Probabilistic model uncertainty

Chris Holmes
with V. Shirvaikar & S. Walker

Ellison Institute of Technology, University of Oxford, The University of Texas at Austin

June 9, 2025

Motivation

- We explore model uncertainty through the lens of **predictive inference**
- The predictive viewpoint highlights observables and missing data as the source of all statistical uncertainty
- If we had access to data on the full population, then any identifiable quantity would be known, including the 'best' model
- Make the missing data the focus of the modelling

Predictive Approach to Uncertainty

- Uncertainty stems from unobserved data $x_{n+1:N}$, given observations $x_{1:n}$, where N is the total population size
- Construct joint $p(x_{n+1:\infty} \mid x_{1:n})$ and impute missing information
- Use recursive updates (chain rule) and **start from the data in-hand** to construct

$$p(x_{n+1:\infty} \mid x_{1:n}) = \prod_{i>n} p(x_i \mid x_{1:i-1})$$

  avoiding the need for prior elicitation as the starting point is $p(x_{n+1} \mid x_{1:n})$
- Inspired by de Finetti (1937) and the focus on observables, and modern predictive inference (relaxing exchangeability).

Key Aspects

- **No priors** required — inference is data-driven and "objective"
- **Simple protocol**: alternate model comparison and data simulation.
- **Efficient and parallelizable**: computational burden is in repeated model updating
- **Bayesian**: in specifying a conditional distribution directly on the model space, where uncertainty arises from the missing data – but we don't assume exchangeability

Traditional Bayesian Model Uncertainty

- Posterior probability of model: $P(\mathcal{M}_k \mid \mathcal{D})$

$$P(\mathcal{M}_k \mid \mathcal{D}) = \frac{P(\mathcal{D} \mid \mathcal{M}_k)P(\mathcal{M}_k)}{P(\mathcal{D})} \tag{1}$$

- Depends on marginal likelihood and model priors
- Used in model averaging for prediction and selection

Marginal Likelihood and Bayes Factors

- Marginal likelihood integrates over model specific parameters, $P(\mathcal{D} \mid \mathcal{M}_j) = \int f_{\theta_j}(x)\, p(\theta_j) d\theta_j$
- Bayes factor compares model evidence

$$BF = \frac{P(\mathcal{D} \mid \mathcal{M}_1)P(\mathcal{M}_1)}{P(\mathcal{D} \mid \mathcal{M}_2)P(\mathcal{M}_2)} \tag{2}$$

  with a value of $BF > 1$ presumably indicating support for $\mathcal{M}_1$ over $\mathcal{M}_2$, and vice versa (Kass and Raftery, 1995)
- Sensitive to priors, can be complex to compute

Bayes Factors

- A well-known problem with the Bayes factor is that it can heavily depend on the model priors, and requires integration over parameter spaces
- Efforts to specify objective prior distributions have led to an array of Bayes factor alternatives, such as intrinsic and fractional Bayes Factors
- These methods use some of the observed data to help specify the prior in some way, such as by weighting the likelihood (O'Hagan, 1995) or setting aside a portion of data for "training" (Berger and Pericchi, 1996)
- This still requires an element of user choice, and can also result in the loss of some information contained within the observed data.

BIC as an Approximation

- The Bayesian information criterion (BIC) provides an approximation to the negative log-evidence, given by

$$\mathrm{BIC} = \mathrm{d}\log\mathrm{n} - 2\log\widehat{\mathcal{L}}$$

  where d is the dimension of the model, n is the sample size, and $\widehat{\mathcal{L}}$ is the maximum likelihood of the model at the optimal parameter values.
- Penalizes model complexity
- A tempting idea is to use $\exp(-\mathrm{BIC})$ as the marginal likelihood, but Kass and Raftery (1995) show that this has a relative error of $O(1)$ in approximating the Bayes factor, meaning that even for large samples, the BIC should not be used directly to extract posterior probabilities

Recursive model sampling

- Views uncertainty as missing data
- With observed $x_{1:n}$, the guiding principle is that uncertainty quantification for any statistical task, including model selection, requires the construction of a model for the data we have not observed, given what has been observed
- Alternates model selection and data simulation
- Estimates $P(\mathcal{M}_k \mid \mathcal{D})$ via Monte Carlo

Predictive Resampling for Model Selection

- We have observed data $x_{1:n}$ and candidate model set $\{\mathcal{M}_k\}$.
- Use a consistent selection criterion C, such as BIC, to choose best model $\mathcal{M}_{\widehat{k}(n)}$.
- Sample:

$$x_{n+1} \sim p(\cdot \mid \mathcal{M}_{\widehat{k}(n)}, \widehat{\theta}_{\widehat{k}(n)})$$

- Update models and select the new best model given the additional information
- Repeat up to x_N, for some large N
- Pick off the final model $\mathcal{M}_{\widehat{k}(N)}$ as a sample from the posterior

Posterior over Model Space

- Monte Carlo estimate of model uncertainty:

$$p(\mathcal{M}_k \mid x_{1:n}) = \frac{1}{B} \sum_{b=1}^{B} \mathbf{1}\left(\mathcal{M}_{\widehat{k}(\infty)}^{(b)} = \mathcal{M}_k\right)$$

- Avoids priors over models or parameters.
- Converts consistent model selection into a posterior over models.

Relation to Existing Work

- Related to prequential forecasting (Dawid, 1984).
- Avoids Bayes factors and need for RJ-MCMC complexities.
- Inspired by Draper's notion of model expansion:
  - Structural uncertainty propagated by predictive updates.

Predictive Resampling: Algorithm

Algorithm Predictive resampling

Specify search space of candidate models $\{\mathcal{M}_k\}$
Set number of trials B and final sample size $N >> n$
for b from 1 to B **do**
    for i from $n+1$ to N **do**
        Calculate consistent model selection criterion $C(\mathcal{M}_{k(i-1)}, x_{1:i-1})$
        Optimize and identify best model $\mathcal{M}_{\widehat{k}(i-1)} = \text{argmax}_k\, C(\cdot,\cdot)$
        (If applicable) Identify parameter MLE $\hat{\theta}_{k(i-1)}$
        Sample $x_i \sim p(\cdot|\mathcal{M}_{\widehat{k}(i-1)})$ and add to training data
    end for
    Record final model $\mathcal{M}_{\widehat{k}(N)}$
end for
Return probabilities $p(\mathcal{M}_k \mid \mathcal{D}) = B^{-1}\sum_{b=1}^{B} \mathbf{1}(\mathcal{M}^{(b)}_{\widehat{k}(\infty)} = \mathcal{M}_k)$

Note on supervised learning

- For supervised learning, with covariates X and outcomes Y we copy $x_{1:n}$ as a block (fixed design), yielding $x_{n+1:2n}$
- predict $y_{n+1:2n}$ using the optimal fitted model; add $(x, y)_{n+1:2n}$ to the observed data; and so forth
- Sampling the entire outcome vector $Y_k = y_{kn+1:n(k+1)}$ for $k = 1, 2, \ldots$ in blocks, rather than sampling individual observations
- keeps with the idea of "repeating the experiment", and also simplifies the update calculation

Link to traditional Bayes

- The approach recovers the conventional Bayesian approach when using the usual posterior predictive to sample new observations
- Rather than optimizing and sampling from the best model, we would draw x_i directly from the posterior predictive using the appropriate model mixing weights
- Each of the B trials would eventually converge to a single model
- The relative proportions of these trials would converge to the initial model mixing weights as $B \to \infty$
- The predictive framework is a generalization of standard Bayesian model uncertainty.

Link to Bayesian Updating

- Generalizes Bayesian inference under predictive viewpoint
- No explicit priors required
- Based solely on observable data and a pre-selected model selection criterion

Simple Example: Hypothesis Testing

- Compare two point hypotheses $H_0 : \theta = \theta_0$ and $H_1 : \theta = \theta_1$, for the unknown mean parameter of a normal distribution, $N(\theta, 1)$, with known variance $\sigma^2 = 1$
- Select model via maximum log-likelihood (BIC penalty with common dimension)
- Propagate uncertainty through simulation

Experimental Setup

- $n = 100$ observations from $\mathcal{N}(0, 1)$, so true hypothesis is H_0
- Vary θ_1 from -0.3 to 0.3
- $B = 1000$ trials, $N = 2600$ samples

Experimental Setup

- We draw $n = 100$ observations from $\mathcal{N}(0, 1)$, so true hypothesis is H_0
- We found $\bar{x} = 0.031$
- We consider the evidence for the alternative $\theta_1 = 0.01$
- $B = 1000$ trials, $N = 2600$ samples
- Note that with $\bar{x} = 0.031$, so in this case we have $\mathcal{M}_{\widehat{k}(100)} = H_0$, always
- That is, we always start the recursive sampling from the optimal model given $x_{1:n}$

Recursive sampling showing sample means $\bar{x}_{1:n+i}$

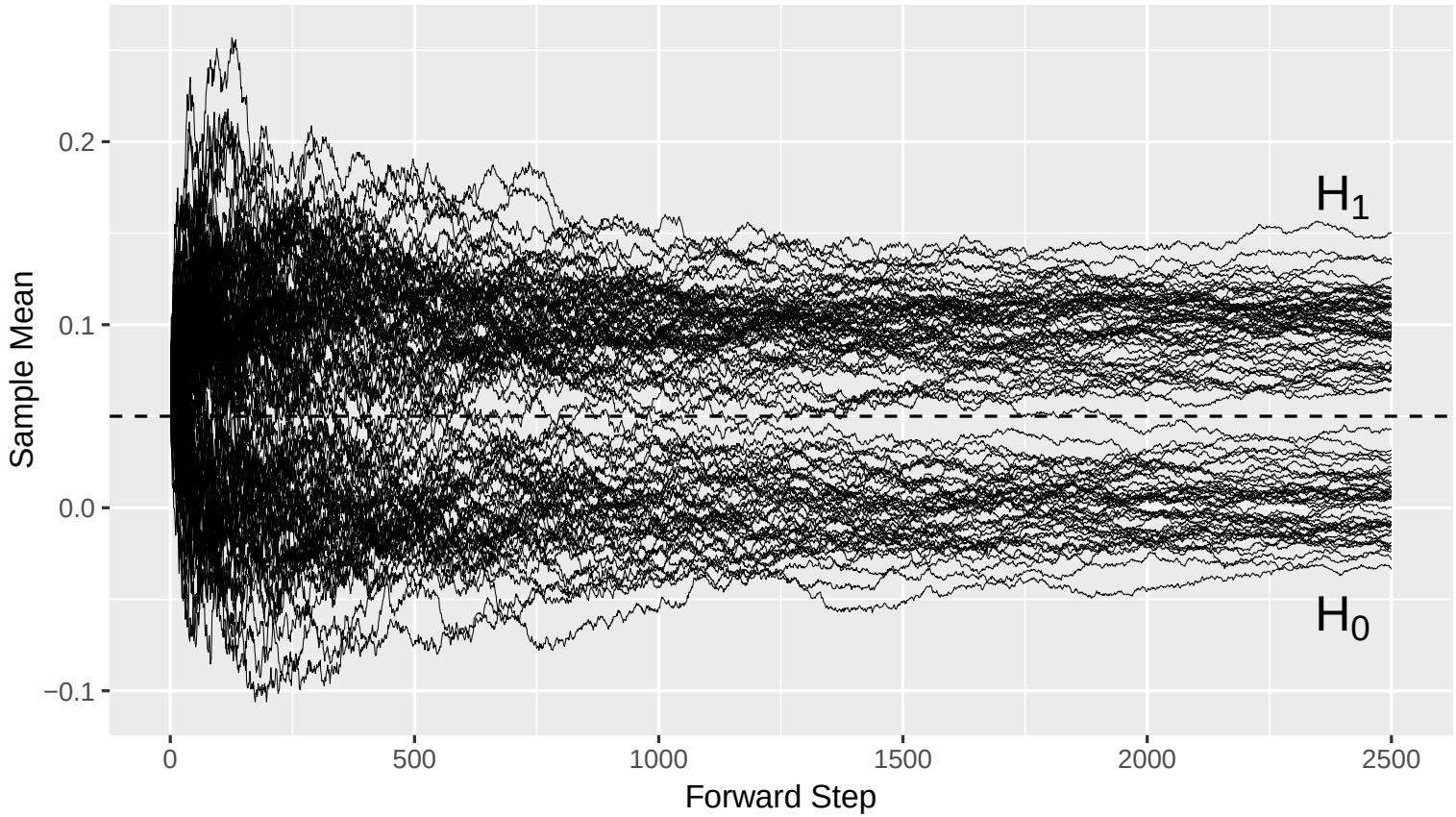


Figure: Uncertainty through sampling of missing observations, where different possible realizations of the complete data start at the same $\mathcal{M}_{\widehat{k(n)}} = H_0$, but individually converge to sampling from H_0 or H_1. The relative number of sample paths in the two regions for large N gives the approximation to $P(H_0 \mid x_{1:n})$

Experimental Setup

- $n = 100$ observations from $\mathcal{N}(0, 1)$, so true hypothesis is H_0
- Vary θ_1 from -0.3 to 0.3
- $B = 1000$ trials, $N = 2600$ samples

Varying H_1

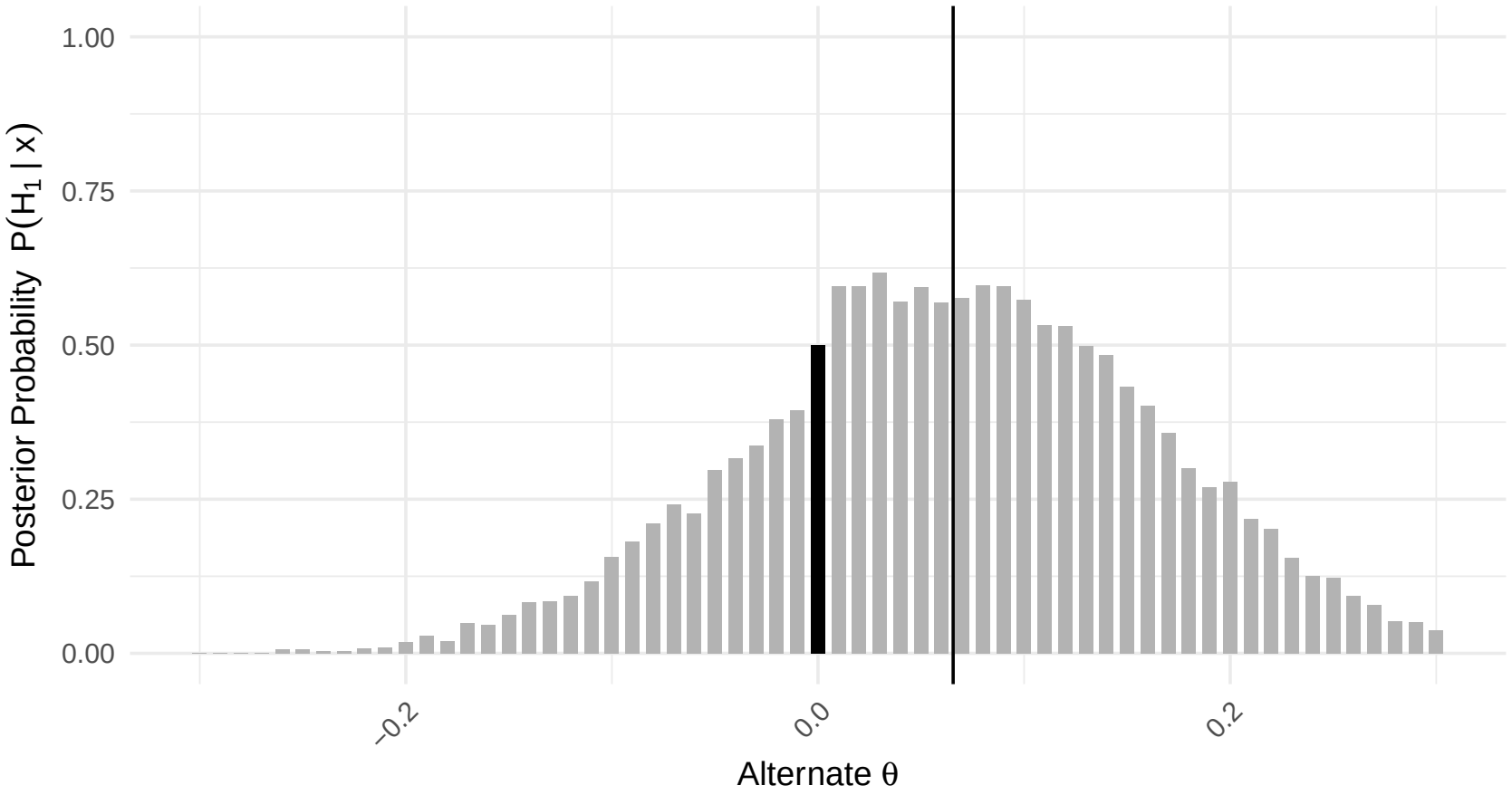


Figure: Proportion of trials in which alternate model H_1 is selected as alternate mean θ_1 varies from -0.3 to 0.3. The observed sample mean $\bar{x}$ is denoted by the vertical line, while the baseline of $\theta_0 = \theta_1 = 0$ is denoted by the black bar.

Convergence

- It's important to understand the convergence properties of the model choice in this simplest setting
- For $k \in \{0, 1\}$, define the likelihood as

$$L_m(k) = \prod_{i=1}^{m} \frac{\mathcal{N}(x_i \mid \hat{\theta}_k, 1)}{\mathcal{N}(x_i \mid \hat{\theta}_{(m-1)}, 1)}$$

where $\hat{\theta}_{(m-1)}$ maximizes $\prod_{i=1:m-1} \mathcal{N}(x_i \mid \theta, 1)$ with $\theta \in \{\theta_0, \theta_1\}$. Then

$$E\left(L_m(k) \mid x_{1:m-1}\right) = \prod_{i=1}^{m-1} \frac{\mathcal{N}(x_i \mid \hat{\theta}_k, 1)}{\mathcal{N}(x_i \mid \hat{\theta}_{(m-1)}, 1)} \leq 1,$$

and also $E(L_m(k) \mid x_{1:m-1}) \geq L_{m-1}(k)$

- Hence, for both $k \in \{0, 1\}$ it is that $L_m(k)$ converges due to the martingale convergence theorem
- The hypothesis selected maximizing $L_\infty(k)$; in the limit, $L_\infty(k)$ will either be 0 or 1

Interpreting Summary Statistics

- $\bar{x}$ drives the model choice
- Partition dataset space via predictive distribution
- Links model uncertainty to decision boundaries in observation space at $x_{1:\infty}$, given $x_{1:n}$
- In contrast to p-values, predictive resampling can provide a direct probability both for and against a null hypothesis without assigning a prior probability over the hypothesis space using just the observed sample

Consistent Model Selection

- The approach requires the specification of a model selection criterion, which will be used to select the best model at each step for the generation of x_{n+1} given observed $x_{1:n}$, and ultimately to select the best final model for each possible realization of the complete data
- The key requirement for this criterion is therefore that it be consistent, i.e. that the probability of selecting the correct model converges to 1 as $N \to \infty$ (Claeskens and Hjort, 2008).
- BIC is consistent and interpretable
- BIC also has a Predictive and Bayesian justification

Cross-Validation Pitfalls

- One might be tempted by cross-validation procedures such as Leave-One-Out (LOO) or AIC
- However, LOO-CV and AIC are inconsistent Shao (1993)
  - from a predictive perspective it seems odd to use an inconsistent selection criteria
- Leave-p-out CV better but expensive
- It has also been highlighted that Bayesian LOO-CV can be unreliable Vehtari et al. (2017) and Sivula et al. (2023)

Convergence under more general models

- Given $x_{1:m}$, let $\mathcal{M}_{\widehat{k}(m)}$ be the optimal candidate model under a consistent model selection criterion, with any necessary parameter MLE(s) denoted by $\widehat{\theta}_{\widehat{k}(m)}$
- We sample x_{m+1} from the predictive $p(\cdot \mid \mathcal{M}_{\widehat{k}(m)}, \widehat{\theta}_{\widehat{k}(m)})$, append it to our dataset yielding $x_{1:m+1}$, and repeat this process.
- As before, a key requirement is that the model choice $\widehat{k}(m)$ converges to some $k(\infty)$ as the sample size grows from $m \to \infty$
- We demonstrate this first for the case where the intermediate models are selected by maximum likelihood, and then for the case where this likelihood is penalized (as in the AIC, BIC, or LASSO)

Convergence for penalized likelihood

Proposition

The model $\mathcal{M}_{\widehat{k}(m)}$ selected by sequential penalized maximum likelihood converges as $m \to \infty$.

Proof.

We now consider

$$L_m(k, \theta_{k(m)}) = c(m, d_k, \theta_{k(m)}) \prod_{i=1}^{m} \frac{p(x_i \mid \mathcal{M}_{k(m)}, \theta_{k(m)})}{p(x_i \mid \mathcal{M}_{\widehat{k}(m-1)}, \widehat{\theta}_{\widehat{k}(m-1)})},$$

and

$$E(L_m(k, \theta_{k(m)}) \mid x_{1:m-1}) = L_{m-1}(k, \theta_{k(m)}) \frac{c(m, d_k, \theta_{k(m)})}{c(m-1, d_k, \theta_{k(m-1)})}.$$

where c is a penalty function

- c is a penalty function in terms of the sample size m, model dimension d, and parameter θ. For AIC, $c(m, d, \theta) = e^{-d}$; for BIC, $c(m, d, \theta) = e^{-d \log m}$; and for Lasso, $c(m, d, \theta) = e^{-\lambda_m |\theta|}$, for some increasing $\lambda_m > 0$; $c = 1$ for ML model
- This remains a supermartingale when c decreases as m increases, which is the case for the key model selection criteria listed
- Hence, for each $(k, \theta_{k(m)})$, we have $L_m(k, \theta_{k(m)}) \to L_\infty(k, \theta_{k(\infty)})$ almost surely for some L_∞
- To extend this to uniform convergence, similar model conditions as for the convergence of an MLE are required, namely that each Θ_k is a compact space and each $p(x \mid \mathcal{M}_{k(m)}, \theta_{k(m)})$ is suitably regular
- BIC is both consistent and converges

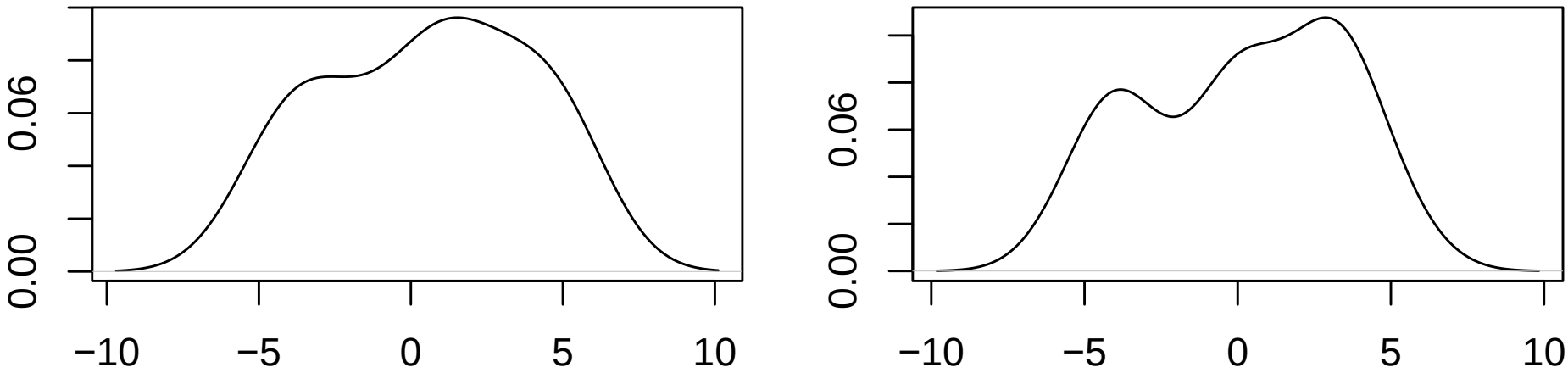
Illustration: Density Estimation

- Code for all illustrations can be found at `https://github.com/vshirvaikar/MPModel`
- A typical model selection question is the number of components required in a finite Gaussian mixture model (GMM) for density estimation
- We generate $n = 20$ and $n = 50$ data points from a GMM with $G = 3$ components

$$f_0(y) = 0.4\mathcal{N}(y \mid -3, 1) + 0.3\mathcal{N}(y \mid 0, 1) + 0.3\mathcal{N}(y \mid 4, 1)$$

where the goal is to identify and return uncertainty around the true value of G

TBC



(a) Density for $n = 20$ observations

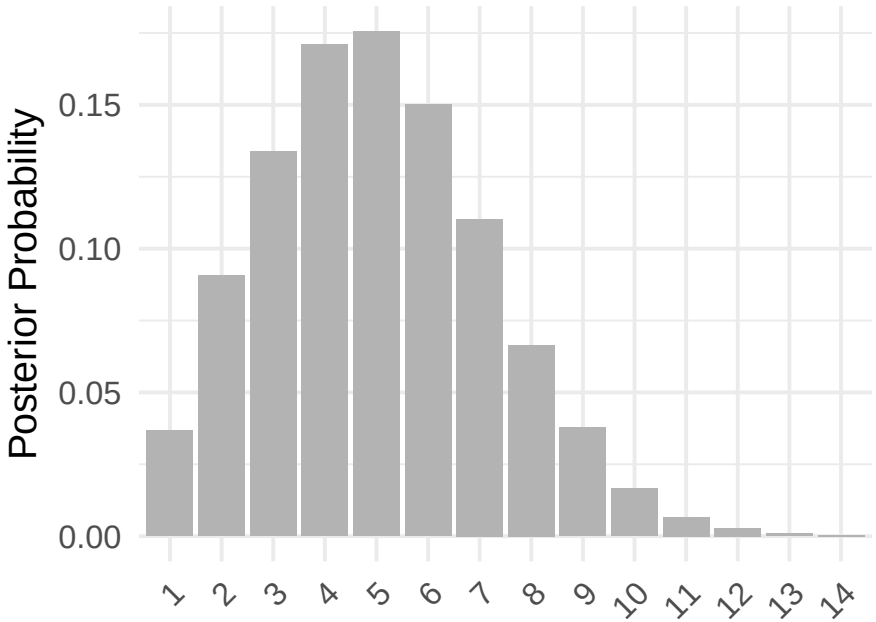
(b) Density for $n = 50$ observations

Figure: Kernel density plots for data generated from GMM with 3 components

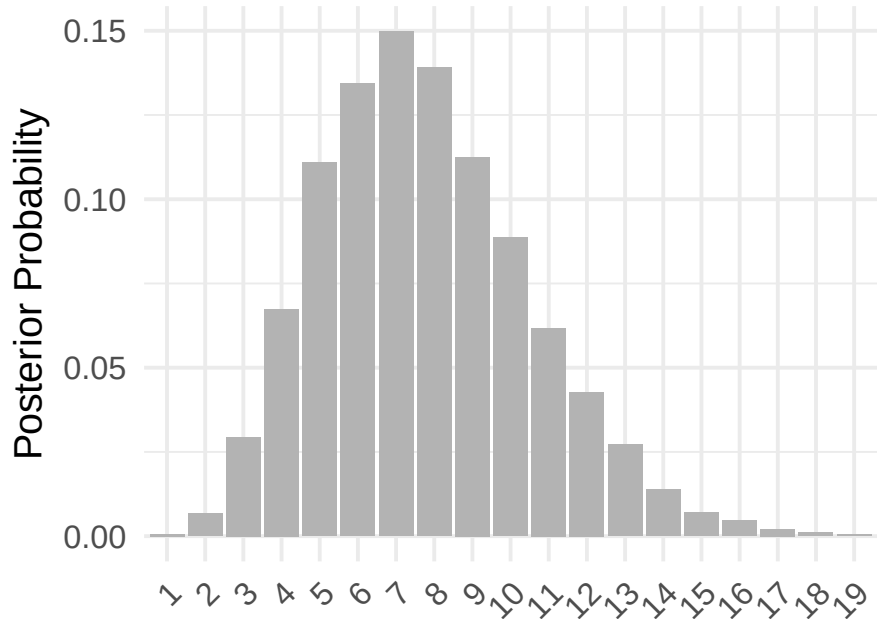
Density estimation

- We implement a DPMM with the **dirichletprocess** package in R using the default prior and hyperparameters (Ross and Markwick, 2019)
- For eight separate Gibbs sampling chains, we discard the first 500 iterations as burn-in and retain the next 2,000 iterations

DPMM



(a) Components for $n = 20$ observations



(b) Components for $n = 50$ observations

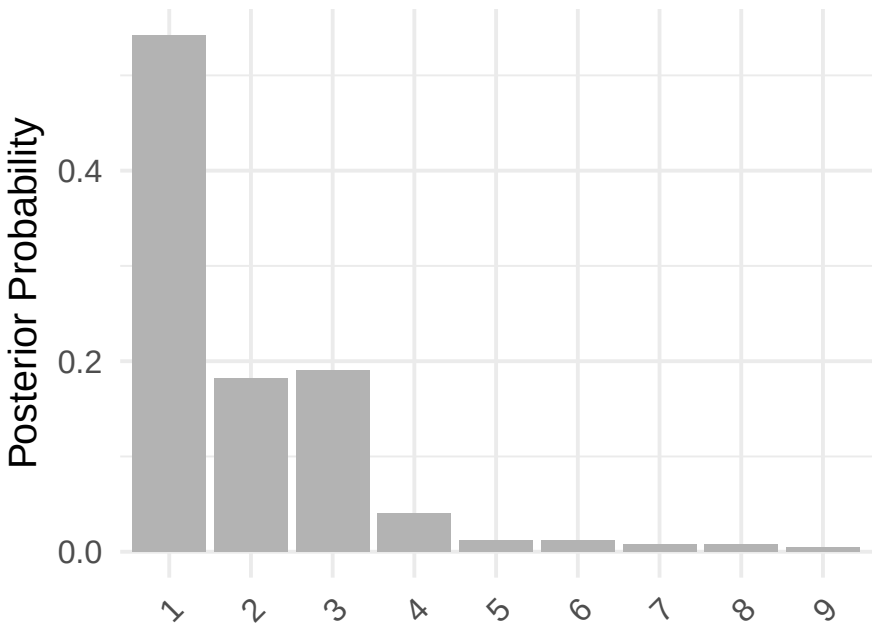
Figure: Posterior uncertainty over number of components G sampled in DPMM

- For the $n = 20$ case (Figure 4a), the mode is 5 components, and for $n = 50$ (Figure 4b) the mode is 7 components, both more complex than the "true" $G = 3$
- This reflects the known result that DPMM should not be used to estimate the number of components, which asymptotically tends towards infinity as n increases (Yang et al., 2019; Cai et al., 2020).

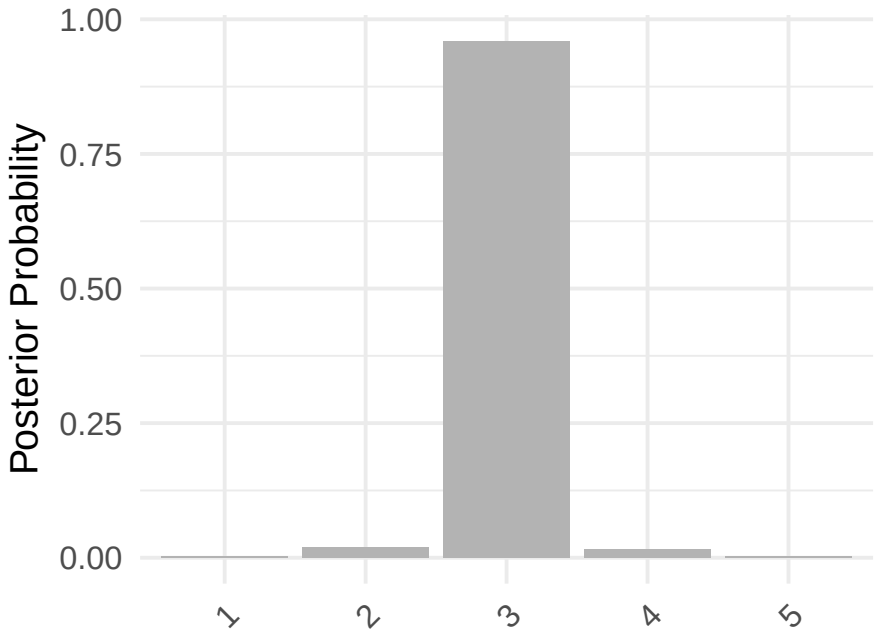
Predictive Resampling

- For the resampling approach, we implement EM clustering with the **mclust** package in R with specified candidate models ranging from 1 to 9 components
- Models with both equal and unequal variances are tested, with differing dimension penalties in the BIC calculation
- Simple to apply our approach as a wrapper around existing software – turning a model selection procedure into a posterior distribution on the model space
- We recursively simulate $N = n + 600$ new observations per trial across a total of $B = 400$ trials; this value of N is again empirically found to be sufficiently large that $\mathcal{M}_{k(N)}$ closely approximates $\mathcal{M}_{k(\infty)}$, with convergence diagrams provided in the paper

Resampling approach



(a) Components for $n = 20$ observations



(b) Components for $n = 50$ observations

Figure: Posterior uncertainty over number of components G via resampling

Conclusions

- We view model uncertainty through the lens of missing information
- With a consistent model selection criterion in hand, we would be able to reliably identify the correct model if we had the complete data
- The imputation of missing observations converts a generative model selection criterion directly into probabilities over the space of candidate models
- The approach serves as a form of model expansion around the initial best model for the observed data, as discussed by Draper (1995), and also echoes the prequential argument of Dawid (1984) with its focus on step-by-step prediction as the fundamental object of statistical modeling

Conclusions II

- The approach is data driven (objective), doesn't use priors, and avoids some of the sensitivities of Bayes Factors
- The approach is Bayesian, in that it provides a probabilistic measure of model uncertainty directly on the model space conditional on the observations $x_{1:n}$
- The method requires rapid optimization over the model space to be efficient

Paper with further details: `Shirvaikar, V., Walker, S. G., Holmes, C. (2024). A general framework for probabilistic model uncertainty. arXiv preprint arXiv:2410.17108.`

References I

Berger, J. O. and Pericchi, L. R. (1996) The Intrinsic Bayes Factor for Model Selection and Prediction. *Journal of the American Statistical Association*, **91**, 109–122.

Cai, D., Campbell, T. and Broderick, T. (2020) Finite mixture models do not reliably learn the number of components.

Claeskens, G. and Hjort, N. L. (2008) *Model selection and model averaging*.

Dawid, A. P. (1984) Present Position and Potential Developments: Some Personal Views: Statistical Theory: The Prequential Approach. *Journal of the Royal Statistical Society. Series A (General)*, **147**, 278–292.

Draper, D. (1995) Assessment and Propagation of Model Uncertainty. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, **57**, 45–70.

References II

de Finetti, B. (1937) La prevision : ses lois logiques, ses sources subjectives. *Annales de l'institut Henri Poincaré*, **7**, 1–68.

Kass, R. E. and Raftery, A. E. (1995) Bayes Factors. *Journal of the American Statistical Association*, **90**, 773–795.

O'Hagan, A. (1995) Fractional Bayes Factors for Model Comparison. *Journal of the Royal Statistical Society. Series B (Methodological)*, **57**, 99–138.

Ross, G. J. and Markwick, D. (2019) dirichletprocess: An R Package for Fitting Complex Bayesian Nonparametric Models.

Shao, J. (1993) Linear Model Selection by Cross-validation. *Journal of the American Statistical Association*, **88**, 486–494.

Sivula, T., Magnusson, M., Matamoros, A. A. and Vehtari, A. (2023) Uncertainty in Bayesian Leave-One-Out Cross-Validation Based Model Comparison.

References III

Vehtari, A., Gelman, A. and Gabry, J. (2017) Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, **27**, 1413–1432.

Yang, C.-Y., Xia, E., Ho, N. and Jordan, M. I. (2019) Posterior Distribution for the Number of Clusters in Dirichlet Process Mixture Models.