Robust Bayesian Hypothesis Testing

John T. Ormerod
(john.ormerod@sydney.edu.au)

School of Mathematics and Statistics
University of Sydney
Sydney 2006, NSW, Australia
Motivation

- One of the main projects of the Binformatics group in the School of Mathematics and Statistics at the University of Sydney is to predict survival times of Stage III melanoma patients.

- For stage III melanoma there are two main survival groups.
  - A poor prognosis group (survival times less than 1 year); and
  - A good prognosis group (survival times more than 4 years).

- We want to test whether there are differences in gene expression levels between the poor and good prognosis groups.
Motivation

□ Let $x_1, \ldots, x_n$ be the gene expression levels for a single gene for $n$ patients.
□ Let $y_1, \ldots, y_n$ be the survival class indicators for each patient.
□ Differential expression is the term used to describe when there are differences in the mean expression level between two classes, e.g., determined by testing

\[ H_0: x_i \sim N(\mu, \sigma^2) \quad \text{vs} \quad H_1: \begin{cases} x_i \sim N(\mu_0, \sigma^2) & \text{if } y_i = 0; \\ x_i \sim N(\mu_1, \sigma^2) & \text{if } y_i = 1, \end{cases} \]

with $\mu_0 \neq \mu_1$, e.g., via a two sample $t$-test.
□ Differential expression has been the primary method used to find biomarkers for various phenotypes.
Motivation

- Recent cancer studies have shown that when transcription factors or epigenetic signals become deregulated, a change in expression variability of target genes is frequently observed.

- Differential distribution is the term used to describe when there are differences in the distribution of expression levels between two classes, e.g., determined by testing

  \[
  H_0: x_i \sim N(\mu, \sigma^2) \quad \text{vs} \quad H_1: \begin{cases} 
  x_i \sim N(\mu_0, \sigma_0^2) & \text{if } y_i = 0; \text{ and} \\
  x_i \sim N(\mu_1, \sigma_1^2) & \text{if } y_i = 1,
\end{cases}
\]

  with \( \mu_0 \neq \mu_1 \) and \( \sigma_0^2 \neq \sigma_1^2 \).

- The likelihood ratio test statistic is

  \[
  \lambda = n \log(\hat{\sigma}^2) - n_0 \log(\hat{\sigma}_0^2) - n_1 \log(\hat{\sigma}_1^2) \quad (\sim \chi^2_2 \text{ under } H_0).
\]

- This test is extremely fragile to even a single outlier.
LRT – “Most different” genes – Stage III melanoma
Bayesian hypothesis testing
Bayes factors and Posterior odds

- Hypothesis test:
  \[ H_0: p(x|\theta_0, H_0) \text{ versus } H_1: p(x|\theta_1, H_1), \]

- Parameter priors: \( p(\theta_0|H_0) \) and \( p(\theta_1|H_1) \).

- Hypothesis priors: \( p(H_0) \) and \( p(H_1) \).

- Bayes factor: (analogous to likelihood ratio statistic)
  \[
  BF_{01}(x) = \frac{p(x|H_0)}{p(x|H_1)} = \frac{\int p(x|\theta_0, H_0)p(\theta_0|H_0)d\theta_0}{\int p(x|\theta_1, H_1)p(\theta_1|H_1)d\theta_1}.
  \]

- Posterior odds of \( H_0 \) to \( H_1 \) is defined by
  \[
  PO_{01}(x) = BF_{01}(x) \times \frac{p(H_0)}{p(H_1)},
  \]

  where the factor \( p(H_0)/p(H_1) \) is the prior odds. Assume prior odds is 1 so we can focus on the Bayes factor.
Bayesian hypothesis testing via Bayes factors

A Bayesian hypothesis test $T(x) \in \{0, 1\}$ is based on

$$T(x) = I(\lambda_{Bayes} > 0)$$

where $\lambda_{Bayes} = -2 \ln BF_{01}(x)$, i.e.,

$$T(x) = \begin{cases} 
1 & H_1 \text{ is preferred} \\
0 & H_0 \text{ is preferred} 
\end{cases}$$
Bayes factors and strength of evidence

Kass and Raftery (1995) offer the following interpretation of Bayes factors in terms of strength of evidence (slight modification of Jeffreys (1961) table).

<table>
<thead>
<tr>
<th>$\lambda_{\text{Bayes}}$</th>
<th>$BF_{10}$</th>
<th>Strength of evidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 to 2</td>
<td>1 to 3</td>
<td>not worth more than a bare mention</td>
</tr>
<tr>
<td>2 to 6</td>
<td>3 to 20</td>
<td>positive</td>
</tr>
<tr>
<td>6 to 10</td>
<td>20 to 150</td>
<td>strong</td>
</tr>
<tr>
<td>&gt; 10</td>
<td>&gt; 150</td>
<td>very strong</td>
</tr>
</tbody>
</table>
Bayesian hypothesis testing using diffuse priors

Let $X_i \mid \mu \sim N(\mu, \sigma^2)$ ($\sigma^2$ known). We want to test

$$H_0: \mu = \mu_0 \quad \text{versus} \quad H_1: \mu \neq \mu_0.$$  

All parameters have fixed values under $H_0$ (no parameters $\Rightarrow$ no priors). Under the alternative $\mu \mid H_1 \sim N(\mu_0, \tau^2)$.

Then

$$BF_{01}(x) = \sqrt{1 + \frac{n\tau^2}{\sigma^2}} \exp \left[ -\frac{n z(x)^2}{2(n + \sigma^2/\tau^2)} \right],$$

where $z(x) = \sqrt{n}(\bar{x} - \mu_0)/\sigma$ is the standard $z$-test statistic.

\begin{itemize}
\item \textbf{Problem:} $\lim_{\tau^2 \to \infty} BF_{01} \to \infty$.
\end{itemize}
Bayesian hypothesis testing with diffuse priors

Problem is associated with Barlett’s paradox (Lindley, 1957; Bartlett, 1957).

Barlett’s paradox: As priors become increasingly diffuse, in an attempt to be uninformative about model parameters, the simpler model is increasingly preferred. These problems were identified as early as Jeffreys (1935, 1946).
Cake prior

Suppose that we want to test

\[ H_0 : p(x|\theta_0, H_0) \quad \text{versus} \quad H_1 : p(x|\theta_1, H_1). \]

We need priors for \( p(\theta_i|H_i), i = 0, 1. \)

- We would like to have diffuse priors.
  - Non-informative.
  - Parameter estimates often mimic MLEs.

- However, due to Bartlett’s paradox, if we did we would not be able to perform hypothesis testing.

- It seems we cannot do both at once: we cannot have our cake and eat it too.
Cake prior – Ingredients

1. Let \( p(\theta_i) \equiv p(\theta_i|H_i), \ i = 0, 1 \). Define the parameter priors

\[
p(\theta_i|g_i) = \exp \left[ -\frac{p_i}{2} \ln(2\pi g_i) + \frac{1}{2} \ln |P_i(\theta_i)| - \frac{1}{2g_i} \theta_i^T P_i(\theta_i) \theta_i \right],
\]

where \( P_i(\theta_i) \) is a prior precision matrix (usually Fisher Information matrix). Assume \( p_0 \leq p_1 \).

2. Set \( g_i = h^{1/p_i} \).

3. Calculate the Bayes Factor as

\[
BF_{01}(h) = \frac{\int p(x|\theta_0)p(\theta_0|h^{1/p_0})d\theta_0}{\int p(x|\theta_1)p(\theta_1|h^{1/p_1})d\theta_1}.
\]

4. Optional: Let \( h \to \infty \) if flat priors are desired.
Cake prior – How it works

Setting \( g_i = h^{1/p_i} \) we have

\[
p(\theta_i; h) = \exp \left[ -\frac{p_i}{2} \ln(2\pi) - \frac{1}{2} \ln(h) + \frac{1}{2} \ln |P_i(\theta_i)| - \frac{\theta_i^T P_i(\theta_i) \theta_i}{2h^{1/p_i}} \right].
\]

The Bayes factor then becomes

\[
BF_{01}(h) = \frac{\int \frac{|2\pi P_0(\theta_0)|^{1/2} p(x|\theta_0) \exp \left[ -\frac{1}{2} \ln(h) - \frac{\theta_0^T P_0(\theta_0) \theta_0}{2h^{1/p_0}} \right]}{\int \frac{|2\pi P_1(\theta_1)|^{1/2} p(x|\theta_1) \exp \left[ -\frac{1}{2} \ln(h) - \frac{\theta_1^T P_1(\theta_1) \theta_1}{2h^{1/p_1}} \right]}{d\theta_0}}
\]

\[
h \to \infty \quad \frac{\int |2\pi P_0(\theta_0)|^{1/2} p(x|\theta_0) d\theta_0}{\int |2\pi P_1(\theta_1)|^{1/2} p(x|\theta_1) d\theta_1}
\]

The prior variance does not dominate the Bayes factor!
Robust Bayesian testing
Bayesian robustness

Often papers replace normal models with either

- Student’s $t$ distributions; or
- Laplace distributions.

These models lead to robust estimators of location, but do not lead to robust estimators of scale.

Our approach is to use a mixture component to pick up outliers.
The model

We will consider the model

\[
p(x_i|y_i, z_i, H_0, H_1, \theta) = \left[ \left\{ \phi_\sigma(x_i - \mu) \right\}^{H_0} \left\{ \phi_{\sigma_1}(x_i - \mu_1)^{y_i} \phi_{\sigma_0}(x_i - \mu_0)^{1-y_i} \right\}^{H_1} \right] z_i \left[ \phi_{\tilde{\sigma}}(x_i - \tilde{\mu}) \right]^{1-z_i}
\]

- $x_i$ – gene expression level for sample $i$ (observed).
- $y_i \in \{0, 1\}$ – survival indicator for sample $i$ (observed).
- $z_i \in \{0, 1\}$ – outlier indicator for sample $i$ (latent).
- $H_0 \in \{0, 1\}$ – indicator for null hypothesis ($H_1 = 1 - H_0$) (latent).
- $\theta = (\mu, \mu_0, \mu_1, \sigma^2, \sigma_0^2, \sigma_1^2)$ – likelihood parameters (latent).
- $\tilde{\mu}$ and $\tilde{\sigma}$ fixed with $\tilde{\sigma}$ large – mixture component for outliers (observed).
Priors

Using Cake priors, for the null hypothesis we arrive at

\[ \mu | \sigma^2 \sim N(0, g_0 \sigma^2), \quad \text{and} \]
\[ \sigma^2 \sim LN(0, 2g_0). \]

For the alternative hypothesis we use

\[ \mu_0 | \sigma_0^2 \sim N \left( 0, \frac{g_1n}{n_0} \sigma_0^2 \right), \quad \mu_1 | \sigma_1^2 \sim N \left( 0, \frac{g_1n}{n_1} \sigma_1^2 \right), \]
\[ \sigma_0^2 \sim LN \left( 0, \frac{2ng_1}{n_0} \right), \quad \text{and} \quad \sigma_1^2 \sim LN \left( 0, \frac{2ng_1}{n_1} \right). \]

Also,

\[ z_i \sim Bernoulli(\rho) \quad \text{and} \quad \rho \sim Uniform(0, 1). \]
Bayesian approach

If we let \( g_0 = c_0 h^{1/2} \), \( g_1 = c_1 h^{1/4} \) and \( h \to \infty \) then

\[
p(z_i | \text{rest}) \approx \left[ \{ \phi_{\hat{\sigma}}(x_i - \hat{\mu}) \}^{H_0} \{ \phi_{\hat{\sigma}_1}(x_i - \hat{\mu}_1)^y_i \phi_{\hat{\sigma}_0}(x_i - \hat{\mu}_0)^{1-y_i} \}^{H_1} \right]^{z_i}
\times \left[ \phi_{\bar{\sigma}}(x_i - \bar{\mu}) \right]^{1-z_i} \rho^{z_i} (1 - \rho)^{1-z_i},
\]

where all “hatted” values are

\[
\hat{\mu} = \frac{\sum_{i=1}^{n} z_i x_i}{\sum_{i=1}^{n} z_i},
\]

\[
\hat{\mu}_1 = \frac{\sum_{i=1}^{n} z_i y_i x_i}{\sum_{i=1}^{n} z_i y_i},
\]

\[
\hat{\mu}_0 = \frac{\sum_{i=1}^{n} z_i (1- y_i) x_i}{\sum_{i=1}^{n} z_i (1- y_i)},
\]

\[
\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} z_i (x_i - \hat{\mu})^2}{\sum_{i=1}^{n} z_i},
\]

\[
\hat{\sigma}_1^2 = \frac{\sum_{i=1}^{n} z_i y_i (x_i - \hat{\mu}_1)^2}{\sum_{i=1}^{n} z_i y_i},
\]

\[
\hat{\sigma}_0^2 = \frac{\sum_{i=1}^{n} z_i (1- y_i) (x_i - \hat{\mu}_0)^2}{\sum_{i=1}^{n} z_i (1- y_i)},
\]

and \( \hat{\rho} = \sum_{i=1}^{n} z_i \). Under some assumptions

\[
p(z_i = 0 | \mathbf{x}, \mathbf{z}_-i) \to 1 \quad \text{as} \quad x_i \to \pm \infty.
\]
Bayesian approach

Hence, if \( x_i \) is an “obvious” outlier it will be detected as one. The Bayes factor can be written as a mixture of likelihood ratio statistics,

\[
BF_{01} = \sum_z \exp \left[ -\frac{1}{2} \lambda(z) + \frac{1}{2} \nu \log(n) + C \right] p(z|x),
\]

where

- \( \sum_z \) is a combinatorial sum over all possible “outlier” configurations \( z \),
- \( \lambda(z) \) is a likelihood ratio test statistic for configuration \( z \),
- \( \nu = 2 \) is the d.o.f. of the likelihood ratio test, and
- \( C = \log(c_0/c_1) \) is a controllable constant.
Bayesian approach

Then $p(z^*|x) \approx 1$ for some outlier configuration $z^*$. So

$$BF_{01} \approx \exp\left[-\frac{1}{2} \lambda(z^*) + \frac{1}{2} \nu \log(n) + C\right].$$

If we choose the prior hyperparameters $c_0$ and $c_1$ such that

$$C = -\frac{1}{2} \nu \log(n) + \frac{1}{2} \chi_{\nu,1-\alpha}^2$$

then

$$BF_{01} \approx \exp\left[-\frac{1}{2} \lambda(z^*) + \frac{1}{2} \chi_{\nu,1-\alpha}^2\right]$$

and so we are choosing a priori to control type I error to be $\alpha$. 
Results
Effect of one outlier when simulating from $H_0$

Perturbation of single observation

Prob. of reject null

size of perturbation

- - - LRT

- - RBT
Simulations
$p$-value histograms

- LRT (unequal variances)
- t-test (unequal variances)
- ks-test
- Robust Bayesian test
BRT – “Most different” genes – Stage III melanoma

- **PAPOLA (6940719)**
- **BRD7 (430040)**
- **ZNF22 (6960768)**
- **ZNF22 (5860546)**
- **TYR (5260253)**
- **PQBP1 (450315)**
Conclusions

- Can avoid problems with diffuse priors using Cake priors.
- Cake priors can also be used to control type I errors.
- Robustness can be achieved by using an additional mixture component to detect outliers.
- Fast – 12,000 tests in approx 8 seconds.
- Early results, still need to figure out how to break it.
Collaborators

Michael Stewart  Weichang Yu  Sarah Romanes
Bibliography


