Borrowing strength: a likelihood ratio test for related sparse signals

Ernst C. Wit\textsuperscript{1,}\textsuperscript{*} and David J. G. Bakewell\textsuperscript{2,} \textsuperscript{[1]}

\textsuperscript{1}Johann Bernoulli Institute, University of Groningen, 9747 AG Groningen, The Netherlands and \textsuperscript{2}Electrical Engineering & Electronics, University of Liverpool, Liverpool L69 3GJ, UK

ABSTRACT

Motivation: Cancer biology is a field where the complexity of the phenomena battles against the availability of data. Often only a few observations per signal source, i.e. genes, are available. Such scenarios are becoming increasingly more relevant as modern sensing technologies generally have no trouble in measuring lots of channels, but where the number of subjects, such as patients or samples, is limited. In statistics, this problem falls under the heading 'large \( p \), small \( n \)'. Moreover, in such situations the use of asymptotic analytical results should generally be mistrusted.

Results: We consider two cancer datasets, with the aim to mine the activity of functional groups of genes. We propose a hierarchical model with two layers in which the individual signals share a common variance component. A likelihood ratio test is defined for the difference between two collections of corresponding signals. The small number of observations requires a careful consideration of the bias of the statistic, which is corrected through an explicit Bartlett correction. The test is validated on Monte Carlo simulations, which show improved detection of differences compared with other methods. In a leukaemia study and a cancerous fibroblast cell line, we find that the method also works better in practice, i.e. it gives a richer picture of the underlying biology.

Availability: The MATLAB code is available from the authors or on \textsuperscript{[2]}

1 INTRODUCTION

In this article, we develop a likelihood ratio (LR) test that aims to detect small concordant changes in a collection of related signals under two experimental conditions. Under certain conditions, it can be shown that the LR test is the most powerful test in many practical testing problems and potentially offers deeper insight into complex phenomena, e.g. the biological processes underlying cancer development. Exact calculation, however, is not always easy. Previous work \textsuperscript{[3,4]} used LR tests for evaluating changes in a single signal in the area of microarray analysis. Our approach here differs from theirs by considering multiple concordant changes. \textsuperscript{[5]} suggested to use Hotelling’s \( Y^2 \) statistic for this purpose, but this is only a viable alternative when the addition of observations exceeds the number of channels. This is in many modern genomic applications not the case. Alternatives are ANCOVA approaches \textsuperscript{[6]} and a model of gene expression estimation method for testing association with clinical outcome \textsuperscript{[7]}.

2 MODEL FOR RELATED SIGNALS

Our aim is to devise a model for the expression of replicates of \( m \)-related signals measured across two different conditions, say \( x \) and \( y \). This is a situation that is common in many high-frequency sensing data, such as in astronomy, geography and finance. The use of traditional mixed-effect models has also proved popular within functional genomics. \textsuperscript{[8,9]} proposed a simple mixed model combined with two-stage effect estimation for detecting individually differentially expressed genes. Their model allows for systematic nuisance effects; however, the estimation of the effects of interest, i.e. the signals, is performed one-by-one using individual gene-specific variances. These individual gene variances ignore the fact that there is something common about the underlying measurements. Others, such as \textsuperscript{[10]} and \textsuperscript{[11]}, have proposed models with a common variance, stressing that one can borrow strength from measurements in the other channels, but ignoring that there can still be orders of magnitude difference between measurements.

Ordinary mixed-effect models are therefore not suited for modelling sets of related, but possibly quantitatively different signals. Instead, we need to think about a model that data adaptively can move between a model that assumes individual variances for each of the signals, to one that takes a common variance. For this
This is for instance the case in microarray experiments, where each presence of ‘similar’ signals, we shall assume the unknown variance condition

be written as

and be the number of replicates for gene \( j \) of channel \( x \). Although in some cases the number of subjects per condition are the same, we do not assume this here. Let \( n_f^j \) and \( n_y^j \) be the number of replicates for gene \( j \) of condition \( x \) and \( y \), respectively.

We assume that channels measure similar kinds of quantities. This is for instance the case in microarray experiments, where each ‘channel’ is a gene-specific probe. In order to gain strength from the presence of ‘similar’ signals, we shall assume the unknown variance \( \sigma_f^2 \) of channel \( j \) to be related to the variance of the other channels on the measuring device, in that all come from a ‘common’ inverse gamma distribution, \( \sigma_f^2 \sim \Gamma^{-1}(\alpha, \delta) \). The within-channel variation for \( m \) channel can be represented as a two-layer hierarchical directed acyclic graph (DAG), shown in Figure 1. The full log-likelihood can be written as

\[
\ell(\alpha, \delta, \mu_f^j, \mu_y^j) = n \left( h(\alpha/2) - \frac{n}{2} \ln(2\pi) \right) - \left( \alpha + \frac{n}{2} \right) \sum_{j=1}^m \ln \left( \frac{\sigma_f^2}{\sigma_y^2} + \delta \right),
\]

where for notational simplicity, we introduce the quantities \( n = n_x + n_y \) and \( \zeta = \sum_{j=1}^m (\mu_f^j - \mu_y^j)^2 + \sum_{j=1}^m (\sigma_f^2 - \sigma_y^2)^2. \)

### 3 Inference

The aim is to test whether the two conditions are same or differ in one or more channels. Traditionally, one considered series of univariate tests to tackle this question. However, modern measurement instruments typically have lots of parallel channels. Besides the difficulty of inferentially combining lots of tests, the fundamental problem is that univariate tests ignore the information, resulting from the similarity of the channels. Therefore, here we consider a single test, which tests the joint equality of all the average signal strength in all the channels across the two conditions, \( H_0: \mu_f^j = \mu_y^j, \forall j \leq m \).

Asymptotically, under certain regularity conditions, the likelihood ratio statistic provides the most powerful test. However, when each of the signal sources has only a small number of observations, the distribution of such statistic can deviate heavily from its asymptotic \( \chi^2 \) distribution. The Bartlett correction is a first-order correction of likelihood ratio test statistic and involves rescaling the test statistic to have the correct \( \chi^2 \) mean under the null hypothesis with a finite sample.

The model we have described is a highly parametric model. However, inference based on the model is not very much affected by violation of these parametric assumptions. In particular, the likelihood ratio statistic will, even for skewed distributions such as log-normal distribution, be approximately \( \chi^2 \) distributed if \( n \geq 30. \) For larger \( n \) the likelihood ratio statistic will become more and more normally distributed, irrespective of the number of observations \( n. \) In realistic scenarios, such as microarray studies where the groups are defined through GO terms or KEGG categories, the group size \( m \) is typically at least 20. For a moderate number of observations, e.g. \( n = 38 \) such as in the leukaemia example in Section 4.3.2, this means that even if the underlying data are not normally distributed the likelihood ratio test is still expected to work. Often data in measurement channels are strictly positive and therefore right-tailed. In such cases, it is obviously possible to transform the data beforehand, e.g. by some power transformation or the log-transformation.

#### 3.1 Likelihood ratio statistic

In general, the logarithm of the likelihood ratio test statistic is defined as

\[
\Lambda(X, Y) = -2 \left( \sup_{\theta \in \Theta_0} \ell(\theta|X, Y) - \sup_{\theta \in \Theta} \ell(\theta|X, Y) \right),
\]

where \( \Theta_0 \) are the hypothesized values under \( H_0. \) We consider separately the maximization of the likelihood under \( H_0 \) and the full parameter space.

\[
l_0 = \sup_{\alpha, \delta, \mu_f^1 = \mu_y^1} \{ \ell(\alpha, \delta, \mu_f^1, \mu_y^1) \},
\]

\[
l_1 = \sup_{\alpha, \delta, \mu_f^1, \ldots, \mu_m} \{ \ell(\alpha, \delta, \mu_f^1, \ldots, \mu_m) \},
\]

Expressions for finding the maximum log-likelihood \( l_0 \) and \( l_1 \) as a function of \( \alpha, \delta \) and \( \hat{\mu} \) can be found by setting the partial derivatives of Equation 1 to zero, i.e. \( \frac{\partial \ell}{\partial \alpha} = 0 \) and \( \frac{\partial \ell}{\partial \mu} = 0 \) lead to

\[
\hat{\mu}^j = \frac{1}{n_x} \sum_{i=1}^{n_x} y_{ij}. \]

\[
\hat{\mu}^j = \frac{1}{n_y} \sum_{j=1}^{n_y} y_{ij}. \]
for the unconstrained solution \( l_1 \), whereas for the constrained solution \( l_0 \), we have

\[
\hat{\mu}_j^0 = \frac{\hat{\alpha}_j \sum_{i=1}^{n_j} y_i + \hat{\delta}_j \sum_{i=1}^{n_j} x_i}{n_j}
\]  

(5)

Furthermore, \( \hat{\alpha}_j (\alpha, \delta, \hat{\mu}_j) = 0 \) results in

\[
\hat{\delta}(\delta) = n \left( \frac{m}{2} \sum_{j=1}^{m} \left( \frac{m}{2} \ln(\hat{\delta}(\delta)) - \frac{n \ln 2 \pi}{2} \right) - 1 \right)
\]  

(6)

and \( \lambda_0 (\alpha, \delta, \hat{\mu}_j) = 0 \) leads to

\[
m(\hat{\phi}(\alpha + \frac{n}{2} - \hat{\phi}(\alpha)) - \sum_{j=1}^{m} \ln(\hat{\delta}) + 1) = 0,
\]  

(7)

where the ‘digamma’ or ‘psi’ function is defined \( \psi(x) = \partial(\ln(\Gamma(x))) / \partial x \). By substituting Equation (6) respectively Equation (5) in \( \hat{\delta}_j \), defined below (1), for all \( j \), and by substituting Equation (3) in Equation (1), we obtain expressions for \( l_0 \) and \( l_1 \) that only depend on \( \delta \), i.e.

\[
l(\delta) = m \left( \ln(\hat{\delta}(\delta)) + \frac{n_m \ln 2 \pi \delta}{2} \right) - \frac{1}{2} \sum_{j=1}^{m} \ln(\hat{\delta}) + 1
\]  

(8)

where \( \hat{\delta}_j \) is defined with respect to the constrained and unconstrained estimates \( \hat{\mu}_j^0 \) and \( \hat{\mu}_j^1 \) for \( l_0 \) and \( l_1 \), respectively. The procedure for finding the supremum of the log-likelihood \( l(\delta) \), i.e. \( l_0 \) and \( l_1 \), is an iterative process. The initialization step finds the approximate location of \( \delta_{k,\text{MLE}} \) as \( \delta_0 \) by maximizing (8) over some grid \( [0, \delta_1, \ldots, \delta_{n_m}] \), where, e.g. \( \delta_{n_m} \approx 10^{10} \). Then, we progressively refine the starting point using Newton-Raphson (NR).

\[
\delta_{k+1} = \delta_k - \frac{\frac{\partial l(\delta_k)}{\partial \delta}}{\frac{\partial^2 l(\delta_k)}{\partial \delta^2}}
\]

Computationally less demanding, but leading to the same results, is iteratively solving Equation (8) for \( \delta \) by fixing \( \alpha = \hat{\alpha}(\delta_k) \) using NR and then replacing \( \alpha \) by \( \hat{\alpha}(\delta_{k+1}) \). The Nelder-Mead (NM) simplex method [Press et al. 1992], that does not require the explicit use of partial derivatives, provides a robust alternative to NR iteration. Using the same starting values \( \delta_0 \), the NM yielded excellent agreement with NR.

In cases where the maximum likelihood for \( \delta \) achieves \( \delta_{k,\text{MLE}} = \delta_{\text{max}} \), then it is clear from Equation (8) that also \( \hat{\delta} \) tends to be large. Practically, this means that the distribution of \( \sigma_j^2 \) is close to being degenerated in a single point around its mean \( \delta(\alpha - 1) \). In other words, each channel has the same signal variance. By letting \( \delta \to \infty \), Equation (8) simplifies to \( \bar{\sigma} = n m \hat{\delta} / \sum_{j=1}^{m} \hat{\delta}_j \), so that \( \forall j \in [1, \ldots, m] \) the channel variance estimate degenerates in

\[
\sigma^2 \approx \frac{\delta}{\alpha}
\]

\[
= 1 \sum_{j=1}^{m} \sum_{i=1}^{n_j} \left( \frac{n_j}{n_j - 1} \sum_{i=1}^{n_j} (y_i - \mu_j^0)^2 + \sum_{i=1}^{n_j} (y_i - \mu_j^1)^2 \right)
\]

which is the usual pooled ML estimate of the variance. In order to be able to evaluate the values of \( l_0 \) and \( l_1 \) within the likelihood ratio test statistic, we require the value of log-likelihood under these estimates. Applying an asymptotic expansion for the gamma function [Abramowitz and Stegun 1965] to Equation (1), the log-likelihood at the maximum is given by.

\[
\lim_{\delta \to \infty} l(\hat{\delta}) = \frac{nm}{2} \ln \left( \frac{nm}{2 \pi \delta} \right) - 1
\]

(9)

### 3.2 Bartlett correction

For large samples, \( n = n_1 + n_2 \), the distribution of \( \lambda(x, y) \) is approximately distributed like a \( x^2_2 \) distribution, where \( q = \text{dim} \Theta - \text{dim} \Theta_0 \) is the difference in the number of free parameters in \( \Theta \) and \( \Theta_0 \). We define the Bartlett correction as

\[
BC = \text{ExpE}(m / \lambda_{\Theta_0}(x, y)) / m
\]

\[
= -2 E_{\Theta_0}(\lambda_{\Theta_0}(x, y)) / m
\]

(10)

with \( l_0 \) and \( l_1 \) defined in Equations (4) and (5). By defining the small sample likelihood ratio statistic as \( \lambda_{\Theta_0}(x, y) = \lambda(x, y) / BC \), we achieve that precisely the Bartlett corrected likelihood ratio statistic has \( E_{\Theta_0}(\lambda_{\Theta_0}(x, y)) = m \) as would be expected for a \( x^2_2 \) distribution. Calculation of these two expectations in general is very involved. If we use the characterization for \( \lim_{\delta \to \infty} l(\hat{\delta}) \) defined in Equation (8), we can get an explicit approximate expression for the Bartlett correction.

\[
BC \approx -2 \text{ExpE}(\lim_{\delta \to \infty} l_0(\delta) - \lim_{\delta \to \infty} l_1(\delta)) / m
\]

\[
= n \ln \left( 1 + \frac{\sum_{j=1}^{m} S_{\Lambda,j}}{\sum_{j=1}^{m} (S_{j,j} + S_{\Lambda,j})} \right)
\]

(11)

(12)

whereby

\[
\varepsilon_j^0 = \sum_{i=1}^{n_j} (y_i - \hat{\mu}_j^0)^2 + \sum_{i=1}^{n_j} (y_i - \hat{\mu}_j^1)^2
\]
Likelihood ratio test for related sparse signals

\[ \chi^2_j = \sum_{i=1}^{n_j} (x_{ij} - \mu_j)^2 / \nu_j + \sum_{i=1}^{n_j} (y_{ij} - \nu_j)^2 / \nu_j \]

with \( \mu_j = n_j \bar{\mu} / n + \mu_j \) and \( \nu_j = 1/n \sum_{i=1}^{n_j} \nu_i \). Since for small values of \( m \), the supremum of the likelihood in Equations 13 and 14 are actually found at that degenerate \( \delta \), the approximation in Equation 15 can be exact in certain cases. The log expressions are normally distributed, so that under the assumption of no differential correction. We compare the three forms of the Bartlett correction for \( \alpha = 0.05 \). The values of the target BC correction columns are Monte Carlo sample averages of the true BC factor based on 300 and 10,000 Monte Carlo runs of the likelihood ratio statistic under \( \alpha = 0.05 \). The approximation BC correction columns are the three derived approximations, where Equation 18 is the simplest and most conservative one.

### 3.3 Evaluation of the Bartlett correction

In this section, we test the small sample behaviour of the Bartlett correction. We compare the three forms of the Bartlett correction with Monte Carlo simulations of the same value. We also check the distribution of the Bartlett-corrected likelihood ratio statistic under the null hypothesis of no differential signals in all of the channels.

### Table 1. Comparison of Bartlett correction approximations

<table>
<thead>
<tr>
<th>( n )</th>
<th>( m )</th>
<th>( \alpha = 4 )</th>
<th>( \delta = 0.25 )</th>
<th>( \delta = 0.1 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>5</td>
<td>1.02</td>
<td>1.73</td>
<td>1.76</td>
</tr>
<tr>
<td>5</td>
<td>6</td>
<td>1.62</td>
<td>1.40</td>
<td>1.40</td>
</tr>
<tr>
<td>8</td>
<td>5</td>
<td>1.35</td>
<td>1.27</td>
<td>1.27</td>
</tr>
<tr>
<td>16</td>
<td>5</td>
<td>1.20</td>
<td>1.11</td>
<td>1.12</td>
</tr>
<tr>
<td>30</td>
<td>2</td>
<td>1.10</td>
<td>1.07</td>
<td>1.07</td>
</tr>
<tr>
<td>5</td>
<td>1.10</td>
<td>1.06</td>
<td>1.06</td>
<td>1.05</td>
</tr>
</tbody>
</table>

The values of the target BC correction columns are Monte Carlo sample averages of the true BC factor based on 300 and 10,000 Monte Carlo runs of the likelihood ratio statistic under \( \alpha = 0.05 \). The approximation BC correction columns are the three derived approximations, where Equation 18 is the simplest and most conservative one.

### 3.4 Dependent channels

The Bartlett-corrected likelihood ratio test (BC-LRT) we proposed in this section makes some allowance for dependence between the channels. In fact, the common variance distribution induces some dependence on the measurements within the same channel. However, conditionally on the variance the data from the individual channels are assumed to be independent. This may be unproblematic in many practical circumstances, especially when \( m \) is small and the channels show only small correlation. However, in many circumstances the dependence between the channels may be substantial. For example, voxels on a MRI scan or messenger RNA (mRNA) data from genes with a common transcription factor will show high interdependence. In such cases, we should make allowance for the fact that the information that comes from the various channels cannot be considered \( m \) pieces of separately supporting evidence. In this section, we describe how this impacts...
Table 2. Subset of $P$-values for the test of differences between ALL and AML leukemia (cf Section 4.3.2) for the BC-LRT without and with ‘dependence correction’ (DC)

<table>
<thead>
<tr>
<th>Gene ID</th>
<th>m</th>
<th>LRT</th>
<th>$m^*$</th>
<th>LRT-DC</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO:003697</td>
<td>19</td>
<td>0.0001</td>
<td>13</td>
<td>0.0007</td>
</tr>
<tr>
<td>GO:004725</td>
<td>19</td>
<td>0.0001</td>
<td>14</td>
<td>0.0006</td>
</tr>
<tr>
<td>GO:005271</td>
<td>27</td>
<td>0.0009</td>
<td>16</td>
<td>0.0019</td>
</tr>
<tr>
<td>GO:005244</td>
<td>19</td>
<td>0.0003</td>
<td>12</td>
<td>0.0027</td>
</tr>
<tr>
<td>GO:004842</td>
<td>20</td>
<td>0.0009</td>
<td>14</td>
<td>0.0040</td>
</tr>
<tr>
<td>GO:005089</td>
<td>13</td>
<td>0.0012</td>
<td>9</td>
<td>0.0052</td>
</tr>
<tr>
<td>GO:001707</td>
<td>6</td>
<td>0.0018</td>
<td>5</td>
<td>0.0036</td>
</tr>
<tr>
<td>GO:004879</td>
<td>11</td>
<td>0.0022</td>
<td>8</td>
<td>0.0068</td>
</tr>
<tr>
<td>GO:005544</td>
<td>9</td>
<td>0.0027</td>
<td>7</td>
<td>0.0064</td>
</tr>
<tr>
<td>GO:004693</td>
<td>8</td>
<td>0.0030</td>
<td>6</td>
<td>0.0077</td>
</tr>
<tr>
<td>GO:001654</td>
<td>6</td>
<td>0.0034</td>
<td>6</td>
<td>0.0086</td>
</tr>
<tr>
<td>GO:005201</td>
<td>23</td>
<td>0.0041</td>
<td>15</td>
<td>0.0147</td>
</tr>
<tr>
<td>GO:005001</td>
<td>7</td>
<td>0.0047</td>
<td>6</td>
<td>0.0075</td>
</tr>
<tr>
<td>GO:003746</td>
<td>7</td>
<td>0.0049</td>
<td>6</td>
<td>0.0078</td>
</tr>
<tr>
<td>GO:003735</td>
<td>17</td>
<td>0.0328</td>
<td>11</td>
<td>0.0631</td>
</tr>
<tr>
<td>GO:003955</td>
<td>21</td>
<td>0.0917</td>
<td>14</td>
<td>0.1298</td>
</tr>
<tr>
<td>GO:0032358</td>
<td>17</td>
<td>0.1527</td>
<td>8</td>
<td>0.2149</td>
</tr>
</tbody>
</table>

The value $m$ refers to the number of genes in the GO-class, whereas $m^*$ is the number of independent channels as estimated from the data according to Equation (18).

The likelihood-ratio statistic and how we can accommodate this in the test.

Crucially, as the likelihood ratio statistic, conditionally on $\alpha$ and $\delta$ is a sum of channel data, dependence between the channels will not affect the mean of the likelihood ratio statistic. Therefore, as the Bartlett is a mean-value correction, it, conditional on $\alpha$ and $\delta$, is also not affected by the dependence. Clearly, the shape of the distribution is affected. In the extreme case, if the data in a particular group consisted of $m$ identical copies, the likelihood ratio statistic for sufficient sample size $n$ would be a rescaled $\chi^2$ variable under $H_0$, in fact $m \times \chi^2$ distributed, rather than a $\chi^2$ random variable.

The following shows a practical guide to adjust the likelihood ratio statistic in the case of dependence between the variables. The idea is to estimate the number of independent variables by the number of channels needed to explain at least, say, 95% of the correlation in the data. This is done by considering eigenvalues of the observed correlation matrix and calculating the number of eigenvalues to exceed 95% of the total sum. Notice that if $n=m$, the rank of the correlation matrix is less than full rank and this method will always conservatively suggest dependence, whereas the data could be fully independent. In fact, when $n=m$ the method work, but will give conservative $P$-values. If one has additional information that the channels are stochastically independent and the number of observations $n$ is of the same order as or smaller than $m$, it would be better not to use the correction.

Consider the following two examples. In two groups of size $m=4$ channels, the observed correlation matrices are, respectively, given as

$$
\begin{pmatrix}
1.00 & 0.09 & 0.00 & 0.00 \\
0.99 & 1.00 & 0.00 & 0.00 \\
0.00 & 0.00 & 1.00 & 0.99 \\
0.00 & 0.00 & 0.99 & 1.00
\end{pmatrix}
\quad \text{and} \quad
\begin{pmatrix}
1.00 & 0.99 & 0.99 & 0.00 \\
0.99 & 1.00 & 0.99 & 0.00 \\
0.99 & 0.99 & 1.00 & 0.00 \\
0.00 & 0.00 & 0.00 & 1.00
\end{pmatrix}
$$

Method for dependent channel data ($n > m$)

1. Consider data $x$ and $y$ from, 2 conditions respectively, with $n_x$ and $n_y$ replicates across $m$ channels.
2. Calculate the observed $m \times m$ correlation matrix
   $$
   R = \text{Cor} \left( \begin{pmatrix} x_1 & \cdots & x_m \\ y_1 & \cdots & y_m \end{pmatrix} \right)
   $$
3. Let $\lambda_1, \ldots, \lambda_m$ be the eigenvalues of $R$.
4. Define the number of independent variables $m^*$ as the smallest value for which the relative sum of eigenvalues exceeds 95%, i.e.
   $$
   m^* = \min_k \left\{ k \mid \sum_{i=1}^k \lambda_i / \sum_{i=1}^m \lambda_i > 0.95 \right\}
   $$
5. Calculate likelihood ratio test statistic with the most conservative Bartlett correction $\Lambda^*_{(17)}$ and transform this into a reduced test statistic:
   $$
   \Lambda^*_{(17)} = \frac{m^* \Lambda^*_{(17)}}{m}
   $$
   Under $H_0$, we have that approximately $\Lambda^*_{(17)} \sim \chi^2_{m^*}$.
6. Calculate $P$-value $P(\chi^2_{m^*} \geq \Lambda^*_{(17)})$.

Then it is clear that there are really only two independent channels in both cases. This is consistent with fact that in both cases ($\lambda_1 + \lambda_2 + \cdots + \lambda_4 = 0.995 > 0.95$). We also apply this to the leukaemia example discussed in Section 4.3.2. Table 2 shows a subset of the 235 GO terms under consideration and how the potential dependence affects the power of the test.

### 4 APPLICATION

This section applies the method both to simulated and real data.

It is important to note that the computational performance of the method does not deteriorate with larger number of observations or channels. In fact, the numerical maximization described in Section 4.2 converges slower with a small number of channels, since the optimum is more likely obtained for a degenerate channel variance ($\delta, \alpha \rightarrow \infty$), which is computationally more expensive.

#### 4.1 Simulation study

In this section, we ascertain two aspects of the Bartlett-corrected test, to wit (i) its nominal coverage probability and (ii) its power.

The Hotelling $T^2$ test (Prokhorov 2001) is a natural alternative to our method as it is similarly multivariate and under full channel-variance inhomogeneity (close to) the optimal test. One important disadvantage of the Hotelling $T^2$ test is that it is only defined for a few more observations than channels, i.e. $n = m + 2$. This means that for the simulations $n_x + n_y = n = 4$, $m = 2$ and $n_x + n_y = n = 4$, $m = 5$ no Hotelling $T^2$ alternative can be calculated. This limits the use of the Hotelling $T^2$, alternative in sparse data situations. In these
Likelihood ratio test for related sparse signals

4.1.1 Is the method unbiased? In this section, we show that the method does not systematically give low $P$-values. This is essential, otherwise it would lead to an unfair comparison to the other methods. A way to test whether there is bias is to apply the method in the case where there are no differences between the channels. In that case, the resulting $P$-value should be an uniformly distributed value between 0 and 1. The ideal line in Figure 2 should therefore be a straight line going from (0,0) to (1,1).

In the first set of simulations, we consider 300 draws from the null model, according to Equation (2), in which $\alpha = 3$ and $\delta = 1/3$. Furthermore, we vary the number of observations ($n = n_x + n_y$) at two levels, 4 and 12, and the number of simultaneous channels ($m$) at 2 and 5. These are challenging conditions for inference due to a large variance heterogeneity—as a result of a small $\delta$—and a small number of observations $n$. The aim is to see whether the method gives indeed rise to approximate uniform $P$-values. The plots in Figure 2 show the results for each of the four scenarios for the raw LR test, for two version of the BC-LRT and for Hotelling $T^2$ test.

What we can see is that the most conservative Bartlett correction (17), i.e. the simple $n/(n-2)$, is very close to the nominal coverage probability in each of the simulations. Notice that for two channels $m = 2$, there is no difference between the Equations (15) and (17) and therefore only the latter is shown. The Hotelling $T^2$ test naturally achieves the nominal coverage probabilities by the very definition of a $T^2$ distribution.

4.1.2 Power of method In order to test the power of the procedure, we perform a simulation where half of the times all the $m$
Effect size through was proposed early on as a way to test a single (joint) hypothesis group of test. Fisher’s combined probability test (Manly, 1985) is a common method is to combine univariate \(P\)-values across joint hypotheses. Under the null hypothesis each \(P\)-value was independently uniform and, therefore, minus the sum of the log-transformed \(P\)-values would be \(\chi^2\) distributed with 2\(m\) degrees of freedom. Despite its easy applicability, this \(P\)-value combination method is not very often applied in a bioinformatics context. More commonly, \(P\)-values of several individual tests are ‘combined’ through some multiple testing procedure to explore whether something, and what, is happening. Of those the Bonferroni correction is most famous, but the Šidák correction is, in the case of independent \(P\)-values, more sensitive. On the other hand, Hotelling \(\mathbf{T}^2\) test was a model-based approach for testing a joint hypothesis directly. This is similar to the likelihood ratio test we proposed and the Global test (Casella and Berger 1990).

4.2 Comparison with other methods

It is crucial to compare the performance of the test with other standard methods for combining \(P\)-values across joint hypotheses. Effectively there are two approaches. Traditionally, the most common method is to combine univariate \(P\)-values into a single group \(P\)-value that represents the overall significance of those group of test. Fisher’s combined probability test (Fisher 1925) was proposed early on as a way to test a single (joint) hypothesis through \(m\) independent \(P\)-values. Under the null hypothesis each \(P\)-value was independently uniform and, therefore, minus the sum of the log-transformed \(P\)-values would be \(\chi^2\) distributed with 2\(m\) degrees of freedom. Despite its easy applicability, this \(P\)-value combination method is not very often applied in a bioinformatics context. More commonly, \(P\)-values of several individual tests are ‘combined’ through some multiple testing procedure to explore whether something, and what, is happening. Of those the Bonferroni correction is most famous, but the Šidák correction is, in the case of independent \(P\)-values, more sensitive. On the other hand, Hotelling \(\mathbf{T}^2\) test was a model-based approach for testing a joint hypothesis directly. This is similar to the likelihood ratio test we proposed and the Global test (Casella and Berger 1990).

4.3 Two microarray applications

Microarray gene expression experiments typically measure the behaviour of a large number of genes with a relatively small number of independent biological samples. The small number of independent samples limits the statistical inference that can be made about the behaviour of individual genes. Consequently, the microarray analyst is faced with bewildering lists of differentially expressed (DE) genes that are typically full of false positives—a direct result of the well-known ‘small-n-large-p’ dimensionality problem: only a few samples available and many genes to consider. At the same time biologists are using microarrays to understand processes involving the collective action of a number of genes, often organized as a complex or pathway (Gaudet and Weinberg 2010).

To address these needs and remedy the problem of dimensionality, we consider applying the BC-LRT to detect differentially expressed (DE) genes that are typically full of false positives—a direct result of the well-known ‘small-n-large-p’ dimensionality problem: only a few samples available and many genes to consider. At the same time biologists are using microarrays to understand processes involving the collective action of a number of genes, often organized as a complex or pathway (Gaudet and Weinberg 2010). To address these needs and remedy the problem of dimensionality, we consider applying the BC-LRT to detect differentially expressed pre-assigned groups of genes. GO (Ashburner et al. 2000) is a bioinformatics initiative to unify the representation of genes and gene products across the whole biological spectrum. At the leaves of this directed acyclic graph or tree-like representation are the
Table 3. Normal versus cancerous fibroblast cells

<table>
<thead>
<tr>
<th>GO term</th>
<th>m Šidák Global Fisher LRT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cystatin</td>
<td>8 0.0002 0.0285 0.0181 ~0.0001</td>
</tr>
<tr>
<td>Fas</td>
<td>12 0.0740 0.0285 0.0027 0.0001</td>
</tr>
<tr>
<td>Gap junction protein</td>
<td>8 0.0156 0.0285 0.0108 0.0006</td>
</tr>
<tr>
<td>Mitochondrial</td>
<td>48 0.6542 0.1142 0.0047 0.0008</td>
</tr>
<tr>
<td>Keratin</td>
<td>18 0.0342 0.0571 0.0032 0.0011</td>
</tr>
<tr>
<td>Proteasome</td>
<td>32 0.4701 0.2000 0.0017 0.0011</td>
</tr>
<tr>
<td>Fibulin</td>
<td>10 0.1830 0.0857 0.0033 0.0014</td>
</tr>
<tr>
<td>Cyclin</td>
<td>76 0.4047 0.2000 0.2090 0.0030</td>
</tr>
<tr>
<td>Claudin</td>
<td>6 0.0252 0.0285 0.0119 0.0031</td>
</tr>
<tr>
<td>VEGF</td>
<td>12 0.7665 0.4857 0.0274 0.0078</td>
</tr>
<tr>
<td>Cell division</td>
<td>30 0.7432 0.1714 0.0231 0.0154</td>
</tr>
<tr>
<td>Helicase</td>
<td>10 0.1332 0.1142 0.0578 0.0160</td>
</tr>
<tr>
<td>Polymerase</td>
<td>46 0.0642 0.1714 0.2610 0.0198</td>
</tr>
<tr>
<td>Laminin</td>
<td>14 0.2014 0.0571 0.3032 0.0274</td>
</tr>
<tr>
<td>Fas</td>
<td>32 0.0081 0.2000 0.7113 0.0009</td>
</tr>
<tr>
<td>Spectrin</td>
<td>8 0.3134 0.2285 0.1327 0.0325</td>
</tr>
<tr>
<td>Translocase</td>
<td>10 0.3108 0.1428 0.0925 0.0416</td>
</tr>
</tbody>
</table>

Comparison of several group testing procedures: (i) Šidák test, (ii) Fisher’s combined probability method based on parallel univariate tests, (iii) Goeman’s Global test and (iv) our Bartlett corrected likelihood ratio test (BC-LRT) based on simultaneous testing. Group-wise \( p \)-values of normal and cancerous fibroblast genes involving eight observations and based on a GO grouping, where \( n \) is the number of genes in each GO term.

individual genes, whereas higher up in the tree so-called annotation terms unify groups of genes. All the genes that fall under a particular annotation term typically relate to some functional property of these genes and are therefore ideal candidates for a group-wise analysis.

4.3.1 Comparison of cancerous and normal fibroblast cells

We focus on a study by Nighean Barr (Wit and McClure, 2004), which aimed to study differences in expression in cancerous and normal fibroblast cells. The fibroblast tissue used was created in vitro from two cell lines—one cancerous and one normal. From each of these two cell lines, four separate replicates were obtained. Pairs of cancerous and normal replicates were then hybridized to four two-channel cDNA arrays, resulting in eight observations. The presence of a slide effect requires some sort of correction. The simplest possible correction, i.e. pairing the data, would reduce the number of independent samples to four. However, the availability of thousands of gene expressions per slide means that we can use a flexible slide correction model (9216) of gene expressions per slide. This is done in order to increase the number of observations and based on a GO grouping, where \( n \) is the number of genes in each GO term.

Of particular interest is the comparison of cancerous and normal fibroblast cells. The fibroblast tissue used was created in vitro from two cell lines—one cancerous and one normal. From each of these two cell lines, four separate replicates were obtained. Pairs of cancerous and normal replicates were then hybridized to four two-channel cDNA arrays, resulting in eight observations. The presence of a slide effect requires some sort of correction. The simplest possible correction, i.e. pairing the data, would reduce the number of independent samples to four. However, the availability of thousands of gene expressions per slide means that we can use a flexible slide correction model (9216) of gene expressions per slide. This is done in order to increase the number of observations and based on a GO grouping, where \( n \) is the number of genes in each GO term.

For Table 3, we can see that only 3 of the 17 GO terms are declared differentially expressed by all four methods. Fisher’s test is marginal for cyclin and helicase. Only the LRT-BC test confirms DE of proteasome, an intracellular scaffold protein that maintains cell membrane and cytoskeletal integrity, and fibulin, an extracellular matrix (ECM) protein secreted by cells.

Cyclin orchestrates the cell cycle by directing cyclin-dependent kinase activity. Production and subsequent degradation of cyclin by proteolysis are critical for cell cycle control. Cyclin genes with pronounced DE are involved with S, G2, and mitosis phases. Proteasomes are cylindrical protein structures within each cell that contain active sites where proteolysis occurs. Up-regulation of proteasome therefore indicates increased proteolysis so ‘new can be made from old’ and is consistent with down-regulation of inhibitors (e.g. cystatin). Similar trends occur for cyclin, cell division and mitochonidria, that supplies ATP and is involved with many cell-signalling pathways; so the GO terms indicate tumour cell proliferation and growth—a hallmark of cancer (Hanahan and Weinberg, 2011). None of the four GO terms is detected by the Šidák or Global tests at the 5% significance level, and Fisher’s test is marginal for cyclin and helicase. Tumour’s exhibit unstable genomes and this is shown by DE of enzymes. Helicase, for example, drives the unwinding of DNA or RNA helices into separate stands and is crucial for replication, transcription, etc. Polymerases are involved in copying and reading of DNA and RNA and often work co-operatively with helicase motor proteins. Only the LRT-BC test confirms DE of the two GO terms; the other three tests exhibit, at best, marginal DE. Similar results are found for the translocase gene family involved with moving molecules across membranes and chromatin remodelling.

Vascular endothelial growth factor (VEGF) is a signalling protein produced by cells that stimulates angiogenesis needed by cells to access oxygen and nutrients from blood. DE of VEGF occurs in premalignant neoplastic lesions as well developed tumours (Hanahan and Weinberg, 2011) and it is noteworthy this angiogenic factor is detected only by the LRT-BC and Fisher’s test. Of further interest is the expression of insulin-like growth factor (IGF) GO, a complex of proteins that cells use to communicate. Fibroblast dermal cells, for example, communicate survival factor IGF-1 to neighbouring epithelial keratinocytes (Lewis et al., 2006) and the reduction of IGF-2 signal in tumourigenic pathways has been shown.
to reduce tumour growth (Hanahan and Weinberg, 2011). Up-regulation of IGF GO, therefore, confirms fibroblast skin cancer growth and proliferation. Importantly, only the LRT-BC detects IGF GO, the other tests being marginal at best. Fibroblasts are stromal cells that typically produce collagen for ECM connective tissue. Positive DE for keratin in the fibroblast cancer suggests a possible mesenchymal-to-epithelial transition (MET) and EMT since keratin can be expressed by fibroblasts recruited into cancer stroma (Ishii et al., 2008). EMT and MET trans-differentiation therefore suggests invasiveness, another hallmark of cancer (Hanahan and Weinberg, 2011). All tests declare keratin as DE, except Global which is marginal. The cell-surface receptor, Fas, also known as CD95, forms the death-inducing signalling complex (DISC) and the ultra-low \( P \)-value indicates apoptosis (programmed cell death or suicide). The conventional view is that apoptosis reduces cancer by attrition, but paradoxically tumours often express high levels of Fas/CD95. Recent evidence suggests that Fas could in fact have multiple roles apart from DISC, for example, crosstalk may be involved with Fas signalling, and importantly that the skin cancer cells have may become apoptosis resistant (Ben et al., 2011; Ferrozzi et al., 2011). Insensitivity to death signalling is yet another hallmark of cancer. Three tests detect Fas DE, whereas Sidák is marginal.

The above analysis of GO terms for the skin cancer dataset clearly shows the LRT-BC detecting many more GO terms than the other three tests and reveals important cancer hallmarks, e.g. tumour cell proliferation, growth, resistance to cell death, angiogenesis and possibly trans-differentiation and invasiveness. Altogether, they indicate the development of a tumour micro-environment and suggest a follow-up study. LRT-BC has been shown in a previous section to be more sensitive and unbiased. It presents, therefore, a balanced and complete picture of significant GO activity of skin cancer fibroblasts.

4.3.2 Leukaemia In leukaemia gene expression study, Golub et al. (1999) considered 38 bone marrow samples obtained from acute leukaemia patients at the time of diagnosis. Of these 38 samples, 11 were from acute myeloid leukaemia (AML) patients and 27 from acute lymphoblastic leukaemia (ALL) patients. RNA prepared from bone marrow mononuclear cells was hybridized to high-density oligonucleotide microarrays, produced by Affymetrix and containing probes for 6817 human genes. For each gene, the experimenters obtained a quantitative expression level. Samples were subjected to a priori quality control standards regarding the amount of labelled RNA and the quality of the scanned microarray image. After preprocessing (Dudoit et al., 2002), the dataset consisted of 3051 genes and 38 tumour mRNA samples. The aim of this study was to obtain a automatically derived class predictor to determine the class of new leukaemia cases. Although it is possible to create such a predictor with a large number of individual genes, in a second stage one is interested in a biological explanation of the difference between the two types of leukaemia.

In this analysis, we considered 235 GO classes, each one with at least five genes present on the microarray. For each GO class, we calculated \( P \)-values according to our LRT with the most conservative Bartlett correction and Goeman’s Global test. A summary of the results are shown in Figure 4 whereas a detailed list of results for all GO classes is given in the Supplementary Materials. From this two important features become clear: (i) although the Global test and the LRT pick up the same signal, the LRT is, in general, more powerful and (ii) the larger the group size \( m \), the more powerful the LRT test becomes. In fact, the correlation between the difference in log \( P \)-value for the two methods and the group size is 0.55. This feature, i.e. more power when \( m \) increases, is exactly what we call ‘borrowing strength’ in the title of this article.

There are important GO classes that the global test would have missed in this case. For example ATP-dependent helicase activity (GO:0008026) was non-significant at the 5% level for the Global test (not even considering multiple testing), but was highly significant (\( P \)-value < 0.001) for the LRT. This class of eight genes drives the unwinding of a DNA or RNA helix and seems to be important in distinguishing between ALL and AML cases.

5 CONCLUDING REMARKS A property of many modern measurement techniques is the ability to measure simultaneously large numbers of features. Often it is of interest to test for concordant changes in particular subgroups of these features. One can think of groups of genes, so-called pathways, in genomic data or groups of labelled pixels in remote sensing data as part of astronomical images, fMRI brain scans or geographic surveys. Typically measurements are roughly on the same scale, but assuming a common variance is too restrictive. Sequential univariate tests with a ‘multiple testing correction’, such as the Šidák-like correction, are commonly used in such situations. Although they are simple, they are not particularly powerful to detecting small concordant changes in many channels. Multivariate tests, such as Hotelling’s \( T^2 \)-test, are traditionally used to deal with testing movement in multiple dimensions, but are not suited when the number of dimensions (\( m \)) exceeds the number of observations (\( n \)). This is typically the case in modern multichannel data. Other methods considered in this article are Fisher’s method of combining \( P \)-values and Goeman’s Global test. In the case of real applications, both methods work reasonably well, but do not achieve the same power as the likelihood ratio test with the most conservative Bartlett correction.
The advantage of the likelihood ratio test proposed in this article consists ‘borrowing information’, i.e. sharing of variance information across the measurement channels. We have shown in a practical example how the power increases when the number of features goes up. Moreover, its ability to detecting small but concordant changes across a large number of signal sources makes it preferable over more commonly used univariate tests. Ultimately, this sensitive test gives us a richer picture of the underlying biology, as we have shown in the comparison of normal and cancerous fibroblast cells.

6 ACKNOWLEDGEMENTS

D.B. acknowledge the Beatson Laboratories (CR UK) and The Wellcome Trust for financial support (Project 062511).

Funding: David Bakewell was partially funded by a Wellcome Trust Technology grant (Project 062511, “A multi-collaborative microbial pathogen microarray facility”, file number 062511/Z/00/Z).

Conflicts of Interest: none declared.

REFERENCES


