Calibrated Bayes Factors for Model Comparison

Xinyi Xu
The Ohio State University

Joint work with Steve MacEachern, Pingbo Lu and Ruoxi Xu

April 9, 2019

Outline

How complex do the models need to be?

- The advent of automated technology for data collection and of cheap storage capacity provides access to a previously undreamt of wealth of data. The parallel development of computational horsepower enables us to fit quite sophisticated models.
- Should we always fit the complex nonparametric models whenever possible?
- Many evidences have shown that although nonparametric models provide great flexibility for modeling, they can also lead to large estimation variances.

Bayes factors

- The Bayes factor is one of the most important and most widely used tools for Bayesian hypothesis testing and model comparison.
- Given two models M_1 and M_2, we have

$$B_{M_1,M_2} = \frac{m(x;M_1)}{m(x;M_2)} = \frac{\int f_1(x|\theta_1)\pi_1(\theta_1)d\theta_1}{\int f_2(x|\theta_2)\pi_2(\theta_2)d\theta_2},$$

  which can be viewed as a “weighted" likelihood ratio of M_1 to M_2.

Bayes factors (Cont.)

- The value of a Bayes factor reflects the *relative* evidence for the two models in the data
- Jeffreys' rule of thumb of interpreting Bayes factors (Jeffreys 1961, app. B)
  - $1 <$ Bayes factor ≤ 3: weak evidence for M_1
  - $3 <$ Bayes factor ≤ 10: substantial evidence for M_1
  - $10 <$ Bayes factor ≤ 100: strong evidence for M_1
  - $100 <$ Bayes factor: decisive evidence for M_1

Outline

Improper objective priors?

- In computing Bayes factors, prior elicitation of the model parameters can have a big impact, even with a large amount of data.
- As the number of parameters grows, careful subjective specification of priors for all the parameters is often precluded. Thus, it is necessary to resort to specifications of priors using some formal methods
- Improper “noninformative" priors are problematic, because they are determined only up to an arbitrary constant.

Proper diffuse priors?

- In practice, many use a standard proper but vague prior distribution, as famously illustrated in the BUGS manual.
- However, the Bayes factors can be very sensitive to the arbitrary diffuseness of the priors.
- **Example 1.** Suppose the $Y_i \,|\, \theta, \sigma^2 \overset{iid}{\sim} N(\theta, \sigma^2)$.

$$M_1 : \theta = 0$$

$$M_2 : \theta \in \mathcal{R}.$$

For each model, assume $\sigma^2 \sim IG(0.1, 0.1)$; and for M_2, further assume $\theta \sim N(0, \tau^2)$, where τ^2 is set to be 10, 100, 1000, and 10,000, respectively.

Proper diffuse priors? (Cont.)

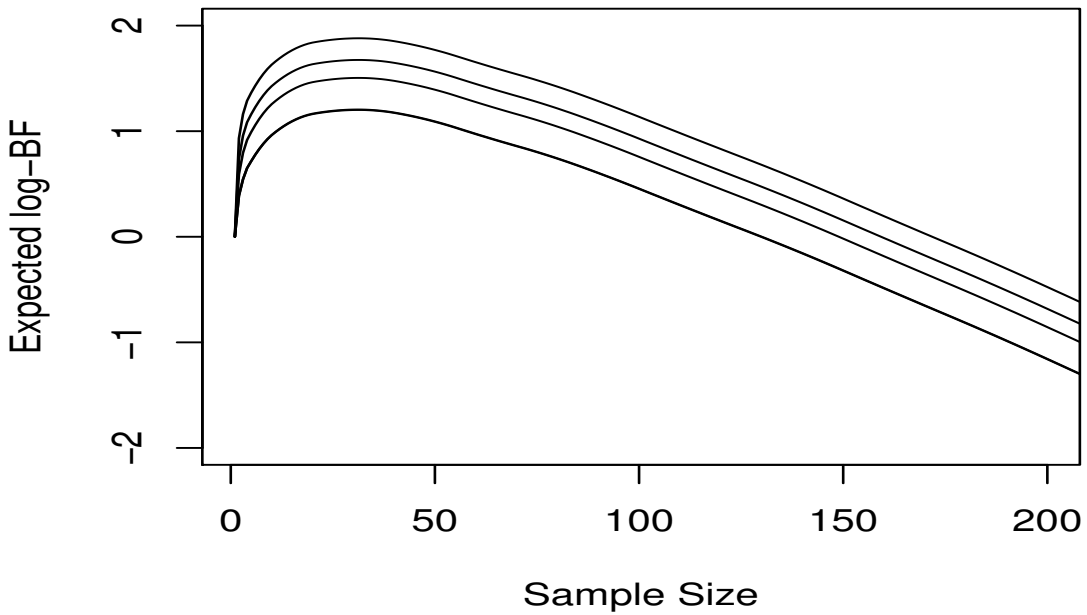


Figure: Set the true parameter $\theta = 1$. The expected log Bayes factor curves from the bottom to top are under the normal priors with variances 10, 100, 1000 and 10000. For each sample size, we simulate 300 data sets and compute the Bayes factors based on 5,000 MCMC iterations following a burn-in of 500 iterations.

A connection with predictive distributions

- An important relationship between the Bayes factor and a sequence of predictive distributions:

$$\log(B_{M_1,M_2}) = \log \frac{m(y_{1:n};M_1)}{m(y_{1:n};M_2)} = \sum_{i=1}^{n-1} \log \frac{m(Y_{i+1} \mid Y_{1:i}, M_1)}{m(Y_{i+1} \mid Y_{1:i}, M_2)}.$$

- Under weak priors, when the sample size is small, the simple model tend to yield better predictive performance. When the sample size is large enough, the complex model can often provide a better approximation to the true model because of its larger support.
- it is worth noting that at moderate sample sizes, the complex model performs better and the log Bayes factor is on a decreasing trend. But the log Bayes factor is still positive and sizable because of the big lead built up by the simple model at the beginning. The size of the lead is determined by the arbitrary diffuseness levels of the priors.

Nonparametric priors

Example 2. Suppose that we have $n = 176$ i.i.d. observations from a *skew-normal(location=0, scale=1.5, shape=2.5)*. Compare a Gaussian parametric model M_1:

$$Y_i \,|\, \theta, \sigma^2 \overset{iid}{\sim} N(\theta, \sigma^2), \qquad \theta \sim N(\mu, \tau^2)$$

with a Mixture of Dirichlet Processes (MDP) nonpara model M_2:

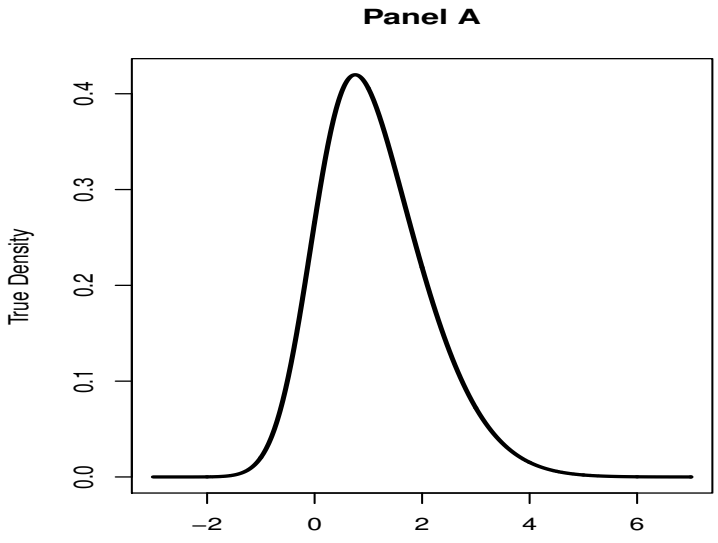
$$Y_i \,|\, \theta_i, \sigma^2 \overset{ind}{\sim} N(\theta_i, \sigma^2), \qquad \theta_i \,|\, G \overset{iid}{\sim} G, \qquad G \sim DP(M = 2, N(\mu, \tau^2))$$

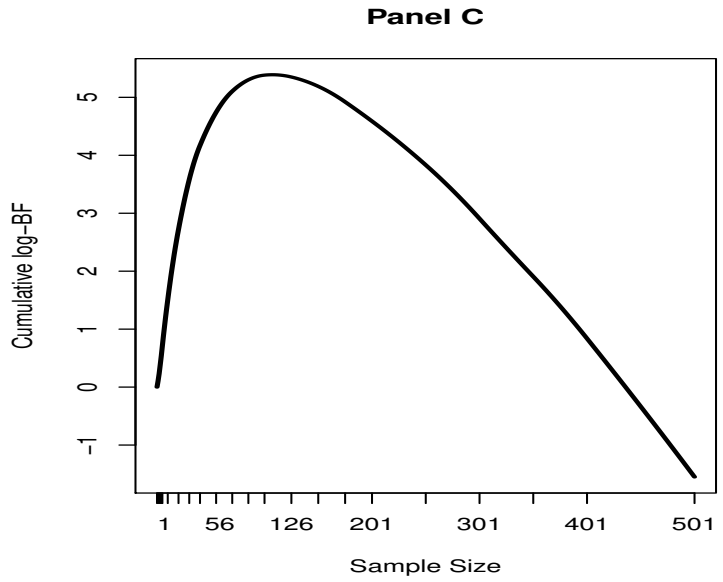
Common priors are placed on hyper-parameters:

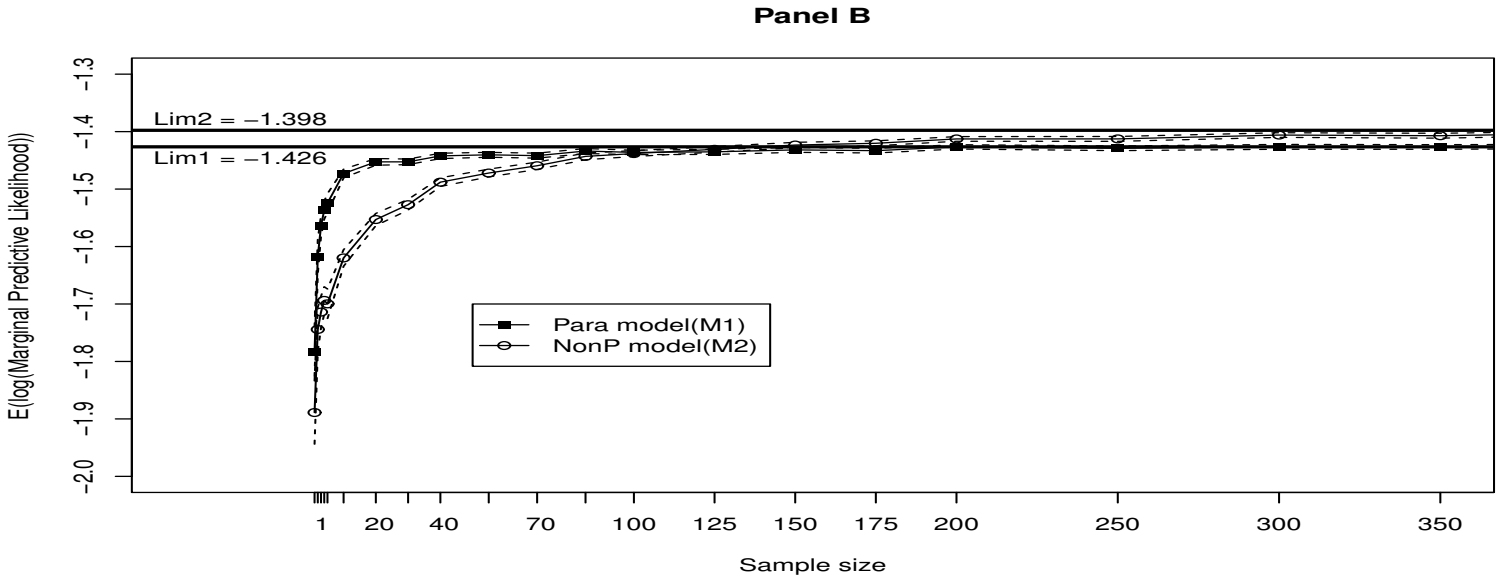
$$\mu \sim N(0, 500), \qquad \sigma^2 \sim IG(7, 0.3), \qquad \tau^2 \sim IG(11, 9.5)$$

We can envision that with sufficiently large sample size, the MDP nonparametric model would outperform the Gaussian model.

Nonparametric priors (Cont.)







Barlett's paradox

- The above phenomenon is closely related to Barlett's paradox (1957), which is a situation where the Bayesian and frequentist approaches to a hypothesis testing problem give opposite results for certain choices of the prior.
- Suppose that in testing $H_0 : \theta = \theta_0$, the data estimate $\hat{\theta}_x$ is far from θ_0, then standard sampling theory suggest to reject H_0
- However, if we use a "noninformative" prior with very large spread on θ under H_a, since this prior assigns little mass around the true parameter value, the data could be even more unlikely under H_a, and thus the Bayes factor favors H_0.
- The fundamental reason underlying the paradox: the prior for the complex model is much more diffuse than that for the simple model

Barlett's paradox (Cont.)

- This phenomenon can be more pronounced in model comparisons involving nonparametric models, where it is usually difficult to evaluate the diffuseness of priors.
- As Jeffreys recognized, to avoid the problem of Bayes factors, priors on model parameters must be proper and not have too big a spread.
- Where should we make the prior on nonparametric models be centered at, that is, where to shrink toward? And how to measure the diffuseness level of nonparametric priors?

Data-dependent shrinkage priors

- We develop data-dependent shrinkage priors, by
  - using part of the data as training samples to update the priors, so that the partial-posteriors have a reasonable level of concentration around good nonparametric models
  - and then computing the Bayes factor based on the remainder of the data.
- The use of training samples has been studied for Bayesian model selection, e.g.
  - Lempers (1970)
  - Berger and Pericchi (1995, 1996): intrinsic Bayes factor
  - O' Hagan (1995): fractional Bayes factor
- There are two main differences between our approach and existing methods:
  - Starting point: we don't require the initial priors to be improper or the models to be parametric
  - Destination: our goal is not only to obtain proper priors, but also to obtain reasonably concentrated priors

Outline

Information metric

- Suppose that $y_i \mid \theta \overset{iid}{\sim} f_\theta$ and $\theta \sim \pi$. It can be hard to measure the amount of information contained in a general prior π.
- Possible approaches:
  - Fisher information?
  - Variance?
  - Effective sample size?
  - The distance between two distributions f_{θ_1} and f_{θ_2}, where θ_1 and θ_2 are two random draws from π

Information metric (Cont.)

- We measure the closeness between f_{θ_1} and f_{θ_2} under the Symmetric - Kullback-Leibler (SKL) divergence

$$SKL(f_{\theta_1}, f_{\theta_2}) = \frac{1}{2}\left[E^{\theta_1}\log\frac{f_{\theta_1}}{f_{\theta_2}} + E^{\theta_2}\log\frac{f_{\theta_2}}{f_{\theta_1}}\right].$$

- The distribution on (θ_1, θ_2) induces a distribution on $SKL(f_{\theta_1}, f_{\theta_2})$
- We evaluate the information contained in π using the percentiles of this distribution in SKL divergence.

Target information level

- To calibrate the Bayes factor and select a training sample size, we need to choose a benchmark prior and then require the updated priors to contain at least as much information as this benchmark prior.

- In order to perform a reasonable analysis where subjective input has little impact on the final conclusion, we set the benchmark to be a “minimally informative" prior – the unit information prior (Kass and Wasserman 1995), which contains the amount of (Fisher) information as that in one observation

- It is easy to verify that under the Gaussian model $Y \sim N(\theta, \sigma^2)$, a unit information prior on θ is $N(0, \sigma^2)$, under which $SKL(f_{\theta_1}, f_{\theta_2})$ follows a χ^2_1 distribution

Constructing data-dependent shrinkage priors

- **Step 1:** Randomly draw a training sample with a pre-specified sample size from the data
- **Step 2:** Update the prior based on this training sample. Take M pairs of (θ_1^j, θ_2^j), where $j = 1, \cdots, M$, and compute $SKL(f_{\theta_1^j}, f_{\theta_2^j})$ based on each pair
- **Step 3:** Repeat Steps 1 and 2 for N times. Pool all MN values of the SKLs to evaluate the information in the posterior
- **Step 4:** Compare the information amount in the posterior with that in the benchmark distribution. If the information amount is comparable, terminate the search and report the current sample size as the calibration sample size. Otherwise reset the sample size and repeat Steps 1 to 4.

Calibrated Bayes factor (CBF)

- The calibration sample size might be larger than the sample size n when n is too small or the initial prior is very diffuse and the model dimension is very large.
- Let s_1 and s_2 represent the calibration sample sizes for models M_1 and M_2 and assume $n > s = \max(s_1, s_2)$. Based on a training sample $X_{(s)}$, the updated Bayes factor is

$$\log B_{12}^*(x \,|\, x_{(s)}) = \log B_{12}(x) - \log B_{12}(x_{(s)}),$$

- Let $\{x_{(s)}^1, x_{(s)}^2, \cdots, x_{(s)}^H\}$ denote all possible subsets of x of size s. Then the calibrated Bayes factor (CBF) is defined by

$$\log CB_{12}(x) = \log B_{12}(x) - \frac{1}{H}\sum_{h=1}^{H} \log B_{12}(x_{(s)}^h).$$

Therefore, it shares the same consistency properties as the original Bayes factor.

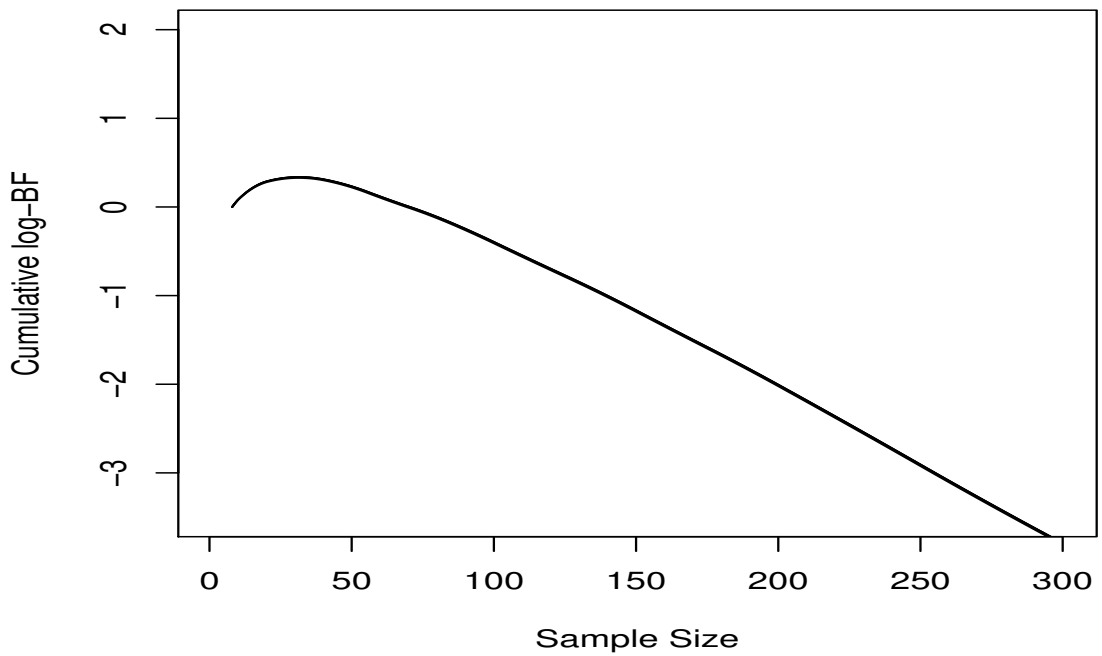
Calibrated Bayes factors (Cont.)

- **Proposition.** Under a set of regularity conditions that ensure the posterior concentration in a neighborhood of the “best fitting" model (see Berk 1966), the calibration sample size is finite. Moreover, as the sample size $n \to \infty$, the CBF is consistent.

- When the models are parametric and the initial priors are improper, the CBF provides similar results as Berger’s intrinsic Bayes factors. Furthermore, it is applicable in situations where the models are nonparametric or the initial priors are proper but over-dispersed.

Revisiting the examples

Example 1 revisited: For the normal mean test, all four diffuse priors (with τ^2 = 10, 100, 1000, and 10000) achieve the “unit information" level with calibration sample size 7. The four CBF curves are almost on top of one another.



Revisiting the examples (Cont.)

- **Example 2 revisited:** For the parametric versus nonparametric density estimation, our search leads to a calibration sample size of 50.
- The peak of the CBF is 2.34 in favor of the parametric model, which is not worth more than a bare mention under Jeffrey's criterion. At the full sample size 350, the CBF is 13.62 in favor of the nonparametric MDP model.
- These CBF are consistent with the posterior predictive performances.

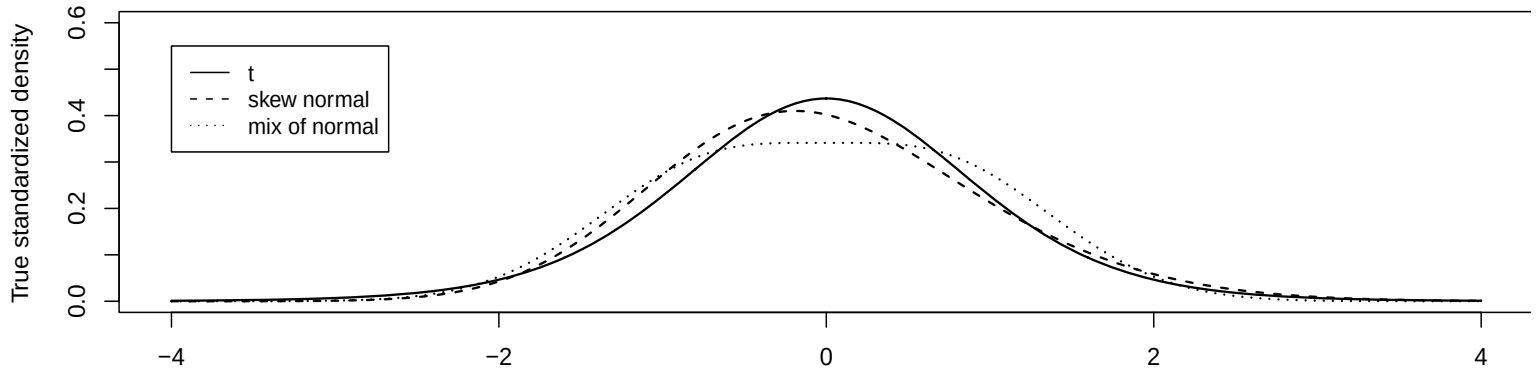
Outline

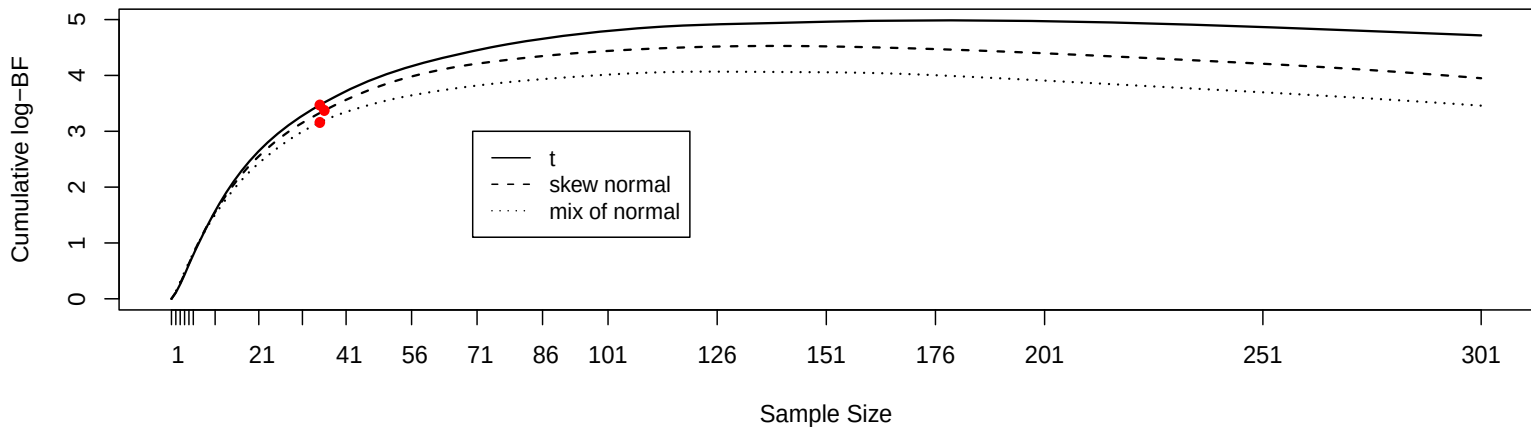
The simulation setup

- To investigate the patterns of log Bayes factors and to illustrate the effect of calibration, we compare the Gaussian parametric model to the MDP model under the following distributions with various shapes:
  - Skew-normal with varying shape parameter α (skewness)
  - Student-t with varying degrees of freedom ν (thick-tails)
  - Symmetric mixture of normals with varying component means $\pm\delta$ (Bimodality)
- In all cases, the distributions have been centered and scaled to have mean 0 and standard deviation 1.
- By specifying α, ν and δ, we tune the KL distances from the true distributions to the best fitting Gaussian distributions.

Simulation results

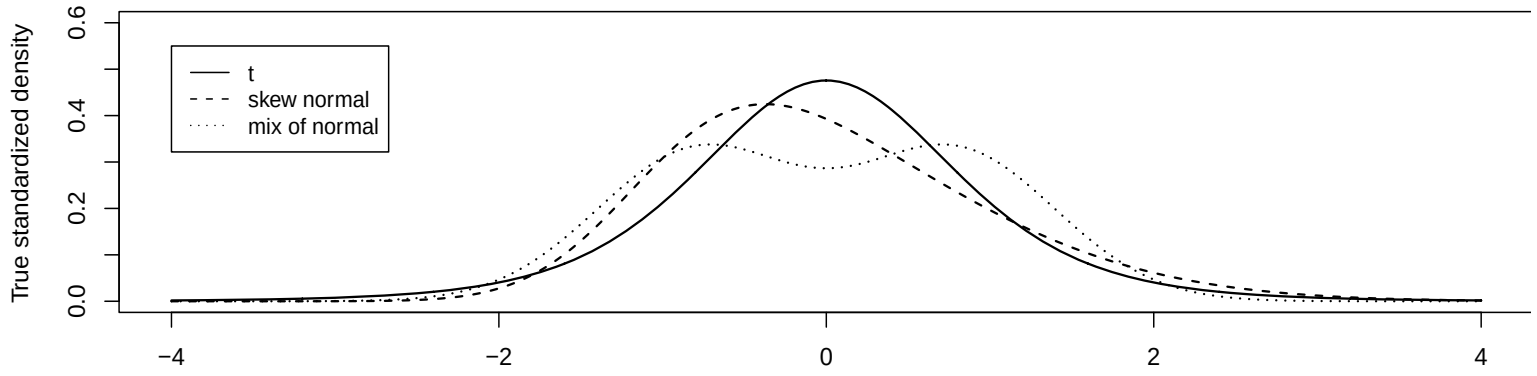
- Small divergences from the Gaussian

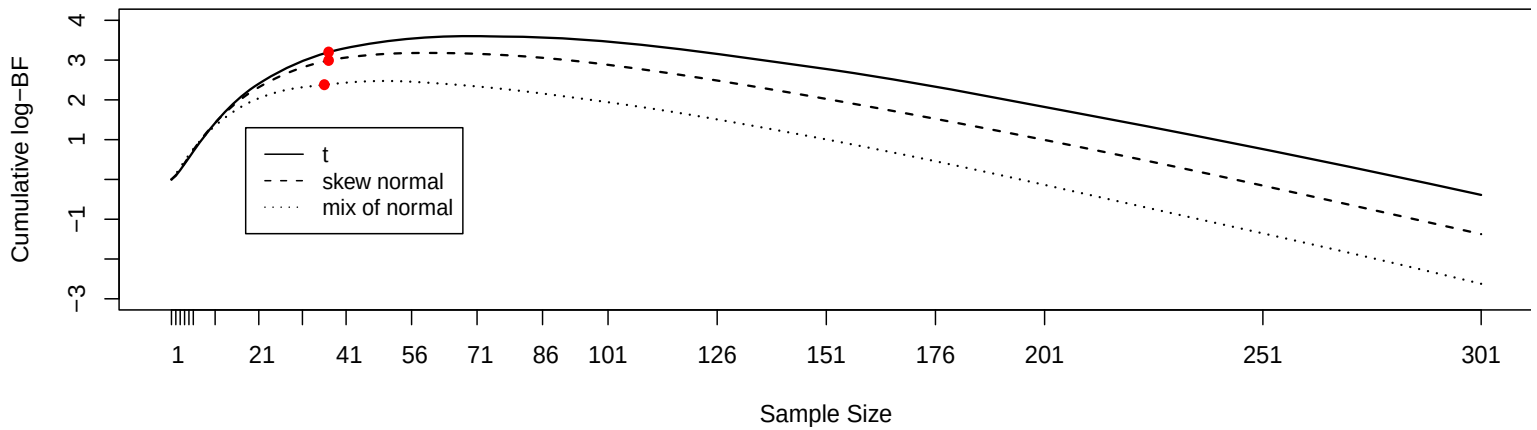




Simulation results (Cont.)

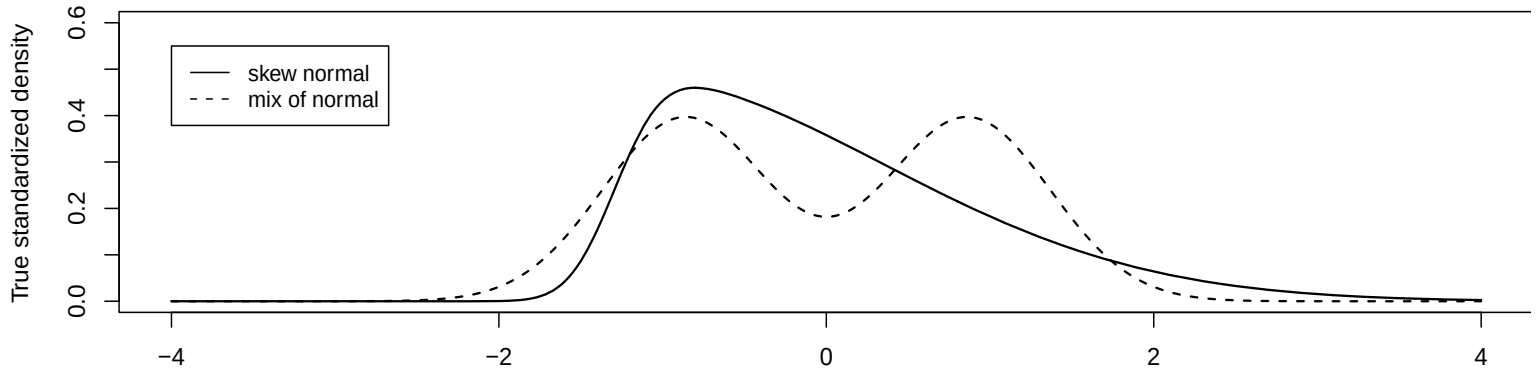
- Moderate divergences from the Gaussian

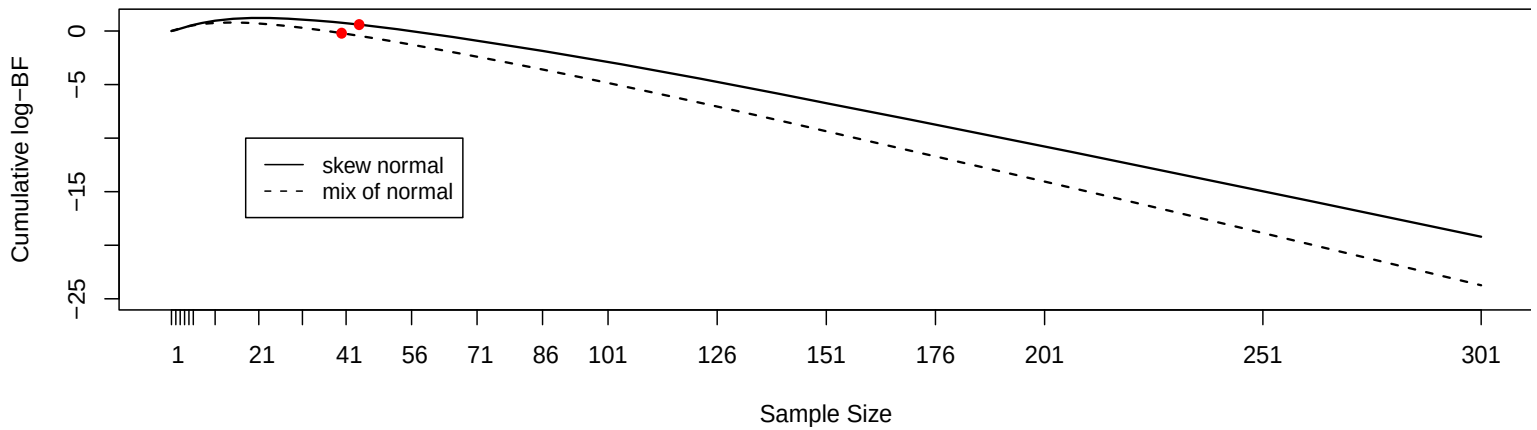




Simulation results (Cont.)

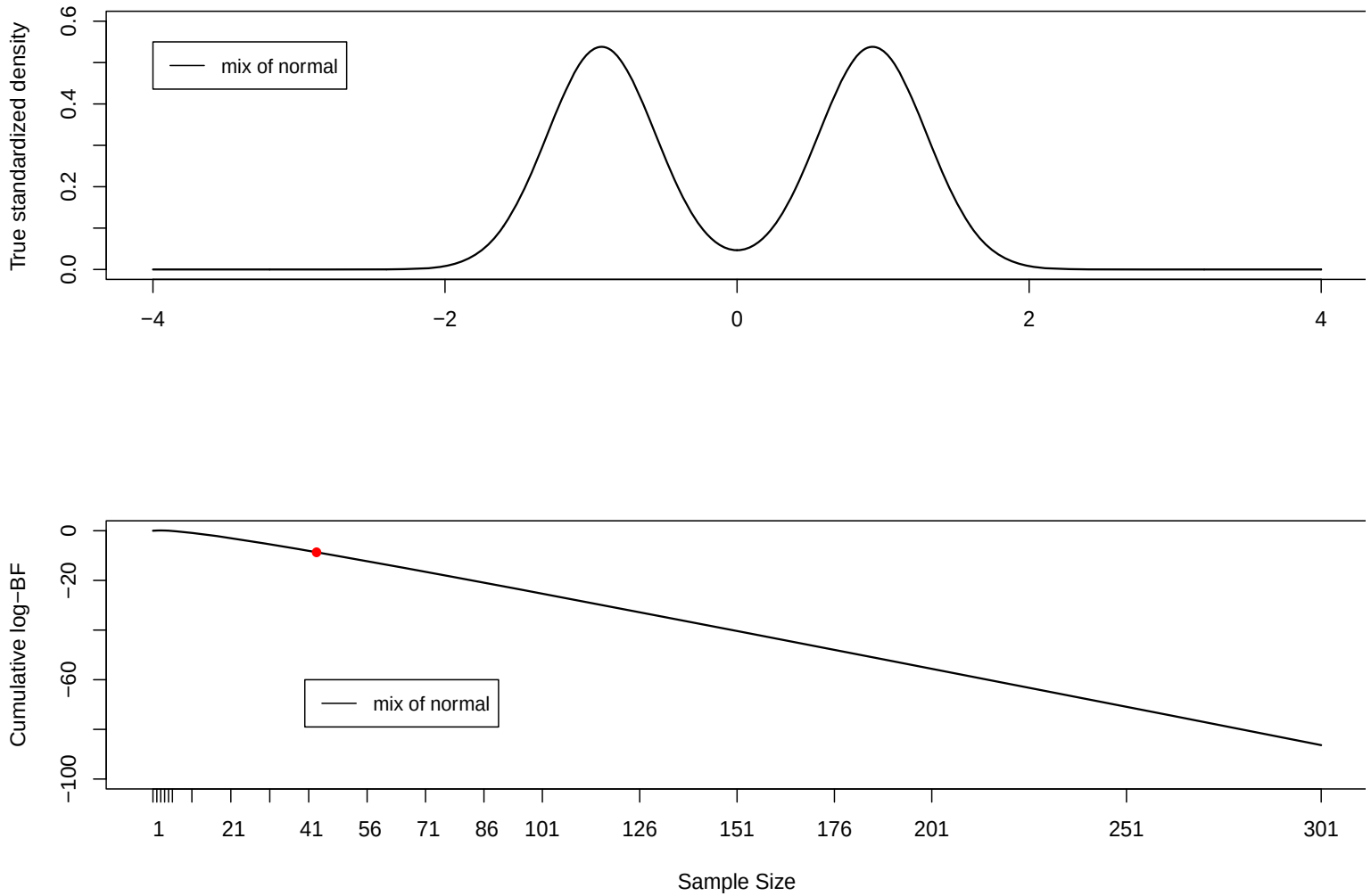
- Large divergences from the Gaussian





Simulation results (Cont.)

- Very large divergences from the Gaussian



Simulation results summary

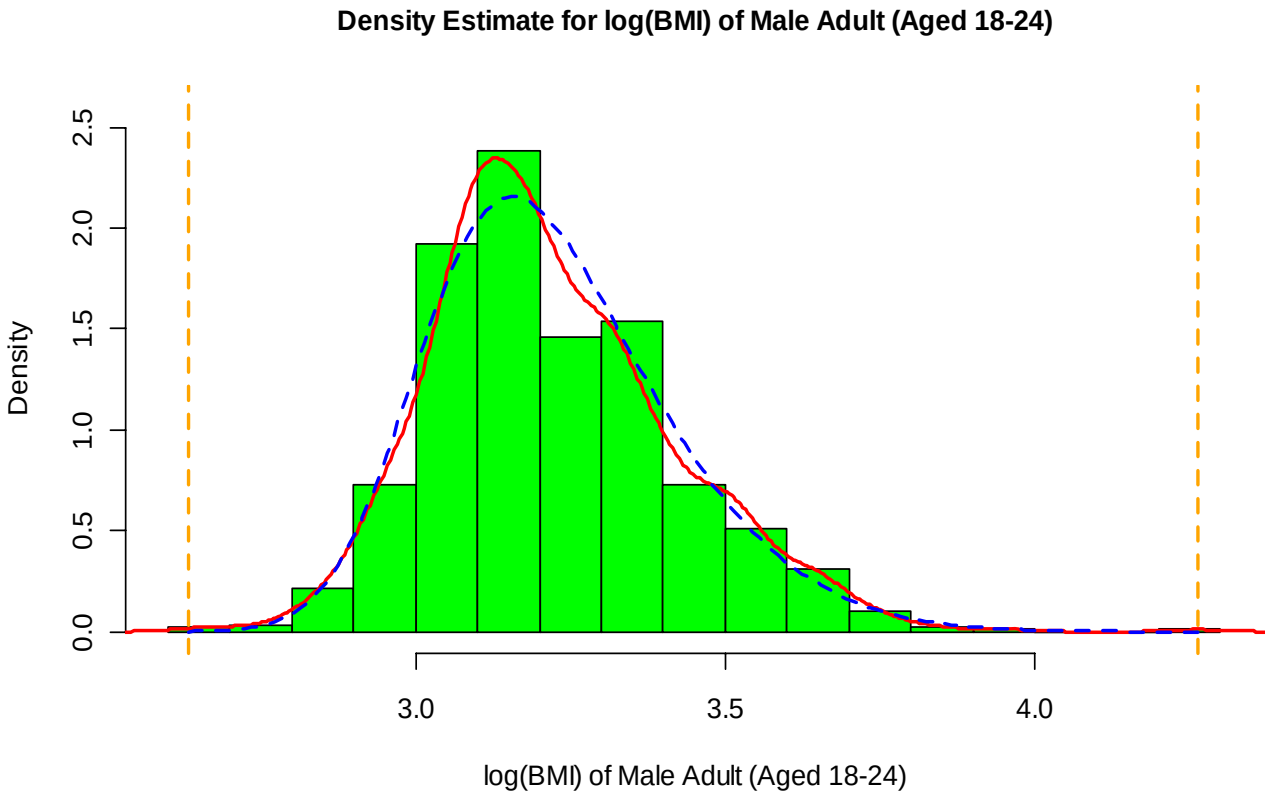
- In all cases, the calibration is driven by the MDP model rather than the Gaussian model (which is typically calibrated after two or three observations)
- In all cases, the peaks of the log calibrated Bayes factors remain below two, leading to better agreement between the Bayes factor and the models' predictive performances.
- In the same scenario (the same KL divergence from the true distribution to the best fitting Gaussian distribution), the calibration sample size varies little
- Across different scenarios, the further the underlying true distribution is from normality, the larger the calibration sample size will be.

Model comparisons in OFHS analysis

- The OFHS (Ohio Family Health Survey) was conducted between August 2008 and January 2009 to study the health insurance coverage for the people in Ohio.
- An important health measurement in this survey is the BMI (Body Mass Index).
- We focus on the subpopulation consists of male adult aged between 18 and 24. There are 895 non-missing BMI values in this group.

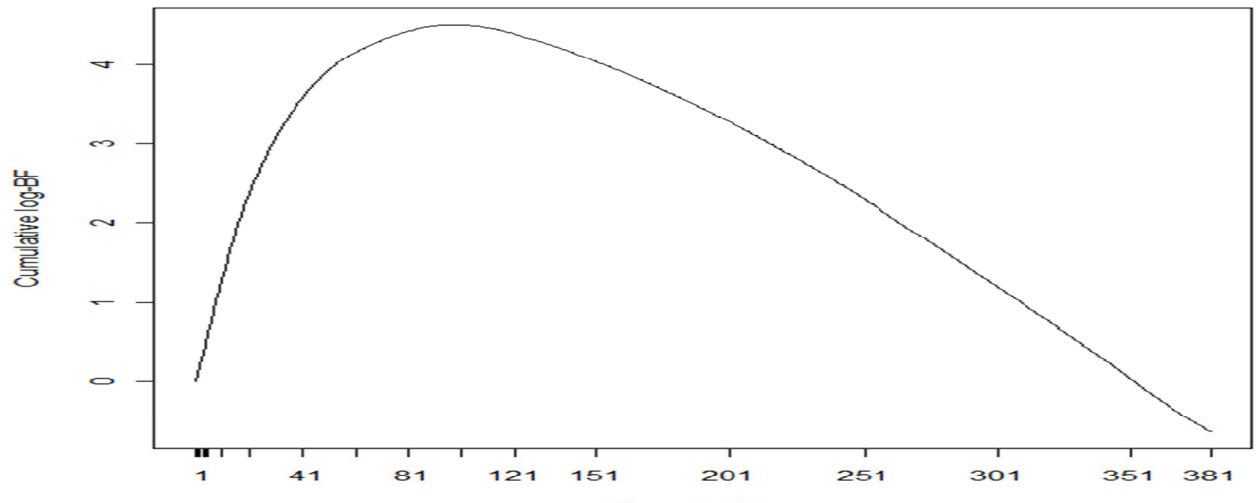
Model comparisons in OFHS analysis (Cont.)

- The log transformed data is close to a skew-normal distribution with the skewness parameter $\hat{\alpha}_{MLE} = 2.41$.



Modle Comparison in OFHS analysis (Cont.)

- Based on the full data set, the log Bayes factor is -12.19, translating to a Bayes factor of $196,811$ favoring MDP.
- We further investigate the expected log Bayes factor for a range of smaller sample sizes. For each sample size, we generate 300 subsamples.
- If we only had a subset of the observations with size $n = 106$, the Bayes factor is $B_{P;NP} \approx e^{4.64} \approx 104$, which provides strong evidence for the Gaussian parametric model



Modle Comparison in OFHS analysis (Cont.)

- After matching the prior concentration with the unit information prior, we calibrate the priors using training samples with size 50.

- At the sample size $n = 106$, the calibrated Bayes factor is $CB_{P;NP} \approx e^{0.64} \approx 1.9$, which provides very weak model preference; at the full sample size, the eventual calibrated Bayes factor is $CB_{P;NP} \approx e^{-16.18}$, which leads to a Bayes factor of 10.6 million to one in favor of the MDP model.

- We find the swing from inconclusive evidence for modest sample sizes to conclusive evidence in favor of the MDP model for the full sample far more palatable than the swing from very strong evidence in one direction to conclusive evidence in the opposite direction.

Outline

Discussions

- The Bayes factor might be unreliable when the prior for one model is much more diffuse than that of the other.
- To make a fair comparison between small and large models, we calibrate the prior distributions using training samples, so that the partial posteriors achieve a reasonable level of concentration. These partial posteriors can be used as new data-dependent shrinkage prior for computing CBFs based on the remainder of the data.
- CBF can also be applied to model comparison among a group of models, and can generate more reliable posterior model weights for Bayesian model averaging.
- The implications of this work extend beyond parametric versus nonparametric model comparisons, instead it is widely applicable in small vs. large model comparisons.