMA and how much to the transformation in Section 3.4?

Jmm: Ummm .... Some of each. This is potentially a very broad question that is (in my opinion) well beyond the scope of this paper. We cite the relevant work on this. I suggest we address it in connection to a lengthier discussion of the meaning of "b" per reviewer 2.

A bit more about presentation: the authors (based on Author and Author in text) should strongly consider allocating Section 3.4 to an appendix.

MDW: Maybe so? Whatddya tink, JM?

JMM: I like it. It is our contribution to this method, but it is minor and really an aside to the whole article. Was more important for the grant.

Specific points: *MDW: beef up 1 as indicated. 2. JM? 3. This is future research. 4. I hate to cite S. Purpura, but we can cite this and could even compute the averages and show differences. How much work, Jacob? 5. Fix the typo on page 10.*

JMM:

1. *MDW takes care of this*
2. *JMM: What? I mean I can ... and I see his point. It is not technically essential, but at some point we are just doing a kabookie methods paper. But I will just use the english version when I revise this section.*
3. *I agree ... future research.*
4. *THis is also the same approach being taken by the pollyvote project, and I have seen simple averages used as benchmarks in some of the other papers we cite. I can do it, and it should not be too much work once I get back into the software. But ... it will add at least a paragraph or two and complicate the presentation. You want just a footnote addressing one of the applications? Seems like a reasonable compromise.*
5. *I will fix typo, but (as stated above) I will try to do an extended discussion with running example at end of section 3.*

1) The introduction *under* sells the paper. Lots of folks have been involved in forecasting. The authors are providing a novel methodology to perform this (introducing ensembles seriously to political scientists). Why not prediction matters, we should be able to do it better, and we're going to show you how to do it better? And as a result, increase the use of prediction, etc, etc 2) Equation 11 doesn't add anything, please remove 3) Section 3.5 you say that you're going to use a gibbs sampler. First, gibbs samplers are hard to apply to mixture models. Second, how would you use the uncertainy from the model? You'd probably want to specify more of a fully Bayesian model with priors on weights (introduce Dirichlet smoothing) 4) You might consider checking out Hillard, Purpura and Wilkerson (2009) in Journal of Information Theory and Politics. They introduce ensembles for classification by just averaging responses. How much better is EBMA than just averaging? 5) On page 10 you make reference to a "country" and I'm not sure where that came from. An extend example would be great, though.

1.3 Reviewer 3

This is an excellent paper on EBMA and how it can be used to answer important questions in political science. I am really looking forward to seeing paper in print as well as working with the replication materials for research and classroom instruction. One of the great things about this paper is that it is engaging some very sophisticated EMBA literature, applying it to political science, and pointing out how not only political science, but also other literatures (see review on page 4) use this. It also does a good job of covering the applications to both cross-sectional and dynamic data (section 3).

With this praise in mind, the comments here are mainly meant to ask for some minor clarifications in places that I think a reader could stumble a bit. None of these are meant to be in any sense fatal and if I read the published version in the future and none of these corrections are made, that is OK.

1) Page 5: when you say that the $w_k$weights are "determined by each model's performance within the training data" can you give a quick example? Say someone were fitting two linear regressions to a training set. How would one optimally weight in this case? (I know the answer, but want the clarification for an

intrepid graduate student who is told 5 years from now to "go do this.") *MDW: FH, intrepid graduate student, this is for you.*

FH: Ok, will do

2) Section 3.2ff: I do not like the triple subscript notation of "kst" because it misses some of the conditioning and confounds the indexing of models, training sets, and the future or cross-validation data. Why not something (using LaTeX math here) like $f^k_{s|t}$ (or you could flip the super / sub to get the weights on the same script as the component – $f_k^{s|t}$)? This then makes explicit that the indexing of the components and the temporal or training set dependence. Just a thought. *MDW: JM, what say you?*

JMM: R3 is right. I thought about that before, but never implemented.

3) Footnote 3: Both here and elsewhere there is no discussion of the size or length of the training set. There are some rules of thumb given, but I have seen more formal presentations and evaluations of this in the literatures you cite. For instance, one way to determine the training set size is to look at a spread error plot or some other metric of model fit over the forecast horizon using recursive estimation over different sized training sets. I am not suggesting that you do that here, but maybe it should be mentioned? *MDW: this point need addressing. I'll look for information on this.*

JMM: I have your email on this topic and will incorporate it.

4) Presentation of the binary example, section 3.4: Can you unpack the first sentence of the last paragraph on page 7? What are you doing that follows Sloughter et al. (2007) and Hamill et al. (2004)? Also, can you give the intuition for what the b term is doing in equation 7? (again, think about that future graduate student!)

MDW: JM: what is the intuition for b?

JMM: The intuition for b is that we are lowering the leverage of extremely likely/unlikely events to prevent overfitting of outliers. I wil add more discussion, especially since this will now be an appendix.

5) Example 1, ICEWS analysis: This is very well done. In discussing Tables 1 and 2, you might cite Gneiting and Raftery (2007) and note that only the Brier scores are proper scoring rules that will correctly rank the density forecasts from the binary model. So ranking / comparing the component models on AUC and PRE are potentially going to have different results. *MDW, yes point this out.....Brier are the best!*

JMM: Flo – this seems like an edit for you. Otherwise I will do.

6) POTUS election example: This is a really nice example and one that should really benefit the scholarship on US election forecasting. Two possible caveats to mention: a) We should not expect RMSE or MAE to consistently rank the models (Armstrong and Collopy 1992) relative to the EBMA weights. b) The better comparison for understanding the weights given to each model are the results in Figure 3, which effectively give a forecast density comparison. And what we see here is that better calibrated, sharper forecasts contribute more to the EBMA prediction in each election. Finally, Table 4 could be done as a sharpness plot showing the coverages of the relevant intervals. *MDW: Steal these ideas and put them in paper, JM.*

JMM: Damn. Almost to the end, and no actual changes to plots or tables. Will do, but this will take longer than revising text

7) SCOTUS forecasting example: This one is the briefest and least discussed. Maybe this can be cut or expanded?

FH, can you expand the discussion of this example? Or JM, do you wanna do it? If neither, I'll give it a shot. FH: I can do it, unless you really want to, Jacob. JMM: Go for it. It might help if one of us was actually familliar with the literature on the supreme court.