The FBST for Model Selection in Mixture of Multivariate Normals

Marcelo de Souza Lauretto

Carlos A. de Bragança Pereira
BioInfo and Statistics Department

Julio Michael Stern
BioInfo and Computer Science Department

University of São Paulo, Brazil.

The FBST Value of Evidence
Full Bayesian Significance Test
(Pereira and Stern, 1999)

Posterior density, likelihood and prior:

$$p_x(\theta) \propto L(\theta \,|\, x)\, p(\theta).$$

Null hypothesis:

$$\Theta_H = \{\theta \in \Theta \,|\, g(\theta) \leq \mathbf{0} \wedge h(\theta) = \mathbf{0}\}$$

Sharp (precise) hypotheses:
$\dim(\Theta_H) < \dim(\Theta)$.

Evidence against the hypothesis:

$$
\begin{aligned}
\text{Ev}(H) &= \int_{T_H} p_x(\theta)\, d\theta \ , \ \text{where} \\
T_H &= \{\theta \in \Theta \mid s(\theta) > s_H\} \\
s_H &= \sup_{\theta \, \in \, \Theta_H} s(\theta) \\
s(\theta) &= \left(\frac{p_x(\theta)}{r(\theta)}\right)
\end{aligned}
$$

$s(\theta)$ is the Posterior Surprise
If the reference density $r(\theta) \propto 1$,
the Tangent set T_H , HRSS = HDPS

Operationally:
Optimization + Integration step

Mixture Models

Sample: x^j , $j = 1 \ldots n$

Coord: x_h^j , $h = 1 \ldots d$

Classes: $c(j) = k$, $k = 1 \ldots m$

Latent Variables: $z_k^j = \mathbf{1}\,(c(j) = k)$

Numb. samp. class k, y_k : $y = Z\mathbf{1}$

Mixture parameters:

$$\text{Pr}\,(c(j) = k) = w_k$$

$$\text{if } c(j) = k \text{ then } x^j \sim f(x^j \,|\, \psi_k)$$

$$\theta = [w_1, \ldots, w_k, \psi_1, \ldots, \psi_k]$$

Conditional on the missing data:

$$\begin{aligned}
f(x^j \,|\, \theta) &= \sum_{k=1}^{m} f(x^j \,|\, \theta, z_k^j) f(z_k^j \,|\, \theta) \\
&= \sum_{k=1}^{m} w_k f(x^j \,|\, \psi_k) \\
f(X \,|\, \theta) &= \prod_{j=1}^{n} f(x^j \,|\, \theta) \\
&= \prod_{j=1}^{n} \sum_{k=1}^{m} w_k f(x^j \,|\, \psi_k)
\end{aligned}$$

Conditional classification probabilities, $P = f(Z \,|\, X, \theta)$:

$$\begin{aligned} p_k^j &= f(z_k^j \,|\, x^j, \theta) \quad = \quad \frac{f(z_k^j, x^j \,|\, \theta)}{f(x^j \,|\, \theta)} \\ &= \frac{w_k f(x^j \,|\, \psi_k)}{\sum_{k=1}^m w_k f(x^j \,|\, \psi_k)} \end{aligned}$$

Likelihood for the "completed" data, X, Z:

$$\begin{aligned} f(X, Z \,|\, \theta) &= \prod_{j=1}^n f(x^j \,|\, \psi_{c(j)}) f(z_k^j \,|\, \theta) \\ &= \prod_{k=1}^m \left[(w_k)^{y_k} \prod_{j \,|\, c(j)=k} f(x^j \,|\, \psi_k) \right] \end{aligned}$$

where $y_k = \sum_j z_k^j$.

Normal-Wishart Distribution

u and S are the statistics:

$$
\begin{aligned}
u &= \frac{1}{n}\sum_{j=1}^{n} x^j = \frac{1}{n}X\mathbf{1} \\
S &= \sum_{j=1}^{n} (x^j - b) \otimes (x^j - b)' \\
&= (X - b)(X - b)'
\end{aligned}
$$

u Normal, mean b, precision nR.

S Wishart, n d.o.freedom, precision R.

$$
\begin{aligned}
N(u \,|\, n, b, R) &= (\frac{n}{2\pi})^{d/2}|R|^{1/2} \\
&\quad \exp(\, -\frac{n}{2}(u-b)'R(u-b)\,) \\
W(S \,|\, e, R) &= c^{-1}\,|S|^{(e-d-1)/2} \\
&\quad \exp(\, -\frac{1}{2}\mathrm{tr}(S\,R)\,)
\end{aligned}
$$

X, unknown mean and precision, b, R

$$\begin{aligned} u &= (1/n)X\mathbf{1} \\ S &= (X-u)(X-u)' \end{aligned}$$

Posterior Normar-Wishart distribution:

$$\begin{aligned} & NW(b, R \,|\, \ddot{n}, \ddot{e}, \ddot{u}, \ddot{S}) \\ &= W(R \,|\, \ddot{e}, \ddot{S}) \; N(b \,|\, \ddot{n}, \ddot{u}, R) \\ \ddot{n} &= \dot{n} + n \\ \ddot{e} &= \dot{e} + n \\ \ddot{u} &= (nu + \dot{n}\dot{u})/\ddot{n} \\ \ddot{S} &= S + \dot{S} + \frac{n\dot{n}}{n+\dot{n}}(u-\dot{u}) \otimes (u-\dot{u})' \end{aligned}$$

One dot $\Rightarrow$ Prior parameters
Two dots $\Rightarrow$ Posterior parameters
Non-informative parameters:
$\dot{n} = 0$, $\dot{u} = 0$, $\dot{e} = 0$, $\dot{S} = 0$.

Dirichlet-Multinomial distribution:

$$M(y \,|\, n, w) = \frac{n!}{y_1! \ldots y_m!} (w_1)^{y_1} \ldots (w_m)^{y_m}$$

$$D(w \,|\, y) = \frac{\Gamma(y_1 + \ldots + y_k)}{\Gamma(y_1) \ldots \Gamma(y_k)} \prod_{k=1}^{m} {w_k}^{y_k - 1}$$

$w > \mathbf{0}$, $w\mathbf{1} = 1$.

Posterior: $\ddot{y} = \dot{y} + y$.

Non-informative prior: $\dot{y} = \mathbf{1}$.

Finally, Dirichlet-Normal-Wishart posterior,

$$f(\theta \,|\, X, \dot{\theta}) = f(X \,|\, \theta) f(\theta \,|\, \dot{\theta})$$

$$f(X \,|\, \theta) = \prod_{j=1}^{n} \sum_{k=1}^{m} p_k^j w_k N(x^j \,|\, b^k, R^k)$$

$$f(\theta \,|\, \dot{\theta}) = D(w \,|\, \dot{y}) \prod_{k=1}^{m} NW(b^k, R^k \,|\, \dot{n}_k, \dot{e}_k, \dot{u}^k, \dot{S}^k)$$

$$p_k^j = \frac{w_k N(x^j \,|\, b^k, R^k)}{\sum_{k=1}^{m} w_k N(x^j \,|\, b^k, R^k)}$$

and completed posterior:

$$
\begin{aligned}
& f(\theta \,|\, X, Z, \dot{\theta}) = f(\theta \,|\, X, Z) f(\theta \,|\, \dot{\theta}) = \\
& \quad = \; D(w \,|\, \ddot{y}) \prod_{k=1}^{m} NW(b^k, R^k \,|\, \ddot{n}_k, \ddot{e}_k, \ddot{u}^k, \ddot{S}^k)
\end{aligned}
$$

$$
\begin{aligned}
y &= Z\mathbf{1} \quad , \quad \ddot{y} = \dot{y} + y \\
\ddot{n} &= \dot{n} + y \quad , \quad \ddot{e} = \dot{e} + y \\
u^k &= \frac{1}{y_k} \sum_{j=1}^{n} z_k^j x^j \\
S^k &= \sum_{j=1}^{n} z_k^j (x^j - u^k) \otimes (x^j - u^k)' \\
\ddot{u}^k &= \frac{\dot{n}_k \dot{u}^k + y_k u^k}{\ddot{y}_k} \\
\ddot{S}^k &= S^k + \dot{S}^k + \frac{\dot{n}_k y_k}{\ddot{n}_k} (u^k - \dot{u}^k) \otimes (u^k - \dot{u}^k)'
\end{aligned}
$$

Model: D-M-N-W mixture, $d = 2, m = 2$,
FBST for Model Selection: $H : m = 1$

Optimization step:
Local: EM or Box-Quacan
Global: MCMC + Cluster Filter, SEM, etc.

Integration step: MCMC

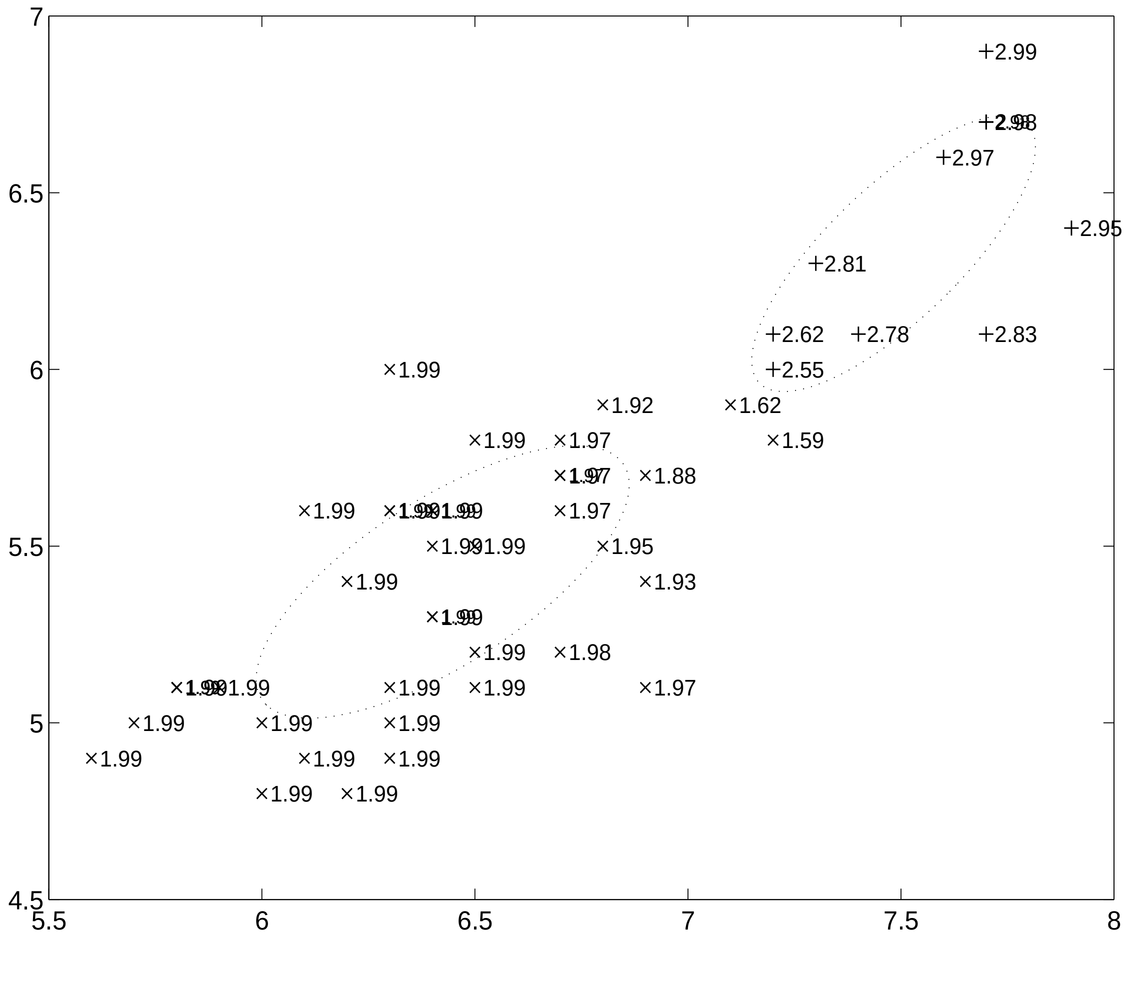
$$
\begin{aligned}
f(z^j \mid p^j) &= M(z^j \mid 1, p^j) \\
f(w \mid Z, \dot{y}) &= D(w \mid \ddot{y}) \\
f(R^k \mid X, Z, \dot{e}_k, \dot{S}^k) &= W(R \mid \ddot{e}_k, \ddot{S}^k) \\
f(b^k \mid X, Z, R^k, \dot{n}_k, \dot{u}^k) &= N(b \mid \ddot{n}_k, \ddot{u}^k, R^k)
\end{aligned}
$$

Label Switching: $perm([1 \ldots m])$
Break all non-identificability symmetries.
Ex: Order components by linear combination of vector means, $c' \, b^k$.

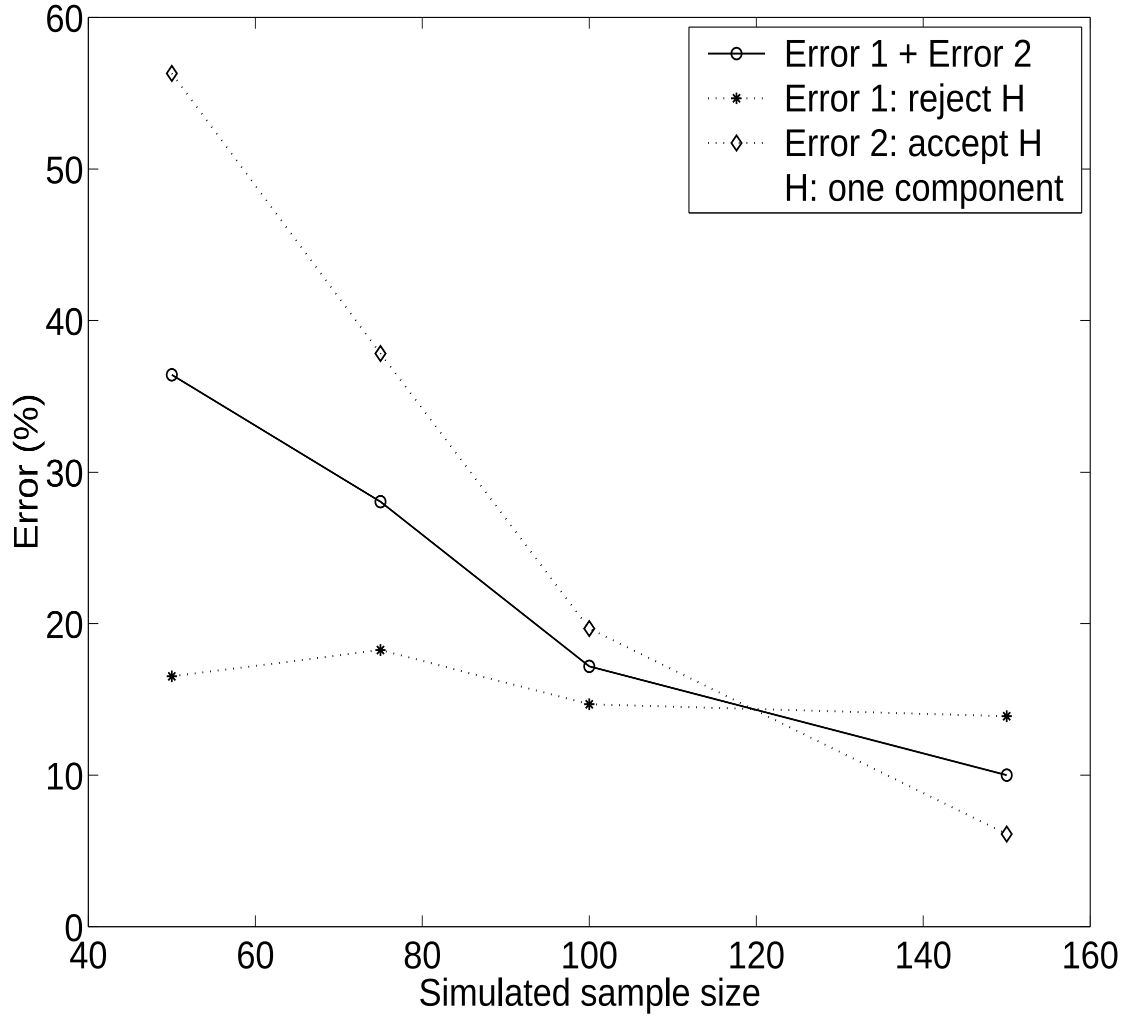
Avoid Trapping states, near-singular V^k
- Vague Priors: Empirical approx.
Forbidden States: - Rejection rules

How many components in the Iris data?

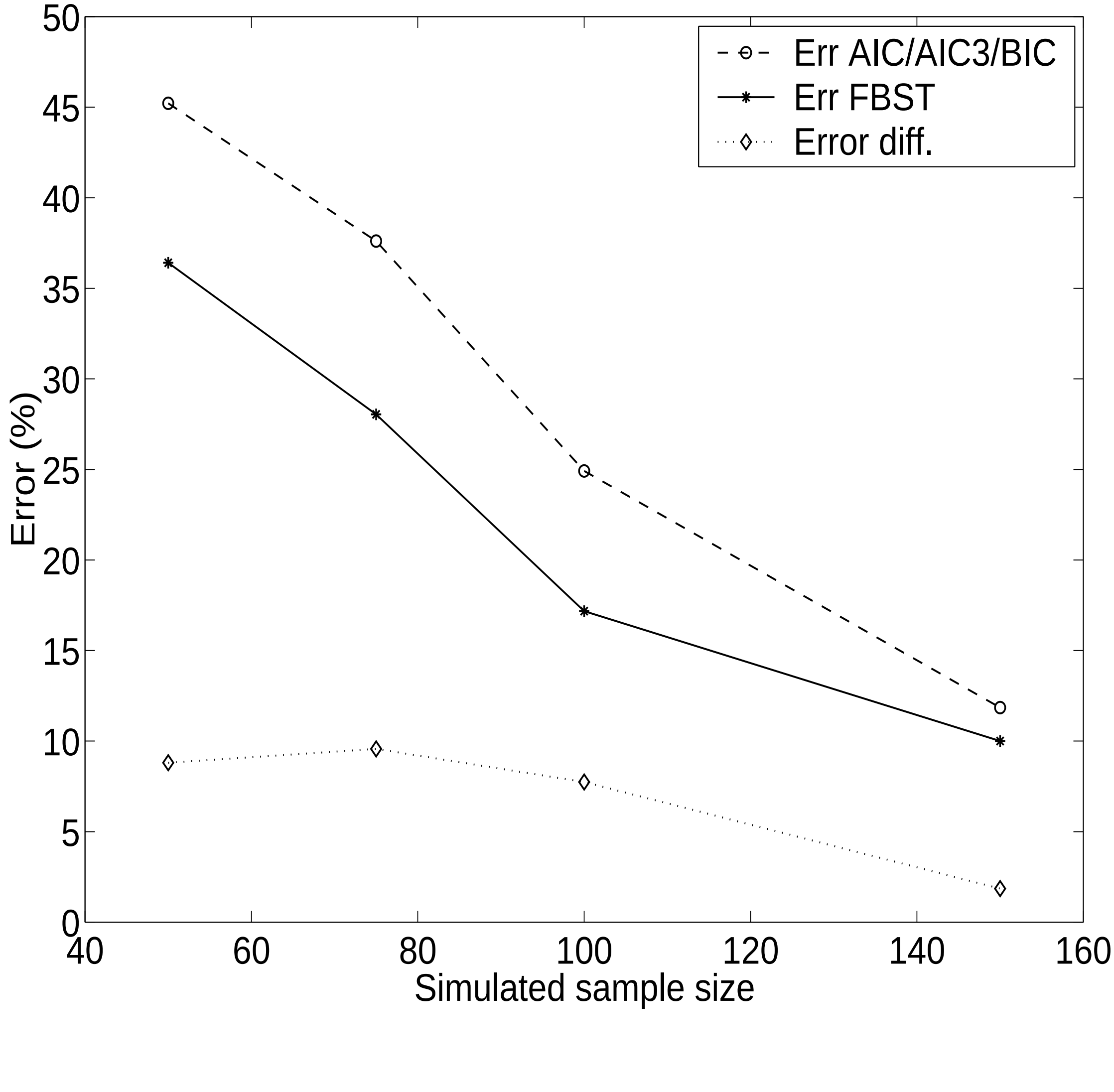
Sepal + pedal length, 49 points



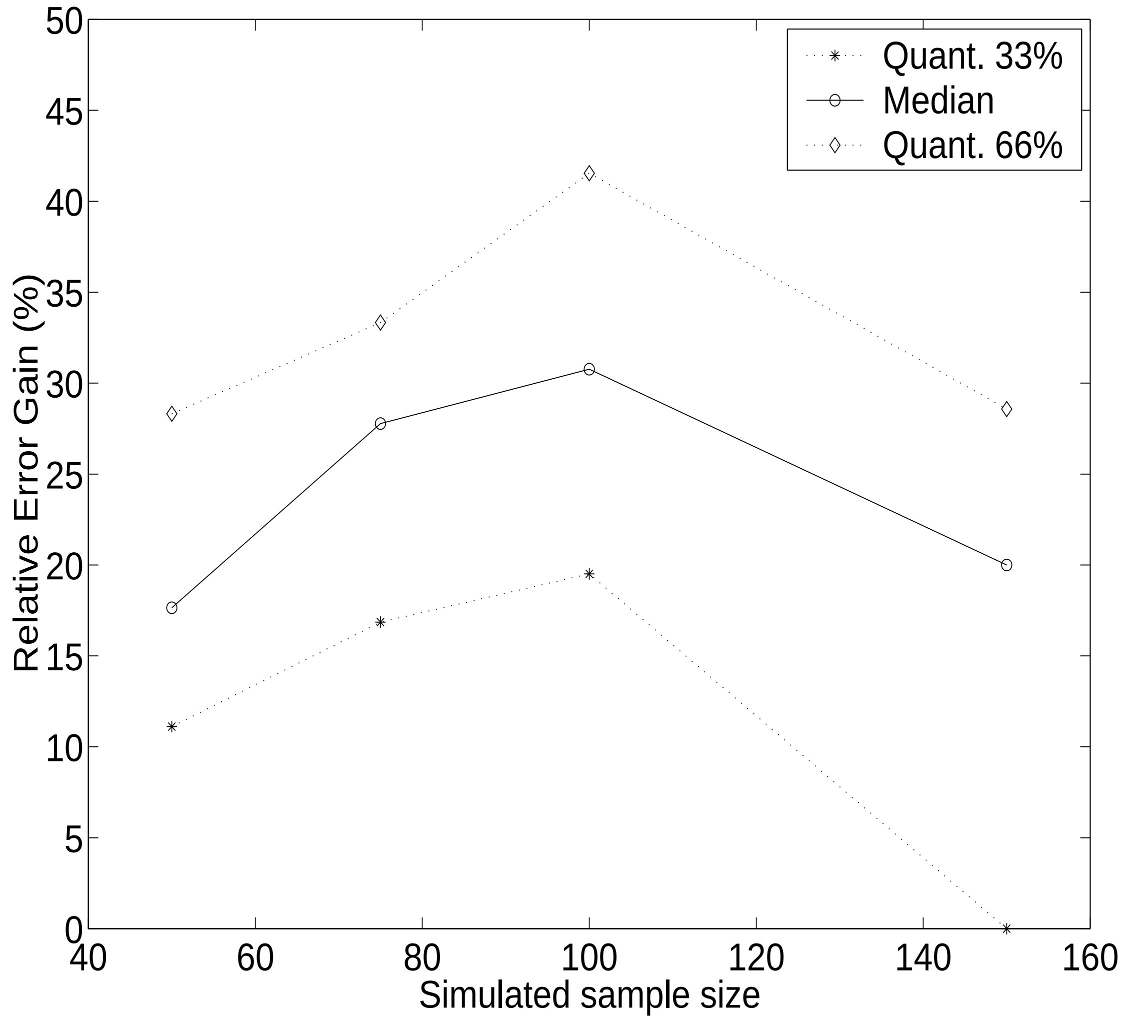
Analysis of empirical errors for FBST:
Sample sizes $n = 50, 75, 100, 150$
α: type 1 error , β: type 2 error
Total error: $\alpha + \beta$



Comparison of empirical errors:
Better of AIC, AIC3, BIC vs. FBST
Sample sizes $n = 50, 75, 100, 150$



Media, 0.33 and 0.66 quantiles for error rates differences



Conclusions:

FBST: Smaller error in model selection (up to 35%) for small (< 150) samples but computationally more expensive

AIC: Same error for large (> 150) samples and computationally cheaper

Further Research:

Mixture of components of different type,

$$c(j) = k \Rightarrow x^j \sim f_k \; (x^j \,|\, \psi_k)$$

Tests for Separate Hypotheses.

Appendix:
Critical evidence for rejecting H

Given a sample X_0, we must establish a critical level cl such that

$$\text{if } Ev(H) > cl \text{ then reject H}$$

Optimal parameters:

$$\begin{aligned} \theta^* &= \arg\max_{\theta \in \Theta_H} f(\theta \mid X_0) \\ &= [w*, b^{k*}, R^{k*}] \\ \widehat{\theta} &= \arg\max_{\theta \in \Theta} f(\theta \mid X_0) \\ &= [\widehat{w}, \widehat{b^k}, \widehat{R^k}] \end{aligned}$$

Simulation of new samples $\{^1_l X\}$ and $\{^2_l X\}$:

$$\begin{aligned} f(^1_l z^j \,|\, \theta^*) &= M(^1_l z^j \,|\, 1, w^*) \\ f(^1_l x^j \,|\, ^1_l z^j, \theta*) &= N(^1_l x^j \,|\, b^{k*}, R^{k*})^1_l z^j \\ f(^2_l z^j \,|\, \widehat{\theta}) &= M(^2_l z^j \,|\, 1, \widehat{w}) \\ f(^2_l x^j \,|\, ^2_l z^j, \widehat{\theta}) &= N(^2_l x^j \,|\, \widehat{b^k}, \widehat{R^k})^2_l z^j \end{aligned}$$

Type 1 error (α) is computed over $\{^1 X\}$
Type 2 error (β) is computed over $\{^2 X\}$

Calibrate cl that minimizes $\alpha + \beta$.

Small run $l = 1, \ldots, r$, poor calibration.