A nonparametric goodness-of-fit test for random effects models via cross-validation Bayes factors

Jeffrey D. Hart
Texas A&M University

Talk given at the conference "Latent Variables 2016," University of South Carolina

October 13, 2016

Collaborators

Collaborator on random effects work:

Matthew Malloure, Texas A&M University.

Paper of Hart and Choi (2016) on the analogous method for *univariate* data:

- To appear in *Bayesian Analysis*.
- Advance version available at `https://projecteuclid.org/adv/euclid.ba`.

Outline

- A simple random effects model
- Models for distributions of random effects
- Bayesian gof with alternatives defined in terms of kdes
- Bandwidth prior
- Simulation results
- A microarray example
- Concluding remarks

A random effects model

Observed data are $\mathbf{X}_i$, $i = 1, \ldots, p$; $\mathbf{X}_i = (X_{i1}, \ldots, X_{in})$ and

$$X_{ij} = \mu_i + \epsilon_{ij}, \quad j = 1, \ldots, n, \ i = 1, \ldots, p.$$

- $\mu_1, \ldots, \mu_p$ are i.i.d.
- ϵ_{ij}, $j = 1, \ldots, n$, $i = 1, \ldots, p$, are i.i.d.
- $\boldsymbol{\mu}$ is independent of $\boldsymbol{\epsilon}$.
- For identifiability, μ_i and ϵ_{ij} have respective location parameters μ and 0.

Note: $n \geq 2$ and until further notice we assume that $n = 2$.

Models for distributions

The purpose of this talk is to present a Bayesian method for testing the fit of a model for the distributions of μ_i and ϵ_{ij}.

- By far the most often used model is the Gaussian model, which says that both μ_i and ϵ_{ij} are normal.
- In the Gaussian model the analysis boils down to estimating μ and the variances of μ_i and ϵ_{ij}, call them σ^2_μ and σ^2_ϵ.
- Non-Gaussian parametric models: A preliminary analysis of the data may suggest a *different* parametric model whose fit one would like to formally test.

An old identifiability result

Let F_μ and F_ϵ be the distributions of μ_i and ϵ_{ij}, respectively.

> Reiersol (1950): Under mild regularity conditions on the characteristic functions of F_μ and F_ϵ, F_μ and F_ϵ are completely determined by the joint distribution F of $(\mu_i + \epsilon_{i1},, \mu_i + \epsilon_{i2})$.

Important implication for our problem: *We may test the fit of a model for F_μ and F_ϵ by testing the fit of the corresponding model for F.*

The Gaussian model

The Gaussian version of our random effects model is "special" in the sense that F_μ and F_ϵ are both normal if and only if $\mu_i + \epsilon_{ij}$ is normally distributed, which follows from Cramér's Theorem.

- Undoubtedly, the easiest and most efficient way to test the Gaussian model is to apply univariate methods to the X_{ij}'s. Lange and Ryan (1989); Hart and Choi (2016)
- Nonetheless, we will use the Gaussian null model as an example of our methodology for the more general case.

KDE

- Kernel density estimates have two appealing properties. They are (1) *simple* and (2) *nonparametric*.
- Kernel estimates have been used to construct frequentist goodness-of-fit tests. Is there a way to use them in a *Bayesian* goodness-of-fit test?
- Bayesian nonparametric procedures based on priors over function spaces can be fairly complicated.
- Bayesian nonparametric gof tests: Verdinelli and Wasserman (1998), Berger and Guglielmi (2001), McVinish, Rousseau, and Mengersen (2009), Tokdar and Martin (2013)

Methodology

- In the Bayesian paradigm the models should come from outside the data on which the models are assessed.
- Kernel estimates do not provide a (simple) a priori model. A kernel *estimate* only becomes a model once it is computed from data.
- Sidestep this problem by using *data splitting.* Divide the data set into a *training* and a *validation* set.
  - Compute a kernel density estimate from the training data.
  - Calculate a Bayes factor from the validation set using the kernel estimate as alternative model.

Methodology

Let $\mathbf{X}^T = (\mathbf{X}_1, \ldots, \mathbf{X}_m)$ serve as the training data.

Denote the rest of the data by $\mathbf{X}^V$, the validation set.

Define a kernel estimate by

$$\hat{f}(x, y | h, \mathbf{X}^T) = \frac{1}{mh^2} \sum_{i=1}^{m} K\left(\frac{x - X_{i1}}{h}\right) K\left(\frac{y - X_{i2}}{h}\right).$$

Treating $\hat{f}(\cdot\,|h, \mathbf{X}^T)$ as a model for $\mathbf{X}^V$, the only parameter is the bandwidth h, and the likelihood is

$$L_V(h) = \prod_{i=m+1}^{p} \hat{f}(\mathbf{X}_i | h, \mathbf{X}^T).$$

Methodology

Let $\mathcal{F} = \{f(\cdot|\theta) : \theta \in \Theta\}$ be a parametric model for the density of $\mathbf{X}_i = (X_{i1}, X_{i2})$.

We wish to test the hypotheses

$$H_0 : f \in \mathcal{F} \quad \text{vs.} \quad H_1 : f \notin \mathcal{F}.$$

Let p and π be priors for h and θ, respectively. A Bayes factor for testing H_0 vs. H_1 is

$$BF(\mathbf{X}_V) = \frac{\int_0^\infty L_V(h)p(h)\,dh}{\int_\Theta \left[\prod_{i=m+1}^p f(\mathbf{X}_i|\theta)\right]\pi(\theta)\,d\theta}.$$

Methodology

We don't want the inference to depend on the chosen data partition, so ideally we would compute

$$BF = \left(\prod_i BF(\mathbf{X}_{V,i}) \right)^{1/N},$$

the geometric mean of Bayes factors over all possible data partitions.

- Not feasible to compute $BF(\mathbf{X}_{V,i})$ for all possible partitions.
- Our experience shows that using more than about 50 randomly chosen partitions yields little new information.

Augmented training set

In our model, the data $X_{i1}, \ldots, X_{in}$ are unconditionally dependent. For $j \neq k$,

$$\mathrm{Cov}(X_{ij}, X_{ik}) = \mathrm{Var}(\mu_i).$$

However, $X_{i1}, \ldots, X_{in}$ are exchangeable and hence the n univariate marginals are the same.

We can guarantee that our kde has equal marginals by computing it from the data

$$(X_{11}, X_{12}), \ldots, (X_{m1}, X_{m2}), (X_{12}, X_{11}), \ldots, (X_{m2}, X_{m1}).$$

Cross-validated bandwidth choice

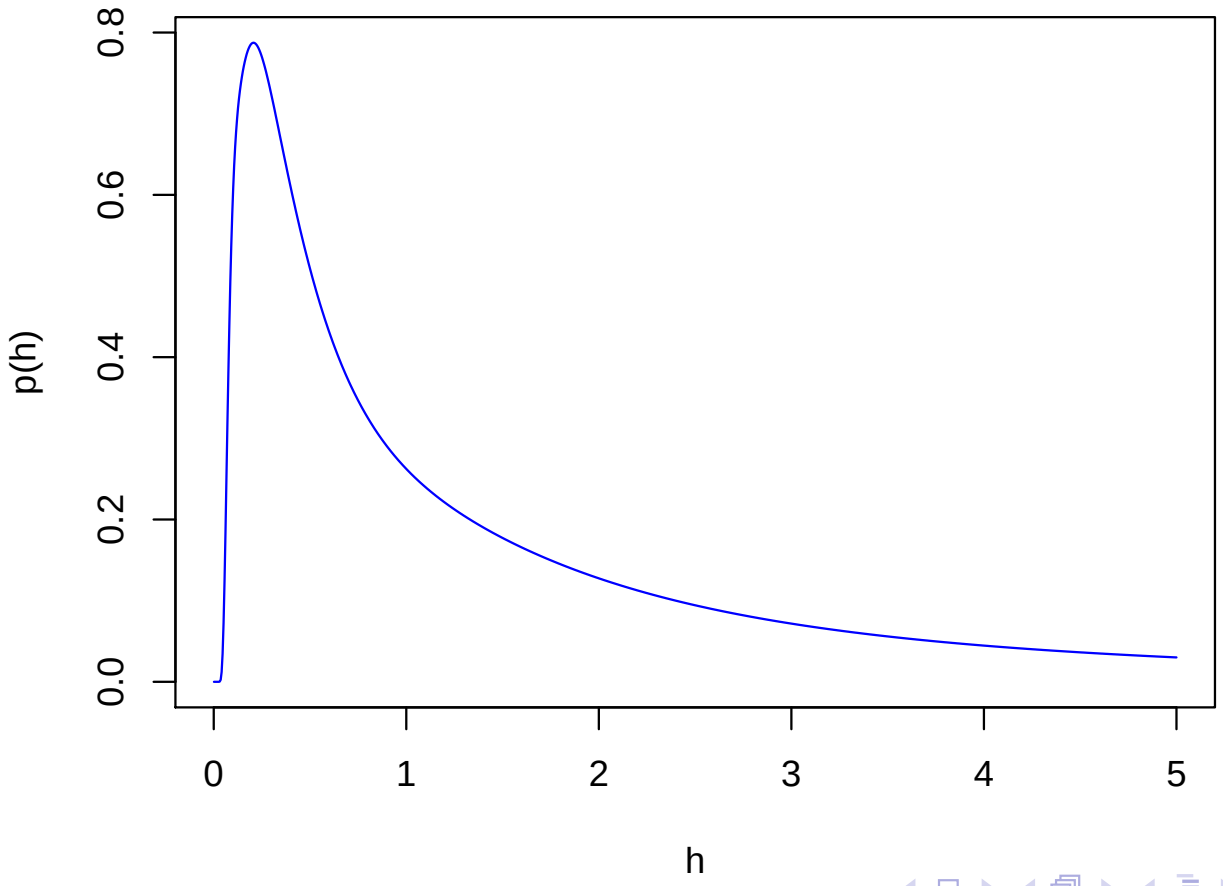
- van der Laan, et al. (2004) propose the cross-validated likelihood criterion as a means of bandwidth selection.
- For bandwidth selection, they ask that the size of the training set be asymptotic to p.
- In contrast, we ask that size of the *validation set* be asymptotic to p.

Prior for h

- We use a unit information prior centered at the validation data.
- Let $\tilde{x}$ be the median of all $2(p-m)$ of the validation data. Then
$$p(h) \propto \hat{f}(\tilde{x}, \tilde{x}|h, \mathbf{X}^T).$$
- If we use a Gaussian kernel, the constant of proportionality is easily determined.

A typical unit information prior

This prior was constructed from data to be discussed later. $(m = 100)$



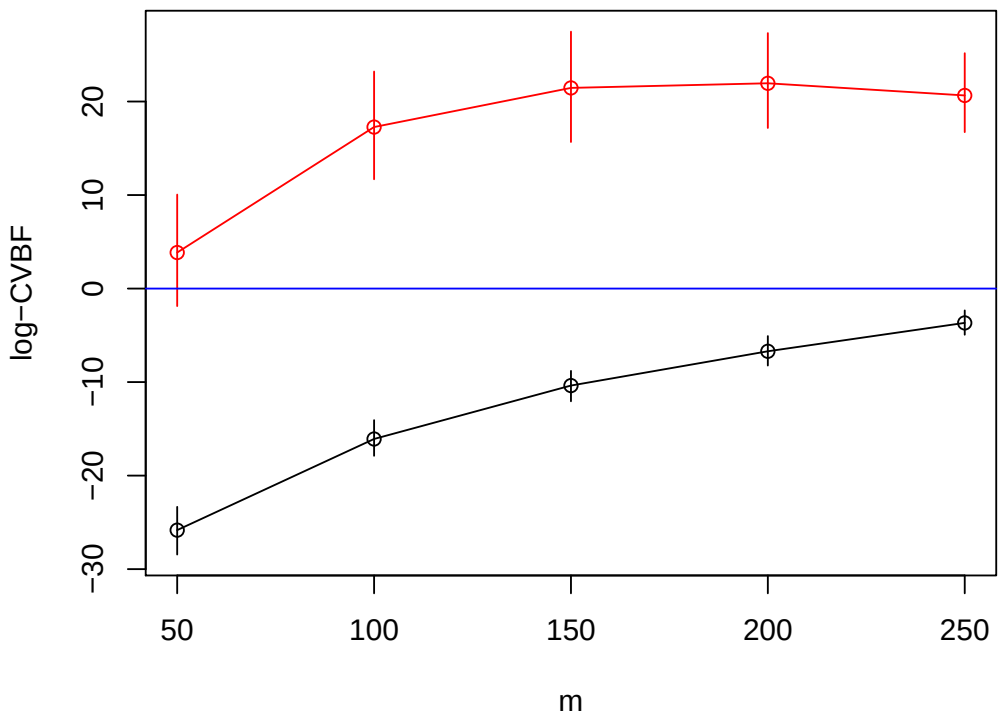
Simulation

- H_0 : Both F_μ and F_ϵ are normal.
- Prior for parametric model:
  - Prior for $(\mu, 1/\sigma_X^2|\rho)$ is a unit reference, conjugate prior centered at validation data.
  - $\rho \sim U(0,1)$
- $p = 500$
- Training set sizes: 50, 100, 150, 200, 250
- 1000 replications
- Twenty-five random splits for each data set and training set size

Log-CVBF as a function of training set size

Black line: $\mu \equiv 0$ and $\epsilon_{ij} \sim$ Normal

Red line: $\mu \equiv 0$ and $\epsilon_{ij} \sim$ Laplace



Effect of training set size

- When H_0 is true, CVBF tends to increase monotonically with training set size. Larger training set size produces better nonparametric model and smaller validation set, both of which are unfavorable to H_0.

- When H_0 is false, CVBF tends to increase to a maximum as training set size increases, and then decrease. Larger training set size produces better nonparametric model but at some point this is offset by validation set getting smaller.

Calibration of CVBF

If a calculated CVBF seems to indicate that H_0 should be rejected, it is advisable to consider the behavior of CVBF when H_0 is true. We call this process *calibration*.

- Generate B samples (of same size as your data set) from the null model.
- For each sample, compute CVBF for a range of training set sizes.
- Choose a training set size for which the average log-CVBF is well below 0.

Rat gene expression data

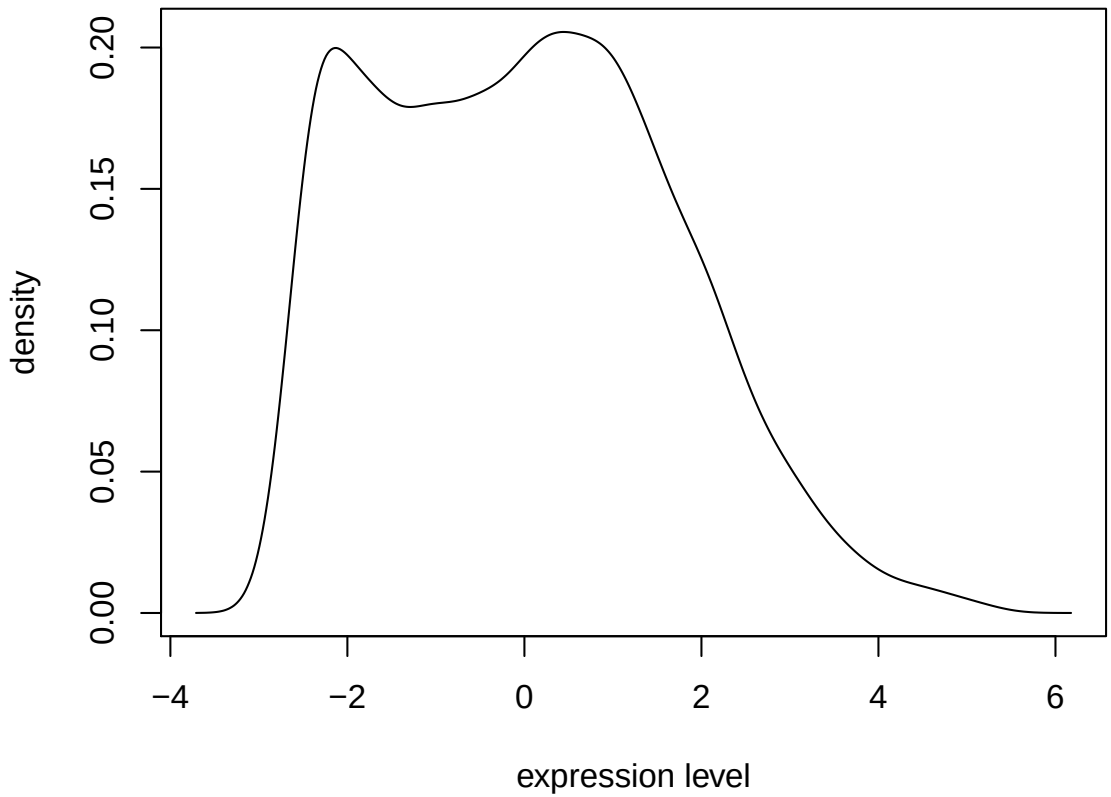
We consider microarray data collected by Robert Chapkin of Texas A&M University; Davidson, et al. (2004).

- Logged expression levels for 8034 genes of five rats: $p = 8034$, $n = 5$
- The mean of the 8034 observations for a given rat was subtracted from each observation for that rat.

Distribution of gene means

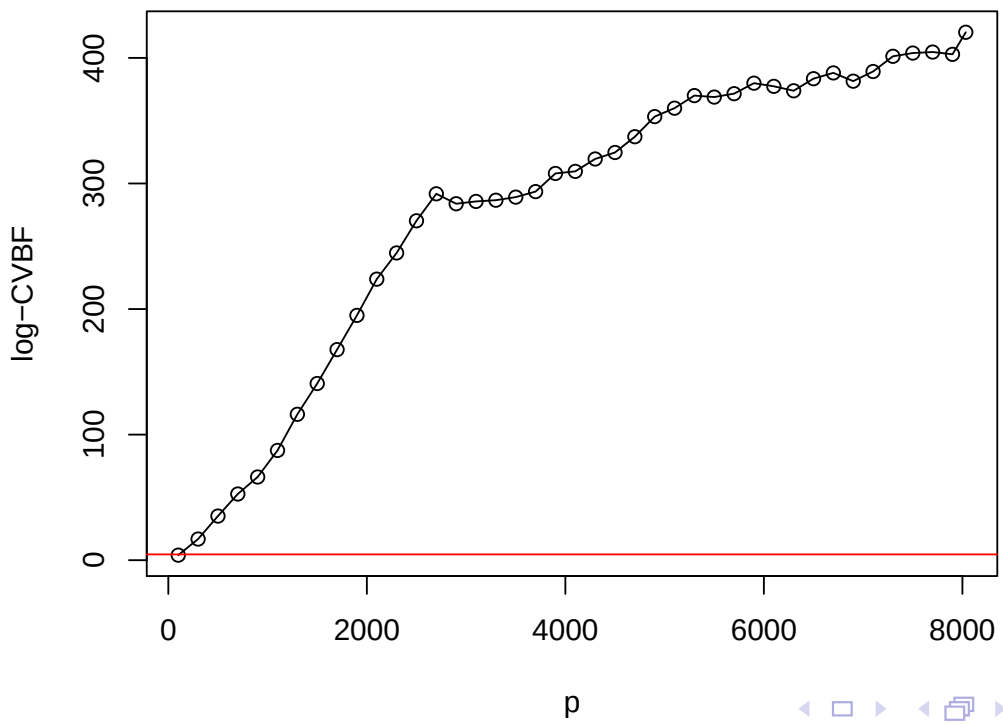
Define $\bar{X}_i = \mu_i + \bar{\epsilon}_i$, $i = 1, \ldots, 8034$.

Density estimate computed from $\bar{X}_1, \ldots, \bar{X}_{8034}$.

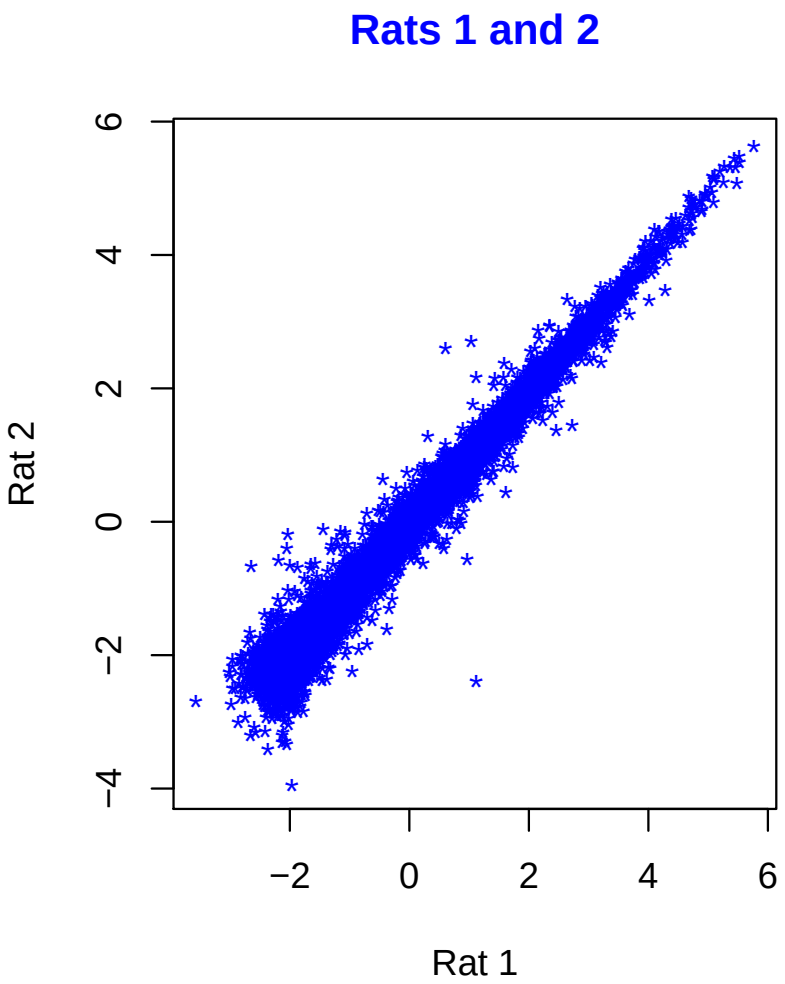


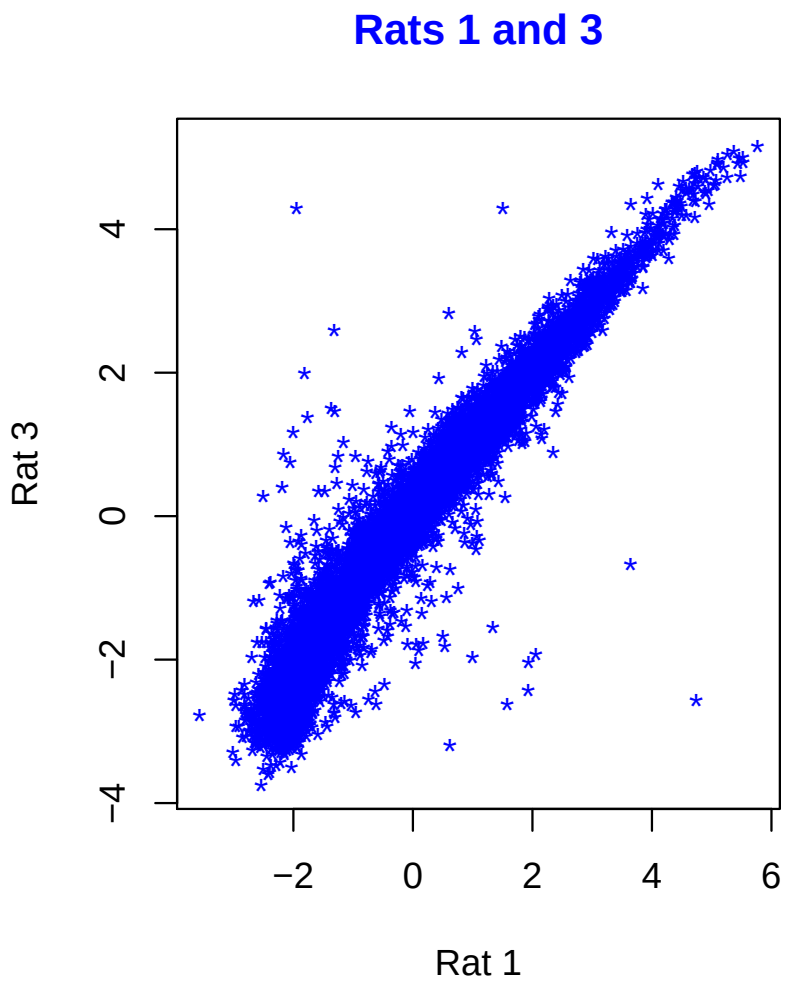
Log-CVBF for univariate gof

- Procedure of Hart and Choi (2016) applied to $\bar{X}_1, \ldots, \bar{X}_{8034}$ to test them for normality.
- Used subsets of data ranging in size from 100 to 8034.
- Training set size was always 20% of sample size; 25 random splits.



Scatterplots





Log-Bayes factors for pairs of rats

Jeffrey D. Hart
Texas A&M University

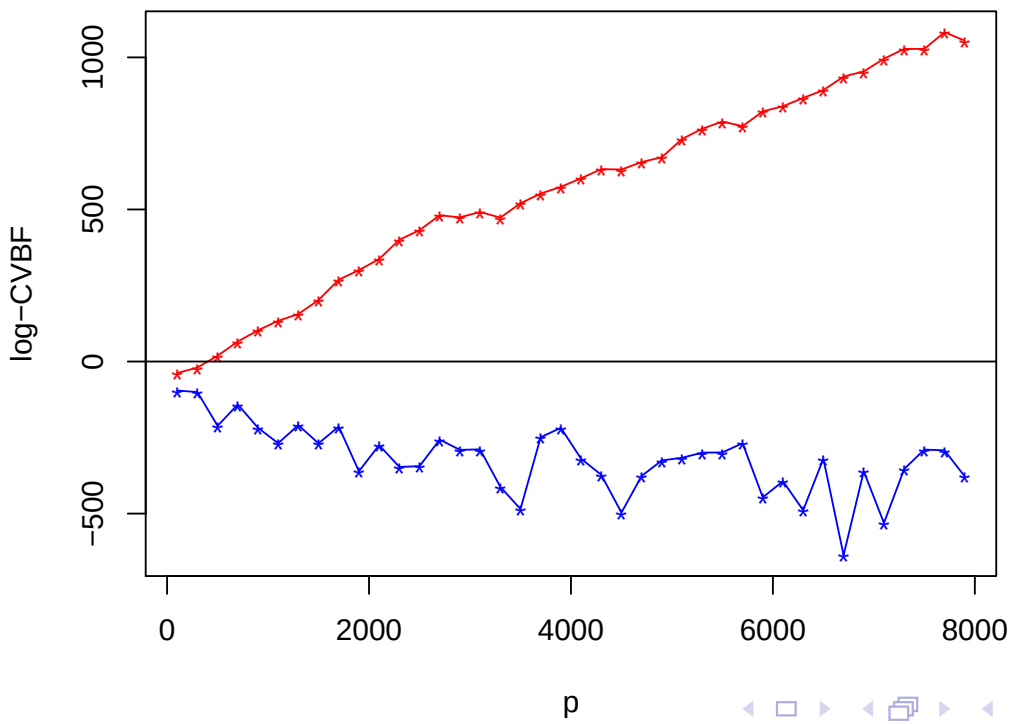
Our procedure for testing the Gaussian model was applied to each pair of rats. The values below are the means of log-CVBF based on training set sizes of 1600 and 25 random splits.

		Rat 1	Rat 2	Rat 3	Rat 4
	2	1080.6			
Rat	3	1096.8	1112.8		
	4	1189.9	1238.1	686.8	
	5	1051.3	1063.6	700.2	647.5

There's overwhelming evidence to reject the Gaussian model.

Log-CVBF for bivariate gof (Rats 1 and 2)

- Red line: log-CVBF for Rats 1 and 2; Blue line: log-CVBF for simulated Gaussian data
- Used subsets of data ranging in size from 100 to 7900.
- Training set size was always 20% of sample size; 25 random splits.



Concluding remarks

- In univariate case, Hart and Choi (2016) provide general conditions under which CVBF is Bayes consistent *at an exponential rate under both hypotheses.*
- No reason to believe that consistency result doesn't extend to random effects model.
- More work needs to be done on the question of choosing m, the training set size. For consistency, theory says that m should tend to ∞ with p in such a way that $m = o(p)$.

Concluding remarks

- When $n > 2$, is it better to use a kde for n-variate data or to average 2-variate results?
- Perhaps there are better ways to construct nonparametric models than by using kdes.
- Extensions to more complex random and mixed effects models.