Using Ensembles of Cognitive Models to Answer Substantive Questions

Henrik Singmann (singmann@gmail.com)
Department of Psychology, University of Zürich
Binzmühlestrasse 14/22, 8050 Zurich, Switzerland

David Kellen (davekellen@gmail.com)
Department of Psychology, Syracuse University
409 Huntington Hall, Syracuse, NY 13244, USA

Eda Mızrak (edamizrak@gmail.com)
Department of Psychology, University of California, Davis
1544 Newton Court, Davis, CA 95616, USA

Ilke Öztekin (ilke.oztekin@nyu.edu)
Department of Psychology, Koç University
Rumelifeneri Yolu, Sarıyer 34450, Istanbul, Turkey

Abstract

Cognitive measurement models decompose observed behavior into latent cognitive processes. For situations with more than one condition, such models allow to test hypotheses on the level of the latent processes. We propose a fully Bayesian ensemble model approach to test hypotheses on the level of the latent processes in situations in which multiple measurement models or model classes exist. In the first step, one needs to perform a Bayesian model selection step comparing the hypotheses within each model class. Aggregating the results of the first step yields *ensemble posterior model probabilities*. We provide an example for a working memory data set using an ensemble of a resource model and a slots model.

Keywords: ensemble models; model selection; Bayesian inference

Introduction

One of the central goals of cognitive science is to develop accurate characterizations of observed behavior in terms of latent cognitive processes. But because it is not possible to directly observe such processes, their existence and interplay needs to be inferred from data coming from specific experimental designs. One privileged approach to evaluate latent processes is through *cognitive measurement models*, which instantiate their relationship with observed data in a clear and general way. These models can be used to directly test process-level hypotheses: For example, if a researcher has data from participants coming from two different populations or under two different conditions she might have specific hypotheses which of the underlying processes (e.g., processes A or B) is responsible for these differences.

One complication often emerges when using some measurement model to guide our inferences: The model we are using is not the only possible account of the data. Many other classes of models also exist, and they can differ substantially in terms of their basic assumptions. If the available evidence suggests that the different models are all at least of approximately equal validity the researcher usually has to make a decision on which of the available models she bases her inference. Here, we propose an alternative solution for such a situation based on *model ensembles*. Model ensembles or ensemble learning is a popular approach in machine learning (e.g., Polikar, 2006) and climate science (e.g., Tebaldi & Knutti, 2007) and usually refers to the process of pooling predictions across different models. A similar approach is taken in *Bayesian model averaging* (Wasserman, 2000). The solution proposed here differs from the existing ensemble or model-averaging approaches in that it is not based on the pooling of model predictions or parameter values, but on results from hypotheses tests.

Specifically, we propose to first perform a model selection step across the different hypotheses – separately for each model class. For example, consider model classes $\mathcal{M}_\bullet^P$ and $\mathcal{M}_\bullet^R$, each with its own sets of assumptions. Within each model class we establish special cases that reflect the different hypotheses of interest (e.g., models $\mathcal{M}_1^P$ and $\mathcal{M}_1^R$). The hypotheses being tested across model classes should reflect analogous statements. For example, the hypothesis that memory is improved across conditions, which can be stated in terms of both model classes, irrespective of the exact way memory processes are established within each class. Using Bayesian model selection, we obtain *posterior model probabilities* for each of the hypothesis of interest, separately for each class. In a second step, these posterior model probabilities are pooled across model classes to produce the *ensemble posterior model probabilities*.

The remainder of this manuscript is organized as follows. We first provide an overview over Bayesian model selection and describe how to calculate ensemble posterior model probabilities. We then provide an example using a working-memory dataset and two popular measurement models in this domain.

Bayesian Model Selection

The main principle underlying the Bayesian statistical framework is the quantification of uncertainty with probabilities (Gelman et al., 2013). Estimating the parameters θ of model $\mathcal{M}$ with likelihood function $p_{\mathcal{M}}(y|\theta)$ in a Bayesian framework requires the specification of a prior distribution $p_{\mathcal{M}}(\theta)$. The prior distribution quantifies the relative uncertainty one has regarding ones parameters prior to seeing the data y. This prior distribution is then updated in light of the data using Bayes' theorem yielding a posterior distribution $p_{\mathcal{M}}(\theta|y)$, with

$$p_{\mathcal{M}}(\theta|y) = \frac{p_{\mathcal{M}}(y|\theta)p_{\mathcal{M}}(\theta)}{\int p_{\mathcal{M}}(y|\theta)p_{\mathcal{M}}(\theta)\,d\theta}. \quad (1)$$

The posterior distribution quantifies the uncertainty or knowledge one has regarding ones parameters after seeing the data. If the parameters of the model are of primary interest the posterior distribution can be directly used for inference.

In most real life cases the posterior distribution cannot be obtained analytically. A common approach to approximate it

is to sample from the unnormalized posterior (i.e., the numerator in Equation 1) using numerical methods such as Markov chain Monte Carlo (MCMC). For parameter estimation this is completely sufficient as the posterior is proportional to the unnormalized posterior for fixed data. The normalizing constant in the denominator,

$$p(y|\mathcal{M}) = \int p_{\mathcal{M}}(y|\theta)p_{\mathcal{M}}(\theta)\,d\theta, \tag{2}$$

usually called the *marginal likelihood*, is independent of θ, and cannot be obtained as easily as the unnormalized posterior. However, Gronau, Singmann, and Wagenmakers (2017) have recently introduced a software package that allows to obtain the marginal likelihood from the unnormalized posterior using *bridge sampling*.

Marginal likelihoods are the central quantity of interest in *Bayesian model selection* (Wasserman, 2000), as they provide a compromise between a model's ability to describe the observed data at hand, and its a priori ability to describe any data. The first multiplicative term in Equation 2 corresponds to the likelihood of the data under a certain set of parameter values θ. The second term corresponds to the prior probability attributed to this set of parameter values. The marginal likelihood therefore corresponds to a weighted average likelihood (averaged across parameter values), where the weights are given by the parameter prior probability distribution. Now, this average likelihood will be smaller if the prior weights are dispersed across parameter values that would make predictions that are far from the observed data. This reduction is a de facto penalty for unwarranted flexibility, thereby implementing the principle of parsimony also known as *Occam's razor* (e.g., Myung & Pitt, 1997). Ideally, a model and its prior parameter distribution should produce a range of predictions that encompasses what is actually observed and excludes what is not. Models which are able to predict almost all possible data will have a lower marginal likelihood than those models that only predict those data that are actually observed.

The model with the largest marginal likelihood is then considered to be the one that gathers the greatest support (i.e., the best trade-off between goodness of fit and flexibility). For any finite set of models $\mathcal{M}_1, ..., \mathcal{M}_k$, their marginal likelihoods can be used to compute *posterior model probabilities*. For model $\mathcal{M}_i$:

$$p(\mathcal{M}_i|y) = \frac{p(y|\mathcal{M}_i)p(\mathcal{M}_i)}{\sum_{j=1}^{k} p(y|\mathcal{M}_j)p(\mathcal{M}_j)}, \tag{3}$$

where $p(\mathcal{M}_i)$ is the prior model probability for model $\mathcal{M}_i$ before seeing the data. In words, the posterior model probabilities are equal to the relative product of marginal likelihood and prior model probabilities across all candidate models.

An important aspect of the marginal likelihood is that posterior model probabilities are not only sensitive to the prior model probabilities, but also depend to a large extent on the parameter priors. For example, it is easy to construct cases for which the choice of priors can determine which model ultimately provides the best account (see e.g., *Lindley's paradox;* Hill, 1982). Consequently, much research is concerned with developing appropriate priors for Bayesian model selection in specific situations. This issue is particularly serious when attempting to compare non-nested models, as it becomes less obvious how priors can bias results.

For nested models corresponding to a standard null hypothesis – that is, model $\mathcal{M}_0$ is a special case of a model $\mathcal{M}_1$ such that fixing one parameter in $\mathcal{M}_1$ to zero produces $\mathcal{M}_0$ – a recipe for producing *default priors* was developed by Jeffreys (1961). The basic idea is to reparameterize the models such that the difference parameter δ that differs between $\mathcal{M}_1$, with parameters $\theta_1 = (\theta_0, \delta)$, and $\mathcal{M}_0$, with parameters θ_0, is scaled by the parameters θ_0 that are shared across $\mathcal{M}_1$ and $\mathcal{M}_0$. For example, for a t-test $\delta = \frac{\mu}{\sigma}$, where μ is the overall mean and σ the overall or pooled variance. As a consequence, the scale of δ becomes independent of the actual data and a prior on θ can be specified in a reasonable manner. Common priors for δ are a normal (i.e., Gaussian) or Cauchy distribution.

Ensemble Posterior Model Probabilities

Ensemble posterior model probabilities are concerned with a situation in which we have at least two experimental conditions and models from at least two model classes that can be applied to each condition and decompose the observed behavior into similar latent processes. We also assume that the interest of the researcher is in inferences on these latent processes across conditions. What constitutes a model class can be understood broadly (e.g., completely different assumption or merely different settings of parameter values). Here it is only relevant that Bayesian model selection between models from these model classes might be seen as problematic.

The first step is to perform a Bayesian model selection step within each model class using the default prior approach of Jeffreys (1961). That is, one needs to calculate marginal likelihoods for a set of models belonging to each model class that correspond to the possible ways in which the parameters representing the latent processes can be restricted between the conditions. Note that the set of models needs to be constructed such that for each model in a given model class an equivalent model exists in the other model classes. Equivalent here means that the parameter restrictions of the model in a given model class correspond to a hypothesis in terms of the latent processes that can also be expressed as a parameter restrictions for each of the other model classes. The marginal likelihoods are then used to calculate posterior model probabilities within each model class.

To calculate the ensemble posterior model probabilities, the posterior model probabilities need to be aggregated across model classes. Specifically, one needs to take the average across models which correspond to the same hypothesis. The result can be used for inferences on the hypotheses on the level of the latent processes.

As before, let us consider a situation with two model

classes, class R with models $\mathcal{M}_\bullet^R$ and class P with models $\mathcal{M}_\bullet^P$, which each decompose the observed behavior into latent processes a and b, via parameters θ_a and θ_b, respectively. Specifically, class R does so via parameters θ_a^R and θ_b^R and class P via parameters θ_a^P and θ_b^P (a substantive example is given in the next section). Further, assume we have obtained data y from two conditions f and g and are interested if and how the latent processes differs between the conditions. Because we have two latent processes that might or might not differ independently across the conditions we need to set up a total of four models for each model class to explore the full hypothesis space. In case parameter differences are assumed to exist between models we introduce standardized difference parameters δ as discussed above.

- $\mathcal{M}_\varnothing$ assumes no differences between f and g and has two parameters, (θ_a, θ_b), with $\theta_{a,f} = \theta_{a,g} = \theta_a$ and $\theta_{b,f} = \theta_{b,g} = \theta_b$.
- $\mathcal{M}_a$ assumes differences in a only (between f and g) and has three parameters, $(\theta_a, \delta_a, \theta_b)$, with $\theta_{a,f} = \theta_a + \frac{\delta_a}{2}$, $\theta_{a,g} = \theta_a - \frac{\delta_a}{2}$, and $\theta_{b,f} = \theta_{b,g} = \theta_b$.
- $\mathcal{M}_b$ assumes differences in b only and has three parameters, $(\theta_a, \theta_b, \delta_b)$, with $\theta_{a,f} = \theta_{a,g} = \theta_a$, $\theta_{b,f} = \theta_b + \frac{\delta_b}{2}$, $\theta_{b,g} = \theta_b - \frac{\delta_b}{2}$.
- $\mathcal{M}_{ab}$ assumes differences in both a and b and has four parameters, $(\theta_a, \delta_a, \theta_b, \delta_b)$, with $\theta_{a,f} = \theta_a + \frac{\delta_a}{2}$, $\theta_{a,g} = \theta_a - \frac{\delta_a}{2}$, $\theta_{b,f} = \theta_b + \frac{\delta_b}{2}$, $\theta_{b,g} = \theta_b - \frac{\delta_b}{2}$.

After calculating the marginal likelihoods for each model, Bayesian model selection provides us with a set of four posterior model probabilities for each model class, $p(\mathcal{M}_i^P|y)$ for class P and $p(\mathcal{M}_i^R|y)$ for class R, where $i \in (\varnothing, a, b, ab)$. The ensemble posterior model probabilities are then given by

$$p(\mathcal{M}_i^e|y) = \frac{1}{2}\left(p(\mathcal{M}_i^P|y) + p(\mathcal{M}_i^R|y)\right). \tag{4}$$

Measurement Models for Detection Experiments

In many experiments across a variety of different domains – such as perception, memory, or reasoning – participants have to decide for each stimulus if it falls into one of two mutually exclusive categories: signal absent (e.g., a not-studied lure in a memory experiment) versus signal present (e.g., a studied stimulus). The observed behavior in such *detection experiments* is usually described in terms of two proportions that sufficiently summarize the data, the probability of correctly detecting a signal present trial as such called *hits* (i.e., p("signal"|signal present)) and the probability of incorrectly denoting a signal absent trial as a signal present trial called *false alarm* (i.e., p("signal"|signal absent)).

We will discuss two prominent measurement models that decompose the performance in detection experiments into two latent processes, *discriminability* and *response bias*. Discriminability is the ability of the decision maker to distinguish the two model classes and therefore reflects task performance. A decision maker who is comparatively good in the task will have a high value on the discriminability parameter. Better discriminability is associated with a decrease in hits and a decrease in false alarms. Response bias captures the propensity of the decision maker to prefer one of the two response options (i.e., "signal" and "no signal") and is in principle independent of task performance. A conservative decision maker will have a propensity for responding "no signal" independent of the actual stimulus category; a liberal decision maker will have a propensity for responding "signal" independent of the actual stimulus category.

The most prominent measurement models for detection experiments are based on *signal detection theory* (Green & Swets, 1966; Kellen & Klauer, in press). Signal detection theory assumes that the decision maker has access to a continuous psychological strength dimension. Signal present and signal absent stimuli are represented as distributions on the strength dimension. At test, each stimulus evokes a strength signal which is compared with an established response criterion c. If the signal surpasses the criterion, the decision maker responds with "signal"; otherwise the response "no signal" is given. The usual assumption is that both stimulus classes follow a normal (i.e., Gaussian) distribution. Furthermore, to establish the scale on the strength dimension, the mean μ_l and variance σ_l^2 of the lure distribution is fixed to 0 and 1, respectively. For the type of design described here it is also common to fix the variance of the signal distribution σ_s^2 to 1 as well. This provides the following model equations:

$$\begin{aligned} p(\text{hit}) &= \Phi(\mu_s - c), \\ p(\text{false alarm}) &= \Phi(-c), \end{aligned} \tag{5}$$

with $\Phi()$ corresponding to the cumulative distribution function of the standard normal distribution.

In terms of the latent processes, discriminability is captured by parameter μ_s. The further apart the two distributions – that is, the larger μ_s – the better the discriminability between the two stimulus classes. The response bias is captured by parameter c. The larger c (given fixed μ_s) the more conservative the response pattern.

Another popular measurement model is based on *threshold theory* (e.g., Rouder & Morey, 2009; Luce, 1963), which assumes that decision makers do not have direct access to a continuous strength signal. Instead, the decision maker only has access to a small number of discrete states. Here, we are concerned with the high-threshold variants which assume that for each item there is a certain probability D with which individuals detect the true status of an item (i.e., signal present or signal absent). In this detection state, the correct answer (i.e., "signal" for signal present trials and "no signal" for signal absent trials) is invariably given. In case the true status of an item is not detected with probability $1 - D$, an uncertainty state is reached. In the uncertainty state response "signal" is guessed with probability g and response "no signal" with probability $1 - g$. This provides the following model equa-

tions:

$$\begin{aligned} p(\text{hit}) &= D + (1-D)g, \\ p(\text{false alarm}) &= (1-D)g. \end{aligned} \tag{6}$$

In terms of the latent processes, discriminability is captured by parameter D. The better the ability of the decision maker to detect the true status of the item the larger D. The response bias is captured by parameter g. A conservative response bias corresponds to a value of $g < .5$ and a liberal response bias corresponds to $g > .5$.

Experiment and Bayesian Modeling

We have collected data from two between-subjects groups using a simple detection experiment within the domain of working memory. The two groups were high working capacity and low working memory capacity. Our substantive question was if there were any differences in terms of the two latent processes (i.e., discriminability and response bias) between the two working memory capacity groups. Note that in the working-memory domain signal-detection models are also known as *resource models* and threshold models as *slots models* (e.g., Donkin, Tran, & Nosofsky, 2013).

To perform the Bayesian model selection step for each of the two model classes (i.e., signal-detection and threshold models) we implemented both as hierarchical Bayesian versions with crossed-random effects for participants and items. The hierarchical-structure was setup in such a way that we not only modeled individual-level random effects, but also the correlation among the individual-level effects using the latent-trait approach of Klauer (2010).[1] Furthermore, we assumed independent group-level variances for the participant-effects, but fixed the correlations across the two groups. For the threshold-model, the individual-level parameters were estimated on the unconstrained-scale and then probit transformed onto the unit range (Klauer, 2010). No transformation was necessary for the signal-detection model.

The models were implemented in `Stan` (Carpenter et al., 2017) which allowed us to put separate priors on the group-level correlations (so-called LKJ-priors with a non-informative scale parameter of 1) and variances (weakly-informative half Cauchy priors with scale 5). For the group-level model parameters, the priors were either non-informative (i.e., normal with mean 1 and variance 1 for the threshold model on the probit scale) or weakly-informative (i.e., Cauchy with location 0.5 and scale 5, truncated at zero).

For each model class we implemented four different models corresponding to the four possible hypotheses as described in the example above: no differences between conditions (model *none*), difference in *discriminability* only, difference in response *bias* only, and differences in *both*, discriminability and response bias. For those models that included difference parameters δ, those were normalized based on the pooled standard deviation from both groups and had a Cauchy prior with location 0. For each model class and each model corresponding to a hypothesis, we estimated two model versions, one with a Cauchy scale of $\frac{2}{\sqrt{2}}$ for "medium" sized effects and one with a Cauchy scale of 0.5 for "small" effects (Morey & Rouder, 2015).[2]

To estimate the models, we obtained a total of 40,000 draws from the posterior distribution. These draws came from four independent MCMC chains which, after 1000 warmup samples, ran for 20,000 samples, retaining every second sample. The models showed excellent convergence, maximal $\hat{R} < 1.001$. This comparatively large number of posterior samples was necessary for obtaining adequate estimates of the marginal likelihood, which we obtained via the `bridgesampling` package (Gronau et al., 2017). For each set of posterior samples we obtained five independent estimates of the marginal likelihoods to check for the stability of the estimates. We performed this check on the level of the posterior model probabilities which showed a maximal difference of 1.7% across marginal likelihood estimates, an acceptable amount of variability. We report results based on the median of the five estimates.

Method

Participants Five-hundred and ninety students of Koç University were screened using the automated operation span task (Unsworth, Heitz, Schrock, & Engle, 2005) to attain working memory capacity measures. Of those, 21 high span individuals (upper quartile of the sample) and 19 low span individuals (lower quartile of the sample) participated in the experiment. Participants received partial course credit for participation in the screening session and monetary compensation for taking part in the experiment.

Working Memory Task The task was a Sternberg paradigm (i.e., short-term recognition) with letters. Stimuli consisted of 18 consonants (`b`, `c`, `d`, `f`, `g`, `h`, `j`, `k`, `l`, `m`, `n`, `p`, `r`, `s`, `t`, `v`, `y`, `z`) displayed in lower case. Each study list comprised 4 consonants drawn randomly without replacement from the stimulus pool that had not appeared in the two preceding lists. Study list items were presented sequentially for 500 ms each at the center of the screen. Following a mask of 500 ms the test probe was shown. Probes were either targets (i.e., shown in the study list) or lures (i.e., not shown in the study list or the previous trial). Participants indicated whether a probe was a target or lure via key press.

Results

Model Fit Figure 1 shows the observed and estimated accuracies for targets and lures as a function of working memory capacity. Each panel shows the estimates of the models corresponding to one substantive hypothesis. When looking at the

[1] To avoid a non-identifiable signal-detection model we only implemented item-effects for μ_s and not for c. Consequently, we only had item-effects for one parameter (i.e., μ_s) and could not estimate a covariance matrix.

[2] Note that in line with Rouder, Morey, Speckman, and Province (2012) we used a scaling factors of $\pm\frac{2}{\sqrt{2}}$ instead of $\pm\frac{1}{2}$ for δ.

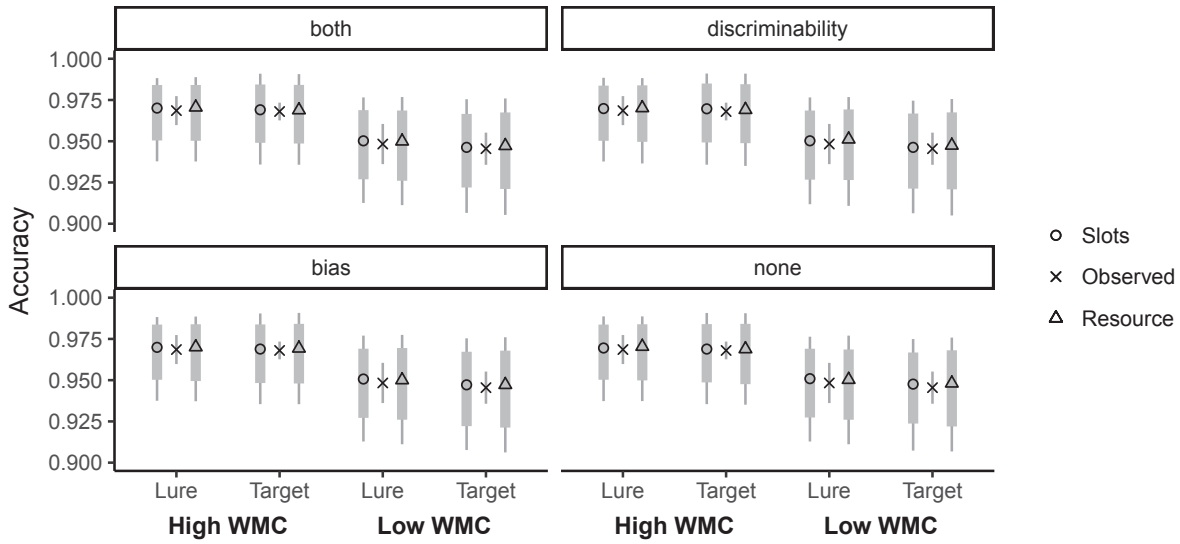

Figure 1: Observed and estimated accuracy as a function of working memory capacity (WMC). Each panel shows the same observed data, but the model estimates (i.e., posterior predictions) from models corresponding to different hypotheses with prior scale of the difference parameters $\delta = \frac{2}{\sqrt{2}}$. For the observed data the error bars shows ± 1 standard error of the mean and the $\times$ the mean of the individual-level means. Model estimates are based on the posterior predictive distributions of the individual-level mean accuracies and the symbols show the mean median, the inner confidence bars shows the mean 80% credibility interval, and the outer confidence bars shows the mean 95% credibility interval.

data it is clear that the differences between the high and low capacity group are rather small ($\approx 2.5\%$), especially when taking the uncertainty of the estimates into account.

Both model classes appear to provide an excellent account to the data, independent of the hypothesis that the model represents; even for the models representing no difference between the two conditions, the models capture the small differences between the conditions basically perfectly. This somewhat intriguing finding can be explained by the by-participant random-effect. For example, the mean of the random-effects estimates for μ_s of the *both* signal-detection model is 0.06 for the high capacity group and -0.03 for the low capacity group. In contrast, the same estimates are 0.21 for the high capacity group and -0.23 for the low capacity group for the *none*-model. In other words, the restricted model can explain the data by pushing the participant random-effects away from the zero-centered prior whereas this effect is captured by δ if present. To distinguish the models in terms of overall adequacy, the marginal likelihoods tell us which of the models provides an on average better account across the whole parameter space, taking the priors into account.

Ensemble Posterior Model Probabilities The ensemble posterior model probabilities are shown as '$\times$'s in Figure 2. The pattern is not overwhelmingly strong, but with over 60% probability the model assuming no difference receives clearly the strongest support. The second strongest support, with 20% probability, goes to the model assuming discriminability differs. Thus, overall (i.e., regardless of whether a threshold model or signal-detection model is a better model of working memory) there is no strong evidence that the small difference between the two conditions reflects a specific difference in terms of the latent processes usually assumed in detection experiment.

Figure 2 also shows the posterior model probabilities which form the basis for the ensemble posterior model probabilities. The two model classes suggest similar conclusions,

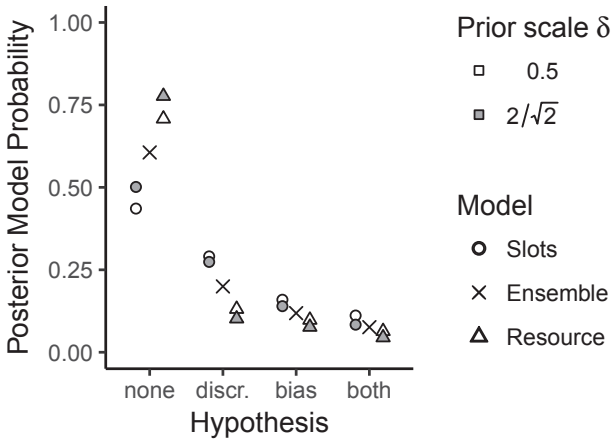

Figure 2: Ensemble posterior model probabilities (i.e., $\times$) and posterior model probabilities (conditional on model class and prior choice) across the four substantive hypotheses.

with the threshold model putting somewhat more probability mass on models assuming an effect compared to the signal-detection models. Furthermore, the impact of model class appears to be larger than the impact of prior width.

Discussion

Box's famous adage "all models are wrong" (Box, 1976) captures the notion that any model-based characterization of data is going to be a caricature at best. Given this situation, there is the need to ensure that any inferences made do not hinge on the characteristics of the particular model used. The ensemble approach proposed here attempts to ensure that, by considering the degree of support for a given hypothesis across different model classes. Although all models are wrong, they are wrong in different ways. Under the assumption that each model captures the true data generating process to some approximate degree, but also overfits the data in its own idiosyncratic way, the consensus across model classes is more likely to reflect the true data-generating processes. In this sense, the ensemble approach can been seen as trying to tap into a "wisdom of the crowd" (Surowiecki, 2004). The success of such ensemble approaches over single models in terms of out-of-sample predictive ability has been empirically demonstrated within cognitive science (e.g., Erev et al., 2010).

Ensemble posterior model probabilities allow to test substantive hypothesis across conditions and model classes using fully Bayesian model selection. This approach is useful when multiple cognitive measurement models that decompose the observed data into similar latent processes exist for a given dataset. The main difference to existing ensemble approaches is that we propose to use the full data (cf. Polikar, 2006, for machine learning approaches that split the data) and build ensembles based on inferential information (i.e., posterior model probabilities) instead of model predictions (e.g., in climate science; Tebaldi & Knutti, 2007). Our approach also does not require any type of model weighting common in other ensemble approaches (e.g., in Bayesian model averaging) as the relevant situation is one where the candidate models are of approximately equal validity.

Acknowledgments

Henrik Singmann and David Kellen received support from the Swiss National Science Foundation Grant 100014_165591.

References

Box, G. E. P. (1976). Science and Statistics. *Journal of the American Statistical Association*, *71*(356), 791–799.

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., … Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, *76*(1), 1–32.

Donkin, C., Tran, S. C., & Nosofsky, R. (2013). Landscaping analyses of the ROC predictions of discrete-slots and signal-detection models of visual working memory. *Attention, Perception, & Psychophysics*, 1–14.

Erev, I., Ert, E., Roth, A. E., Haruvy, E., Herzog, S. M., Hau, R., … Lebiere, C. (2010). A choice prediction competition: Choices from experience and from description. *Journal of Behavioral Decision Making*, *23*(1), 15–47.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian Data Analysis, Third Edition*. Hoboken: CRC Press.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York: Wiley.

Gronau, Q. F., Singmann, H., & Wagenmakers, E.-J. (2017). bridgesampling: An R Package for Estimating Normalizing Constants. *arXiv:1710.08162 [stat]*.

Hill, B. M. (1982). Lindley's Paradox: Comment. *Journal of the American Statistical Association*, *77*(378), 344.

Jeffreys, H. (1961). *Theory of probability* (3ed. ed.). Oxford: Clarendon Press.

Kellen, D., & Klauer, K. C. (in press). Elementary Signal Detection and Threshold Theory. In *The Stevens Handbook of Experimental Psychology and Cognitive Neuroscience.*

Klauer, K. C. (2010). Hierarchical Multinomial Processing Tree Models: A Latent-Trait Approach. *Psychometrika*, *75*(1), 70–98.

Luce, R. D. (1963). A Threshold Theory for Simple Detection Experiments. *Psychological Review*, *70*, 61-79.

Morey, R. D., & Rouder, J. N. (2015). *BayesFactor: Computation of Bayes factors for common designs.* (R package version 0.9.12-2. https://CRAN.R-project.org/package=BayesFactor)

Myung, I. J., & Pitt, M. A. (1997). Applying Occam's razor in modeling cognition: A Bayesian approach. *Psychonomic Bulletin & Review*, *4*(1), 79–95.

Polikar, R. (2006). Ensemble based systems in decision making. *IEEE Circuits and Systems Magazine*, *6*, 21-45.

Rouder, J. N., & Morey, R. D. (2009). The nature of psychological thresholds. *Psychological Review*, *116*(3), 655–660.

Rouder, J. N., Morey, R. D., Speckman, P. L., & Province, J. M. (2012). Default Bayes factors for ANOVA designs. *Journal of Mathematical Psychology*, *56*(5), 356–374.

Surowiecki, J. (2004). *The wisdom of crowds*. New York: Random House.

Tebaldi, C., & Knutti, R. (2007). The use of the multi-model ensemble in probabilistic climate projections. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, *365*(1857), 2053–2075.

Unsworth, N., Heitz, R. P., Schrock, J. C., & Engle, R. W. (2005). An automated version of the operation span task. *Behavior Research Methods*, *37*(3), 498–505.

Wasserman, L. (2000). Bayesian Model Selection and Model Averaging. *Journal of Mathematical Psychology*, *44*(1), 92–107.