# The Calibrated Bayes Factor for Model Comparison

Steve MacEachern
The Ohio State University

Joint work with Xinyi Xu, Pingbo Lu and Ruoxi Xu
Supported by the NSF and NSA

Bayesian Nonparametrics Workshop
ICERM 2012

# Outline

- The Bayes factor – when it works, and when it doesn't
- The calibrated Bayes factor
- Ohio Family Health Survey (OFHS) analysis
- Wrap-up

# Outline

# 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

$$BF = \frac{m(y; M_1)}{m(y; M_2)},$$

- Some rules of thumb for using Bayes factors (Jeffreys 1961)
  - $1 <$ Bayes factor $\leq 3$: weak evidence for $M_1$
  - $3 <$ Bayes factor $\leq 10$: substantial evidence for $M_1$
  - $10 <$ Bayes factor $\leq 100$: strong evidence for $M_1$
  - $100 <$ Bayes factor: decisive evidence for $M_1$

# Monotonicity and the Bayes factor

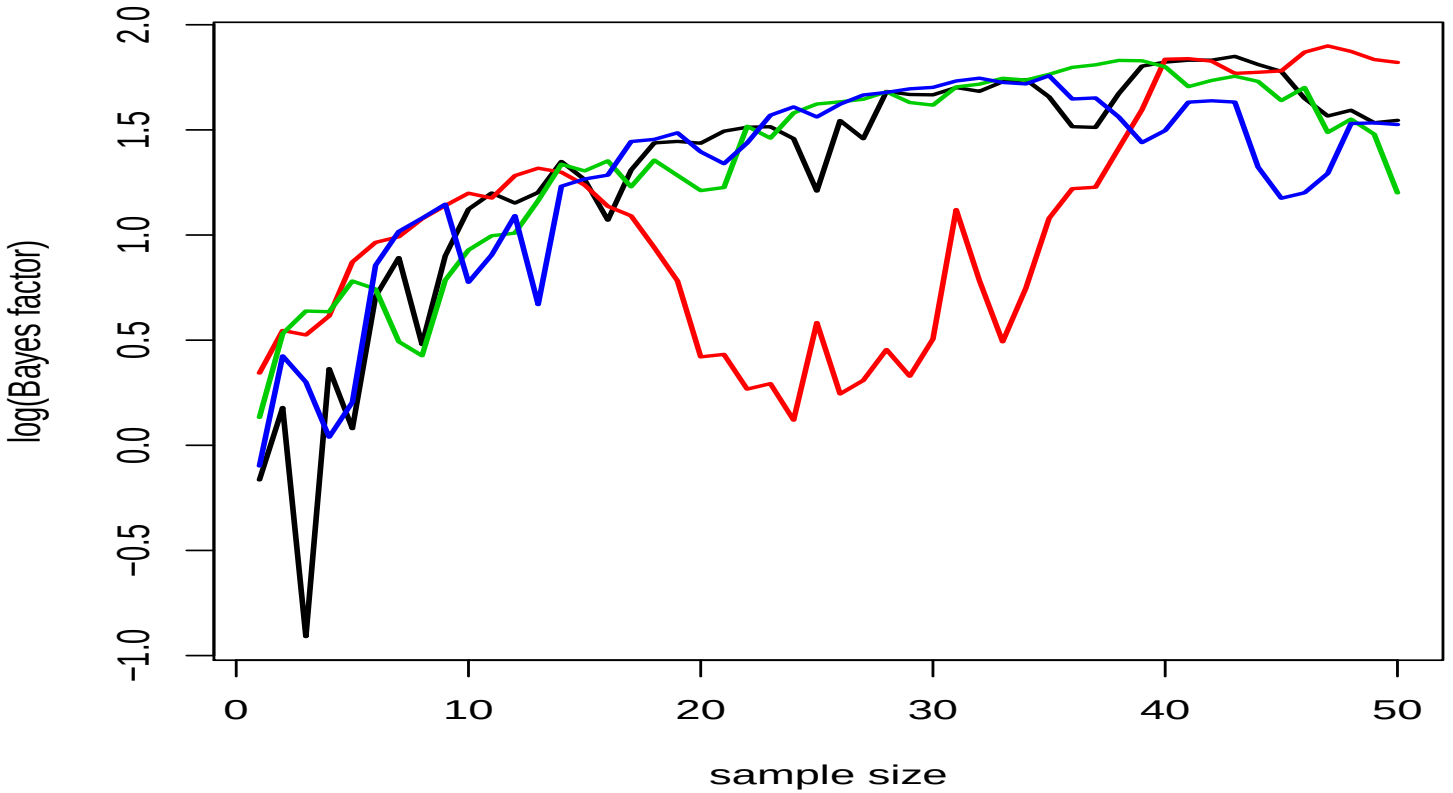
- The Bayes factor is best examined by consideration of

$$\begin{aligned} log(BF) &= log(m(y;M_1)) - log(m(y;M_2)) \\ &= \sum_{i=0}^{n-1} \log \frac{m(Y_{i+1} \mid Y_{0:i}, M_1)}{m(Y_{i+1} \mid Y_{0:i}, M_2)}. \end{aligned}$$

- The expectation under $M_1$ is non-negative, and is postive if $M_1 \neq M_2$.
- Consider examining the data set one observation at a time.
  - If $M_1$ is right, each obs'n makes a positive contribution in expectation.
  - "Trace" of GMSS is similar to Brownian motion with non-linear drift.

# Bayes factors (Cont.)

- log(BF) versus sample size



**Example 1.** 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 vs. a Mixture of Dirichlet processes (MDP) nonparametric model.

- Gaussian parametric model:

$$\begin{aligned} y_i \,|\, \theta, \sigma^2 &\overset{iid}{\sim} N(\theta, \sigma^2), \qquad i = 1, \ldots, n \\ \theta &\sim N(\mu, \tau^2) \end{aligned}$$

- DP nonparametric model:

$$\begin{aligned} y_i \,|\, \theta_i, \sigma^2 &\overset{iid}{\sim} N(\theta_i, \sigma^2), \qquad i = 1, \ldots, n \\ \theta_i \,|\, G &\overset{iid}{\sim} G \\ G &\sim DP(M = 2, N(\mu, \tau^2)) \end{aligned}$$

Common priors on hyper-parameters:

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

# Model comparison results

- Using the Bayes factor:

$$B_{P,NP} = e^{4.92} \approx 137$$

$\Rightarrow$ Decisive evidence for the parametric model!

- Using posterior predictive performance:

$$E[\log m(Y_n \mid Y_{1:(n-1)}; P)] = -1.4267$$
$$E[\log m(Y_n \mid Y_{1:(n-1)}; NP)] = -1.3977$$

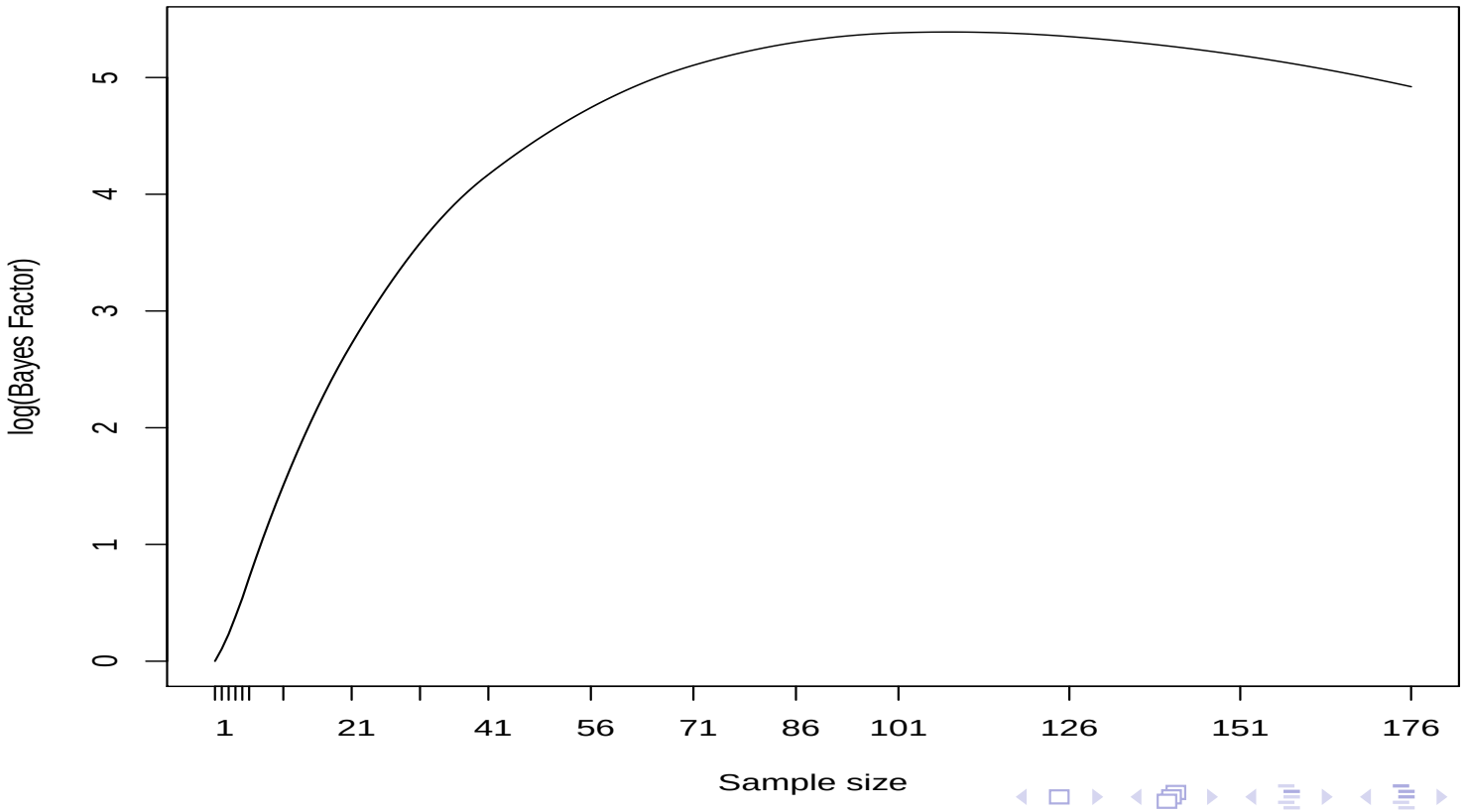
$\Rightarrow$ The nonparametric model is better!

- Given the same sample size, why do the Bayes factor and the posterior marginal likelihoods provide such very different results?

# A motivating example (Cont.)

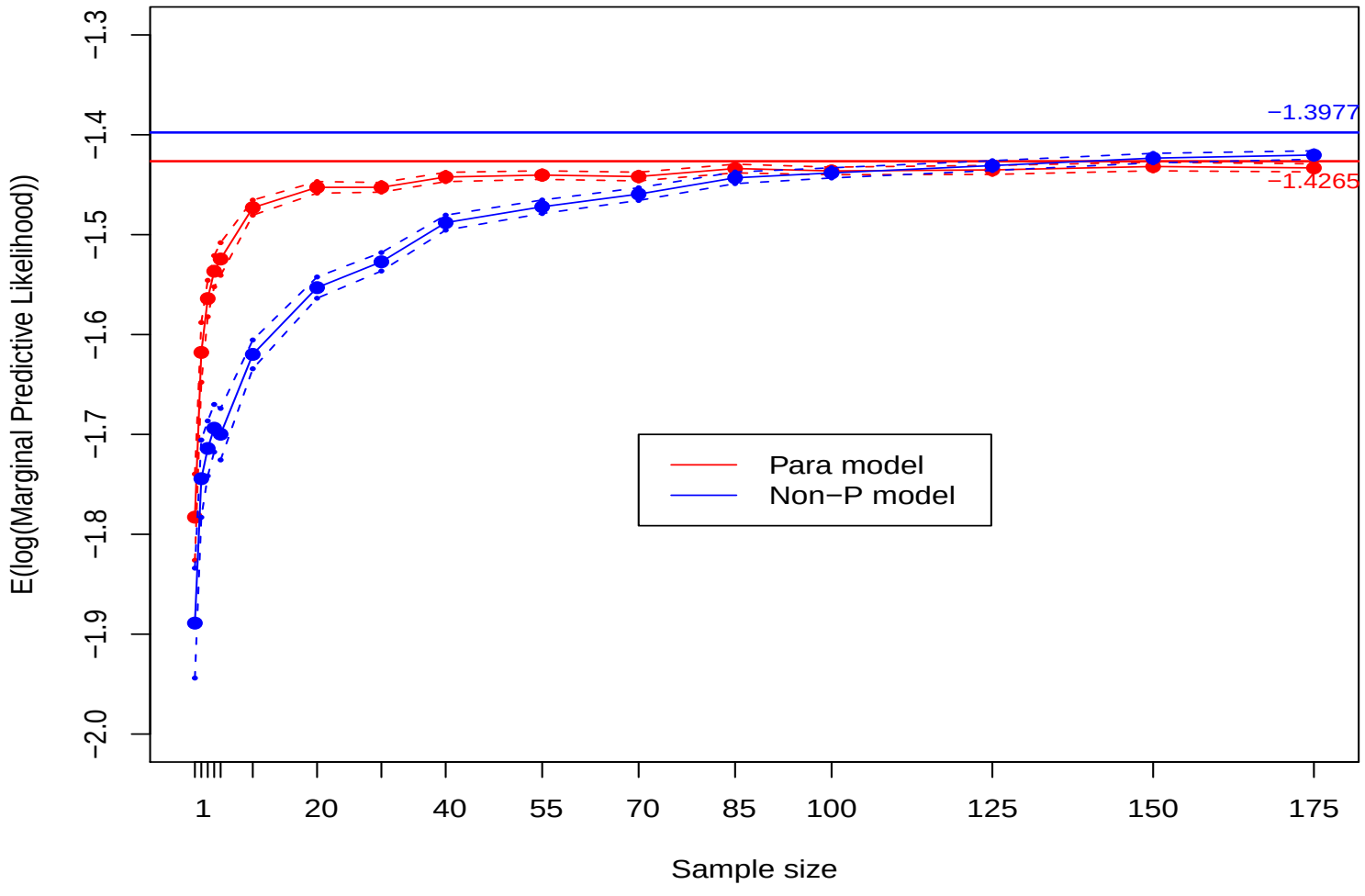
Here is the whole story... We randomly select subsamples of smaller size, compute the log Bayes factor and log posterior predictive distribution based on each subsample, and then take averages to smooth the plot

- $E[\log(\text{Bayes factor})]$ vs. sample size



# A motivating example(Cont.)

- $E[\log(\text{posterior predictive density})]$ vs. sample size



# Bayes factors and predictive distributions

- The small model accumulates an enormous lead when the sample size is small. When the large model starts to have better predictive performance at larger sample sizes, the Bayes factor is slow to reflect the change in predictive performances!
- The Bayes factor follows from Bayes Theorem. It is, of course, exactly right, provided the inputs are right
  - right likelihood
  - right prior
  - right loss

# Bayes factors – do they work?

- The Bayes factor works well for the subjective Bayesian
  - The within-model prior distributions are meaningful
  - The calculation follows from Bayes theorem
  - Estimation, testing, and all other inference fit as part of a comprehensive analysis
- The Bayes factor breaks down for rule-based priors
  - "Objective" priors (noninformative priors)
  - High-dimensional settings (too much to specify)
  - Infinite dimensional models (nonparametric Bayes)
- Many, many variants on the Bayes factor
  - Most change the prior specifically for model comparison/choice
  - One class of modified Bayes factors stands out
    - within-model priors specified by some rule
    - partial update is performed
    - then the Bayes factor calculation commences

# Bayes factors and partial updates

- Several different partial update methods (e.g., Lempers (1970), Geisser and Eddy (1979, PRESS), Berger and Pericchi (1995, 1996, IBF), O'Hagan (1995, FBF))
  - Training data / test data split
  - Rotation through different splits to stabiliize calculation
  - Prior before updating is generally "noninformative"
  - Minimal training sets have been advocated
- Questions !
  - Why a *minimal update?*
  - What if there is no noninformative prior?
  - Do the methods work in more complex settings?
- In search of a principled solution: The question is *not* Bayesian model selection, but Bayesian model selection which is consistent with the rest of the analysis.

# Outline

# Toward a solution - the calibrated Bayes factor

- Subjective Bayesian analyses work well in high-information settings, much less well in low-information settings (try elicitation for yourself)
- We begin with the situation where all (Bayesians) agree on the solution and use this to drive the technique
- We propose that one start the Bayes factor calculation after the partial posterior resembles a high-info subjective prior
- Elements of the problem
  - Measure the information content of the partial posterior
  - Benchmark prior to describe adequate prior information
  - Criterion for whether partial posterior matches benchmark
- We recommend calibration of the Bayes factor in any low-information or rule-based prior setting
- In these settings, elicited priors are unstable (Psychology)

# Measurement of prior information

- Measure the proximity of $f_{\theta_1}$ and $f_{\theta_2}$ through the the Symmetric - Kullback-Leibler (SKL) divergence

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

- SKL driven by likelihood, appropriate for Bayesians
- The distribution on $(\theta_1, \theta_2)$ induces a distribution on $SKL(f_{\theta_1}, f_{\theta_2})$
- This works for infinite-dimensional models too, unlike alternatives such as Fisher information
- Criterion: Evaluate the information contained in $\pi$ using the percentiles of the distribution of SKL divergence

# A benchmark prior

- To calibrate the Bayes factor and select a training sample size, we 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 information in a single observation
- Under the Gaussian model $Y \sim N(\theta, \sigma^2)$, a unit information prior on $\theta$ is $N(\mu, \sigma^2)$, inducing a $\chi^2_1$ distribution on $SKL(f_{\theta_1}, f_{\theta_2})$.

# Calibrating the priors

The overall strategy:

- **Step 1:** For a single model, randomly draw a training sample of a pre-specified sample size from the data
- **Step 2:** Update the prior based on this training sample. Take $M$ pairs of $(\theta_1^j, \theta_2^j)$, 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 $N$ times. Pool all $MN$ values of the SKLs to evaluate the information in the posterior
- **Step 4:** Compare the amount of information in the posterior to that in the benchmark distribution. If the amount of information 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 factors

- Let $s_1$ and $s_2$ represent the calibration sample sizes for models $M_1$ and $M_2$. Take $s = \max(s_1, s_2)$.
- Based on a training sample $y_{(s)}$, the updated Bayes factor satisfies

$$\log B_{12}^*(y \mid y_{(s)}) = \log B_{12}(y) - \log B_{12}(y_{(s)}),$$

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

$$\log CB_{12}(y) = \log B_{12}(y) - \frac{1}{H}\sum_{h=1}^{H} \log B_{12}(y_{(s)}^h)$$

# Asymptotic properties

- The calibrated Bayes factor shares the qualitative asymptotic properties of the Bayes factor.
- The main condition is that the calibration sample size (under both models) be finite with probability one. This removes an expected log(Bayes factor) based on the calibration sample size.
- If log(BF) tends to some infinite limit, then so does log(CBF).
- If log(BF) tends to some finite number (odd, but examples exist), then log(CBF) will tend to a number differing by the offset.

# Perils avoided

- We have avoided two huge pitfalls:
- Change the form of the prior distribution?
  - Destroys the cohesiveness of the analysis, including both selection and estimation
  - Hampers use of low-information priors
  - Adding an extra model may change the priors!
- Adjust the hyper-parameters in the prior?
  - Where should this more concentrated prior be centered?
  - We use the data to drive the centering via the training sample

- Instead, we use training samples to update the priors
  - What within-model prior do you want to use for estimation?
  - Training sample size is chosen to mimic Bayes factor when they work well
  - Driven by actual data without double use of the data

# The motivating example revisited

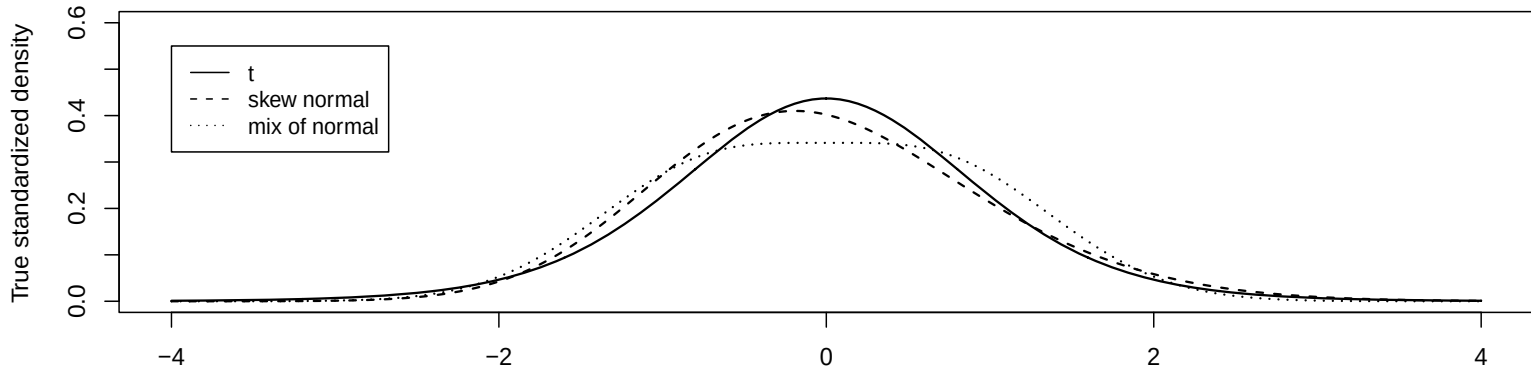
- In the skew-normal example, our search leads to a calibration sample size of 50, based on the MDP model.
- The calibrated Bayes factor $\approx e^{4.92-4.54} \approx 1.46$, which is not worth more than a bare mention under Jeffrey's criterion. This result is consistent with the posterior predictive performances.
- Under the calibrated Bayes factor, the small parametric model never accumulates a significant lead!

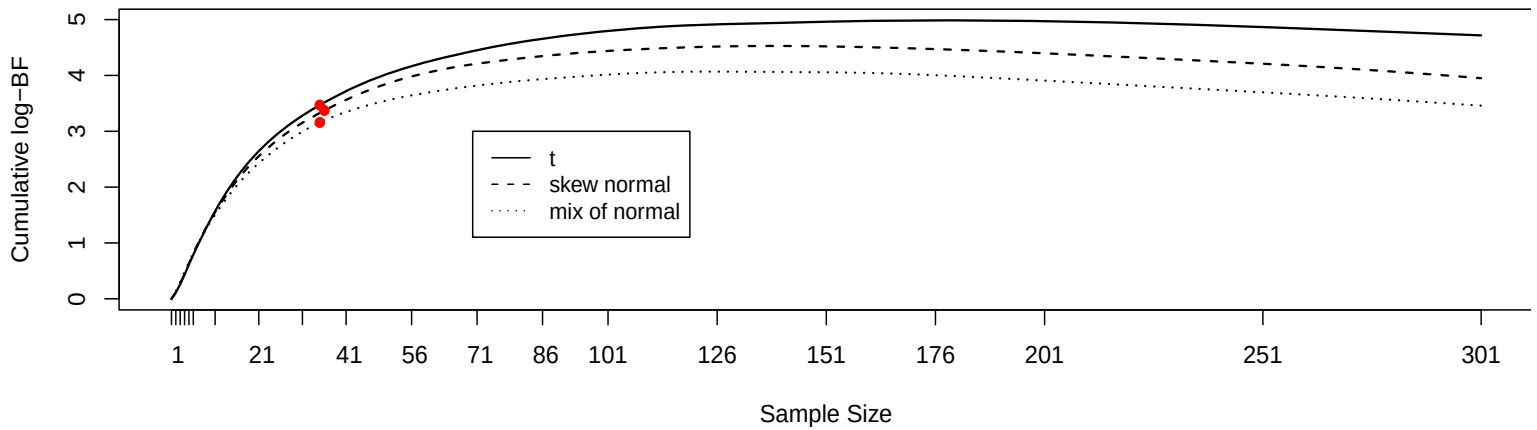
# The simulation setup

- 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 $\alpha$ (skewness)
  - Student-t with varying degrees of freedom $\nu$ (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 $\alpha$, $\nu$ and $\delta$, we tune the KL distances from the true distributions to the best fitting Gaussian distributions.

# Simulation results

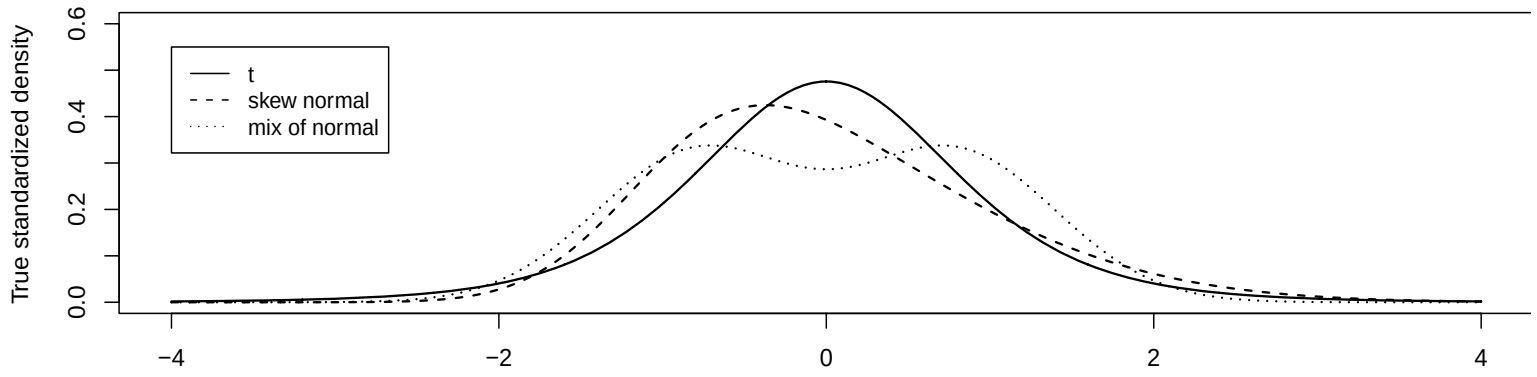
- Small divergences from the Gaussian

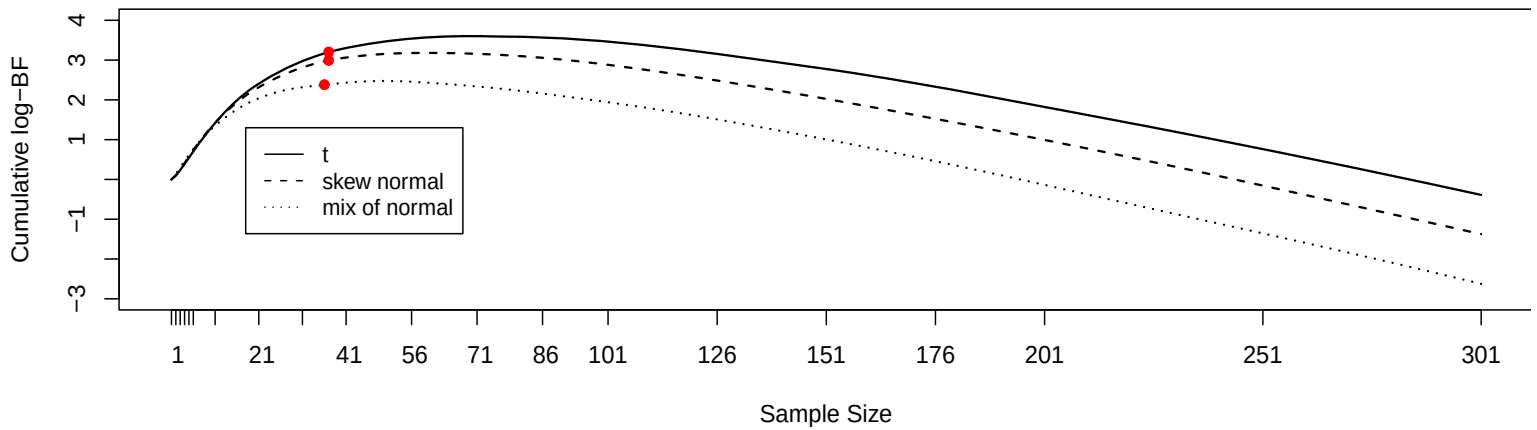


# Simulation results (Cont.)

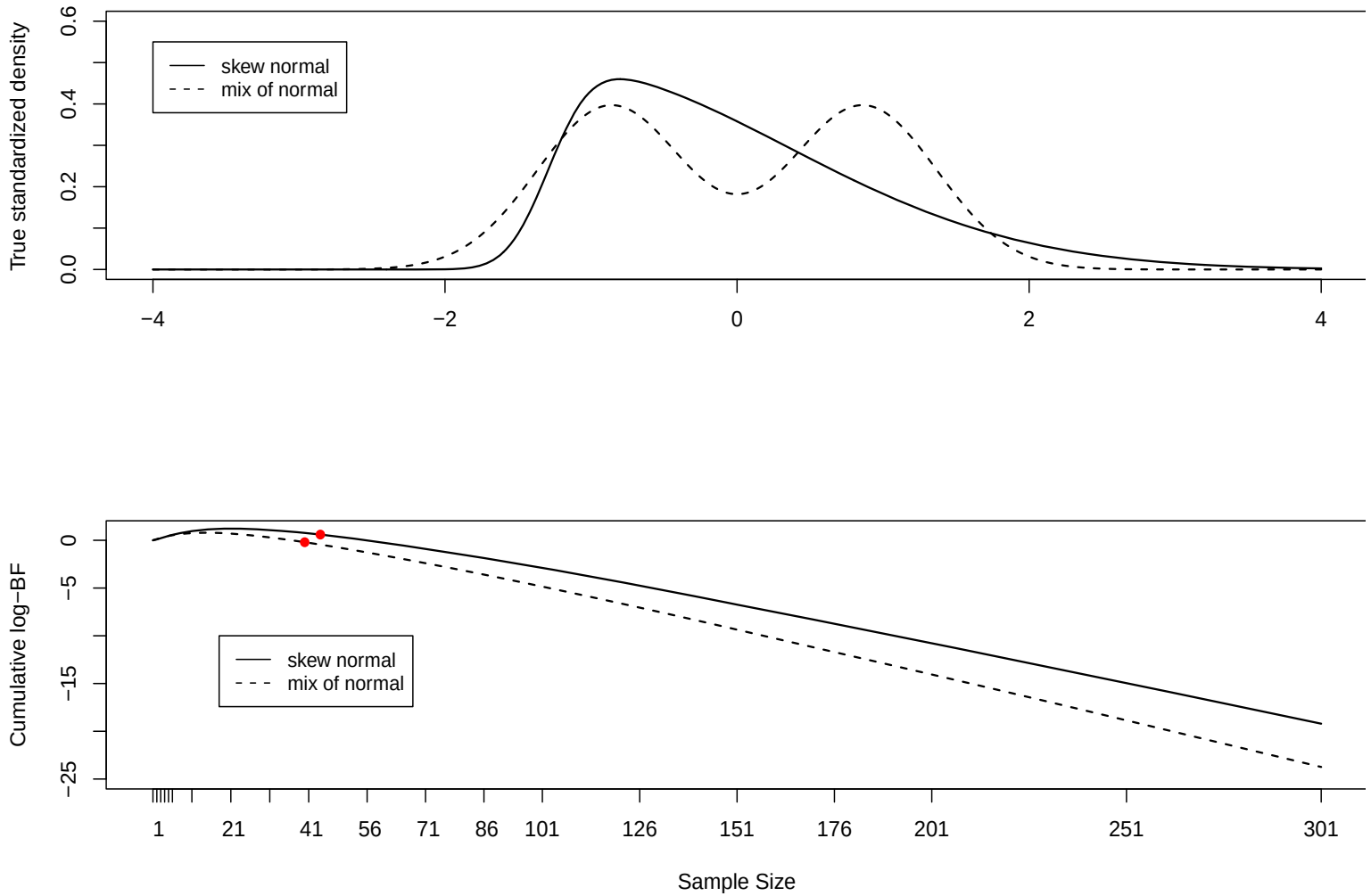
- Moderate divergences from the Gaussian



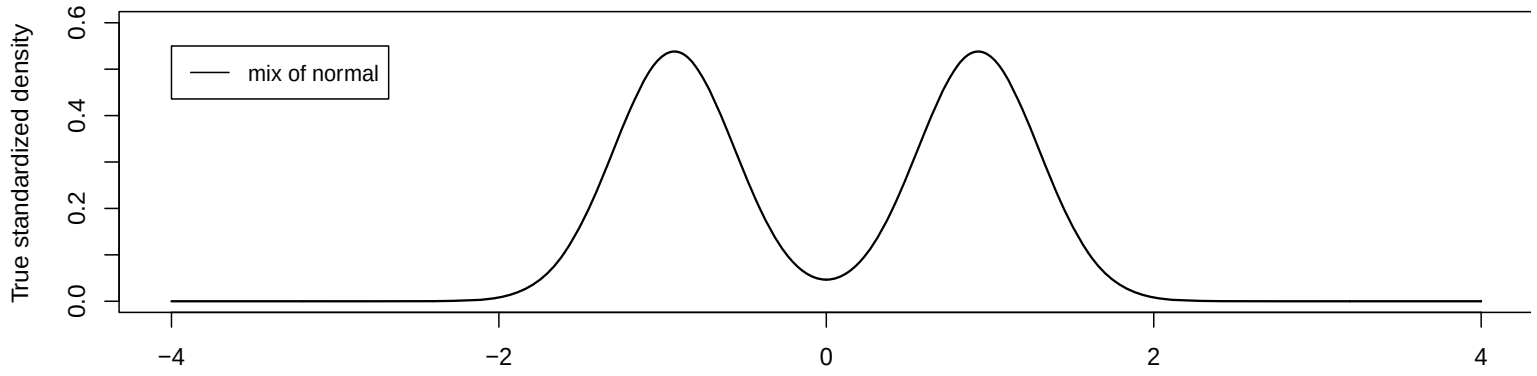
# Simulation results (Cont.)

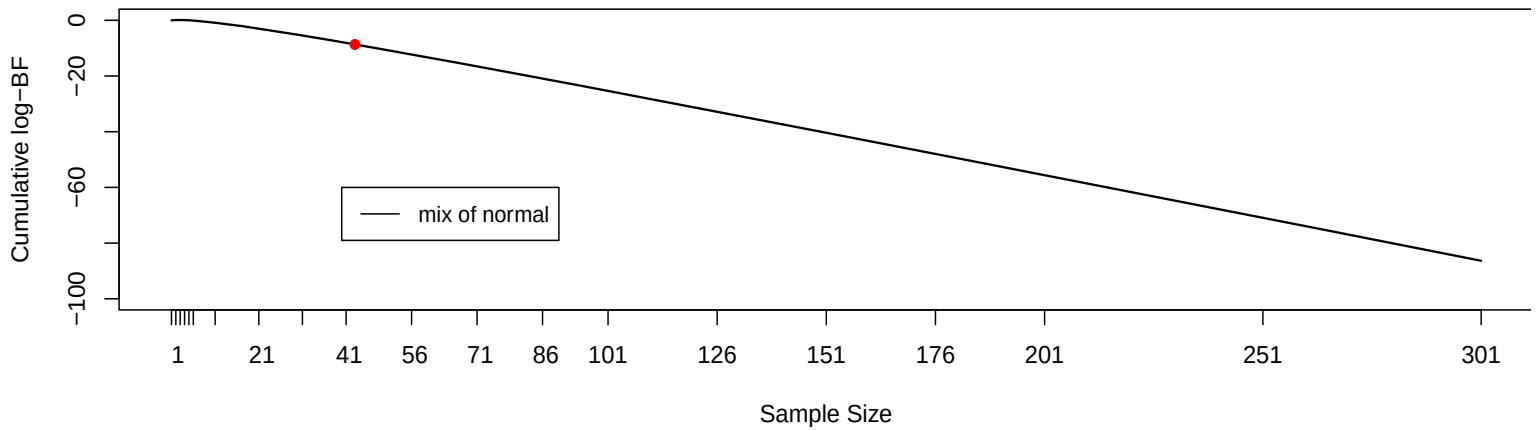
- Large divergences from the Gaussian



# 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.

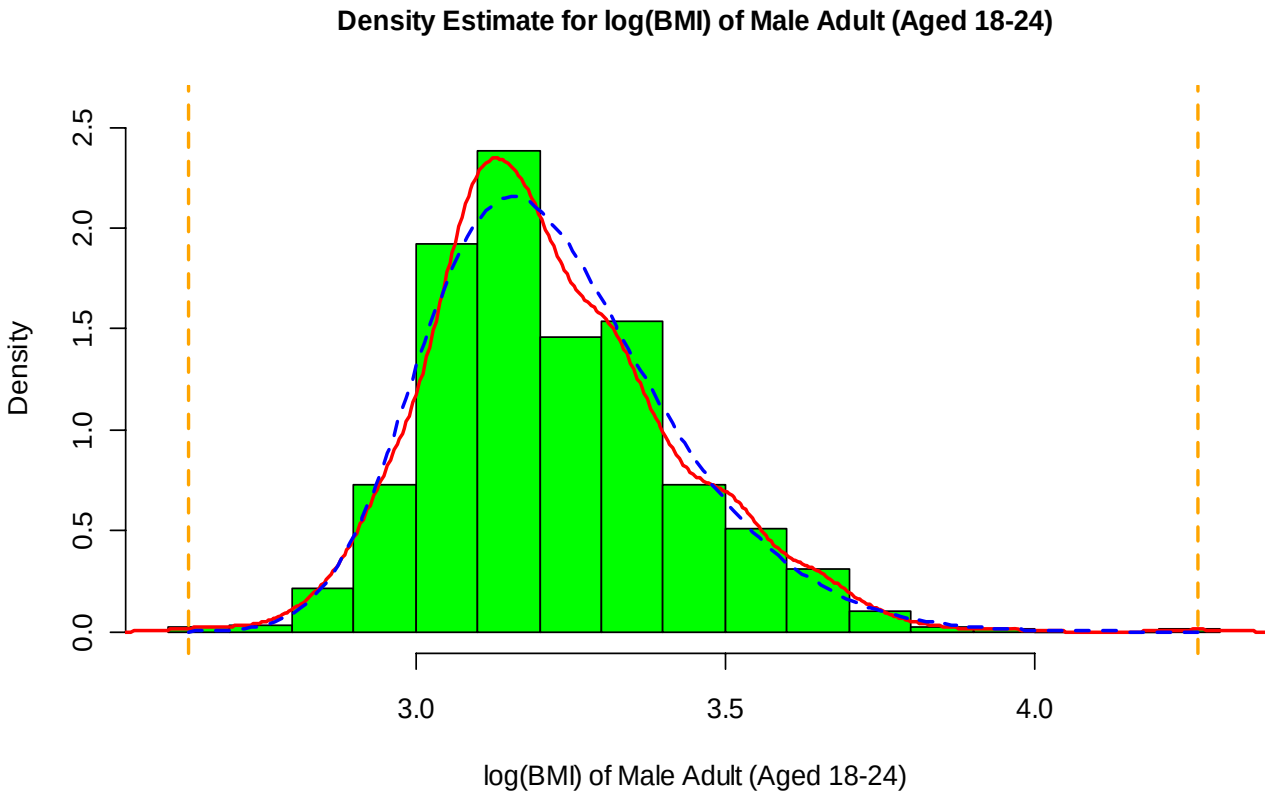
# Outline

# Model comparisons in OFHS analysis

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

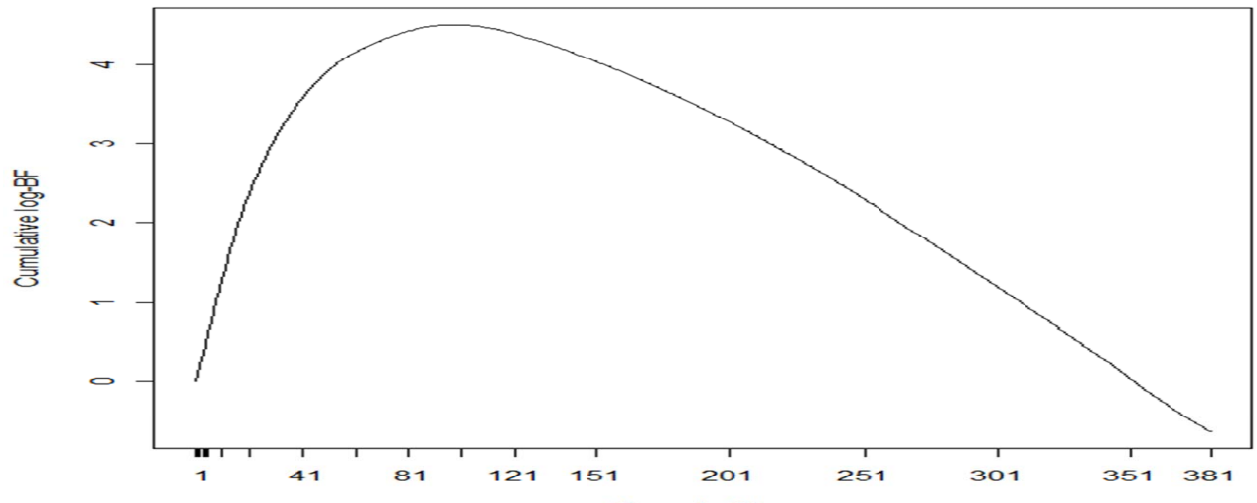
# Model comparison in the OFHS analysis (Cont.)

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



# Model Comparison in the 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



# Model Comparison in the OFHS analysis (Cont.)

- After matching the prior concentration to the unit information prior, we calibrate the priors with training samples of size $50$.
- At sample size $n = 106$, the calibrated Bayes factor is $CB_{P;NP} \approx e^{0.64} \approx 1.9$, which provides a very weak model preference; at the full sample size, the eventual calibrated Bayes factor is $CB_{P;NP} \approx e^{-16.18}$, or about $10.6$ 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

# Concluding remarks

- Huge swings in the Bayes factor are not unique for parametric vs. nonparametric model comparisons. They are prevalent in small vs. large model comparisons.
- The original Bayes factor can be very misleading in this situation. There seems to be no sound remedy through alteration of the prior distributions. Such methods destroy the cohesiveness of the analysis.
- To make a fair comparison between small and large models, careful calibration of the Bayes factor is needed, and this can be done through the use of training samples. Such adjustment is also needed whenever priors are poorly specified or are rule-based.
- Regression models, discrete data models, dependence models (spatial, temporal, other), complex hierarchical models, ...

# A disturbing message

- The story has been that the Bayes factor is inadequate for model comparison in “hard” problems.
- But the Bayes factor is merely an expression of Bayes’ Theorem.
- Model comparison involves a contrast across submodels of a “hyper-model”
- So, what is to say that we should consider the MDP component of our hypermodel a submodel?
- If we split our model up differently, should we calibrate our Bayes Theorem calculations differently as well?