BIOINFORMATICS ORIGINAL PAPER Vol. 28 no. 15 2012, pages 1980–1989
doi:10.1093/bioinformatics/bts316

Structural bioinformatics

Advance Access publication June 4, 2012

Borrowing strength: a likelihood ratio test for related sparse signals

Ernst C. Wit[1,*] and David J. G. Bakewell[2,*]

[1]Johann Bernoulli Institute, University of Groningen, 9747 AG Groningen, The Netherlands and
[2]Electrical Engineering & Electronics, University of Liverpool, Liverpool L69 3GJ, UK

Associate Editor: Alfonso Valencia

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 http://www.math.rug.nl/stat/Software.

Contact: e.c.wit@rug.nl d.bakewell@liv.ac.uk

Received on February 15, 2011; revised on April 18, 2012; accepted on May 23, 2012

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 (Ideker *et al.*, 2000) 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. (Kong *et al.*, 2006) suggested to use Hotelling's T^2 statistic for this purpose, but this is only a viable alternative when the number of observations exceeds the number of channels. This is in many modern genomic applications not the case. Alternatives are ANCOVA approaches (Hummel *et al.*, 2008; Mansmann and Meister, 2005), but compared with the Hotelling T^2 test they lose some power for correlated channels. Our approach is further motivated by recent work that developed a global gene expression estimation method for testing association with clinical outcome (Goeman *et al.*, 2004), although that test either makes asymptotic assumptions for calculating significance, which may be too liberal, or uses permutation, which may not be optimal. The test we propose takes into account the '*magnitude* and *uncertainty*' of the changes.

In the field of microarray analyses, two-stage methods have been proposed that look for over-represented functional classes, e.g. as defined by gene ontology (GO), among differentially expressed genes (Al-Shahrour *et al.*, 2005; Breitling *et al.*, 2004; Martin *et al.*, 2004). Although these methods are very powerful and simple, they tend to look at clearly differentially expressed genes among which patterns are to be found. These methods are exploratory in nature and often require an arbitrary cutoff for the second stage analyses. Our method is more a higher level exploratory method, which is more sensitive to concordant small changes and does not require arbitrary dichotomies.

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. (Wolfinger *et al.*, 2001) 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 (Kerr *et al.*, 2000) and (Rosa *et al.*, 2005), 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

*To whom correspondence should be addressed.

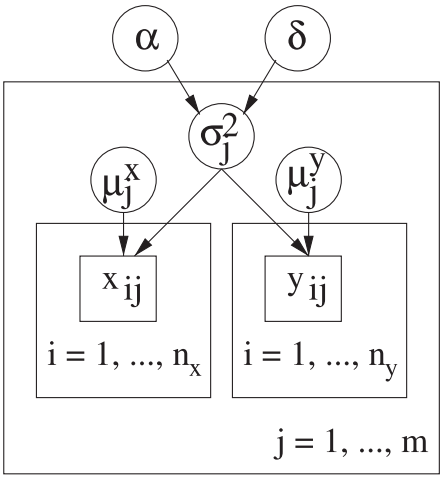

Fig. 1. Directed acyclic graph of the hierarchical model for modelling the difference between two conditions across a set of signals from m similar channels, measured across $n=n_x+n_y$ samples. Hyper-parameters α and δ (top layer) govern the variance σ_j^2 for each gene j. Signals x_{ij} and y_{ij} (bottom layer) are governed by σ_j^2 and mean μ_j^x and μ_j^y, respectively

purpose, we propose a hierarchical model, in which the variance term is modelled explicitly. Initially, we assume that after some suitable transformation, the signals can be considered to be normally distributed with a common variance across the conditions, but potentially varying across the signal sources, $x_{ij}\sim N(\mu_j^x,\sigma_j^2)$ and $y_{ij}\sim N(\mu_j^y,\sigma_j^2)$ where μ_j^x and μ_j^y describe the average signal strength under the two conditions for channel j. Although in some cases the number of subjects per condition are the same, we do not assume this here. Let n_j^x and n_j^y 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 σ_j^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, $\sigma_j^2\sim\Gamma^{-1}(\alpha,\delta)$. 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

$$l(\alpha,\delta,\mu_j^x,\mu_j^y) = m\left(\ln\Gamma(\alpha+\frac{n}{2})-\ln\Gamma(\alpha)-\frac{n}{2}\ln(2\pi\delta)\right) -\left(\alpha+\frac{n}{2}\right)\sum_{j=1}^{m}\ln\left(\frac{\zeta_j}{2}+\delta\right), \quad (1)$$

where for notational simplicity, we introduce the quantities $n=n_x+n_y$ and $\zeta_j=\sum_{i=1}^{n_x}(x_{ij}-\mu_j^x)^2+\sum_{i=1}^{n_y}(y_{ij}-\mu_j^y)^2$.

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, $H_0:\mu_j^x=\mu_j^y,\forall j\le m$.

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 χ^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 χ^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 χ^2 distributed if $n\ge 30$. For larger m the likelihood ratio statistic will become more and more normally distribution, 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

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

where Θ_0 are the hypothesized values under H_0. We consider separately the maximization of the likelihood under H_0 and the full parameter space.

$$l_0 = \sup_{\alpha,\delta,\mu_j^x=\mu_j^y}\{l(\alpha,\delta,\mu_1,..,\mu_m)\} \quad (2)$$

$$l_1 = \sup_{\alpha,\delta,\mu_1,..,\mu_m}\{l(\alpha,\delta,\mu_1,..,\mu_m)\}. \quad (3)$$

Expressions for finding the maximum log-likelihood l_0 and l_1 as a function of α, δ and $\vec{\mu}$ can be found by setting the partial derivatives of Equation (1) to zero, i.e. $\forall_j:\partial_{\mu_j^x}l(\alpha,\delta,\vec{\mu})=0$ and $\forall_j:\partial_{\mu_j^y}l(\alpha,\delta,\vec{\mu})=0$ lead to

$$\hat{\mu}_j^x=\frac{1}{n_x}\sum_{i=1}^{n_x}x_{ij},\quad \hat{\mu}_j^y=\frac{1}{n_y}\sum_{i=1}^{n_y}y_{ij}, \quad (4)$$

for the unconstrained solution l_1, whereas for the constrained solution l_0, we have

$$\hat{\mu}_j^x = \hat{\mu}_j^y = \frac{\sum_{i=1}^{n_x} x_{ij} + \sum_{i=1}^{n_y} y_{ij}}{n}. \tag{5}$$

Furthermore, $\partial_\delta l(\alpha, \delta, \vec{\mu}) = 0$ results in

$$\widehat{\alpha}(\delta) = \frac{n}{2}\left(\frac{m}{\sum_{j=1}^{m}\left(1+\frac{2\delta}{\zeta_j}\right)^{-1}} - 1 \right) \tag{6}$$

and $\partial_\alpha l(\alpha, \delta, \vec{\mu}) = 0$ leads to

$$m\big(\psi(\alpha+\frac{n}{2}) - \psi(\alpha)\big) - \sum_{j=1}^{m} \ln(\frac{\zeta_j}{2\delta}+1) = 0, \tag{7}$$

where the '*digamma*' or '*psi*' function is defined $\psi(x) = d(\ln(\Gamma(x)))/dx$. By substituting Equation (4) respectively Equation (5) in ζ_j, defined below (1), for all j, and by substituting Equation (6) in Equation (1), we obtain expressions for l_0 and l_1 that only depend on δ, i.e.

$$\begin{aligned} l(\delta) = m\left(\ln\Gamma\left(\hat{\alpha}(\delta)+\frac{n}{2}\right) - \ln\Gamma(\hat{\alpha}(\delta)) - \frac{n\ln 2\pi\delta}{2} \right) \\ -\left(\hat{\alpha}(\delta)+\frac{n}{2}\right)\sum_{j=1}^{m}\ln\left(\frac{\hat{\zeta}_j}{2\delta}+1\right), \end{aligned} \tag{8}$$

where $\hat{\zeta}_j$ is defined with respect to the constrained and unconstrained estimates $\hat{\mu}_j^x$ and $\hat{\mu}_j^y$ for l_0 and l_1, respectively. The procedure for finding the suprema 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 $\hat{\delta}_{\mathrm{MLE}}$ as δ_0 by maximizing (8) over some grid $\{0, d_1, \ldots, d_{\max}\}$, where, e.g. $d_{\max} \cong 10^{10}$. Then, we progressively refine the starting point using Newton–Raphson (NR),

$$\delta_{k+1} = \delta_k - \frac{l'(\delta_k)}{l''(\delta_k)}.$$

Computationally less demanding, but leading to the same results, is iteratively solving Equation (7) for δ by fixing $\alpha = \hat{\alpha}(\delta_k)$ using NR and then replacing α 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 δ_0, the NM yielded excellent agreement with NR.

In cases where the maximum likelihood for δ achieves $\hat{\delta}_{\mathrm{MLE}} = \delta_{\max}$, then it is clear from Equation (6) that also $\hat{\alpha}$ tends to be large. Practically, this means that the distribution of σ_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 $\hat{\delta} \to \infty$, Equation (6) simplifies to $\widehat{\alpha} = nm\widehat{\delta}/\sum_{j=1}^{m}\zeta_j$, so that $\forall j \in \{1, \ldots, m\}$ the channel variance estimate degenerates in

$$\begin{aligned} \hat{\sigma}^2 &\approx \frac{\hat{\delta}}{\hat{\alpha}} \\ &= \frac{1}{nm}\sum_{j=1}^{m}\zeta_j \\ &= \frac{\sum_{j=1}^{m}\sum_{i=1}^{n_x}(x_{ij}-\mu_j^x)^2 + \sum_{j=1}^{m}\sum_{i=1}^{n_y}(y_{ij}-\mu_j^y)^2}{m(n_x+n_y)}, \end{aligned}$$

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-likehood 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,

$$\lim_{\delta\to\infty} l(\delta) = \frac{nm}{2}\left(\ln\left(\frac{nm}{2\pi\sum_{j=1}^{m}\zeta_j} \right) - 1 \right). \tag{9}$$

3.2 Bartlett correction

For large samples, $n = n_x + n_y$, the distribution of $\Lambda(x, y)$ is approximately distributed like a χ_q^2 distribution, where $q = \dim\Theta - \dim\Theta_0$ is the difference in the number of free parameters in Θ and Θ_0. We define the Bartlett correction as

$$\begin{aligned} \mathrm{BC} &= E_{H_0}\{\Lambda(x,y)\}/m \\ &= -2E_{H_0}(l_0 - l_1)/m \end{aligned} \tag{10}$$

with l_0 and l_1 defined in Equations (2) and (3). By defining the small sample likelihood ratio statistic as $\Lambda_{\mathrm{BC}}(x, y) = \Lambda(x, y)/BC$, we achieve that precisely the Bartlett corrected likelihood ratio statistic has $E_{H_0}\{\Lambda_{\mathrm{BC}}\} = m$ as would be expected for a χ_m^2 distribution.

Calculation of these two expectations in general is very involved. If we use the characterization for $\lim_{\delta\to\infty} l(\delta)$ defined in Equation (9), we can get an explicit approximate expression for the Bartlett correction,

$$\mathrm{BC} \approx -2E_{H_0}(\lim_{\delta\to\infty} l_0(\delta) - \lim_{\delta\to\infty} l_1(\delta))/m \tag{11}$$

$$= nE\ln\left(1 + \frac{\sum_{j=1}^{m} S_{\Delta,j}}{\sum_{j=1}^{m}(S_{x,j}+S_{y,j})} \right) \tag{12}$$

whereby

$$\zeta_j^0 = \sum_{i=1}^{n_x}(x_{ij}-\widehat{\mu}_j)^2 + \sum_{i=1}^{n_y}(y_{ij}-\widehat{\mu}_j)^2$$

$$\zeta_j^1 = \underbrace{\sum_{i=1}^{n_x}(x_{ij}-\bar{x}_{\cdot j})^2}_{S_{x,j}} + \underbrace{\sum_{i=1}^{n_y}(y_{ij}-\bar{y}_{\cdot j})^2}_{S_{y,j}}$$

$$S_{\Delta,j} = \frac{n_x n_y(\bar{x}_{\cdot j}-\bar{y}_{\cdot j})^2}{n},$$

with $\widehat{\mu}_j = n_x\bar{x}_{\cdot j}+n_y\bar{y}_{\cdot j}/n_x+n_y$ and $\bar{x}_{\cdot j}=1/n_x\sum_{i=1}^{n_x}x_{ij}$. Since for small values of m, the supremum of the likelihood in Equations (2) and (3) are actually found at such degenerate δ, the approximation in Equation (11) can be exact in certain cases. The log expressions are normally distributed, so that under the assumption of no differential expression and in the degenerate case $\sigma_j^2=\sigma^2$, we have that $\bar{x}_{\cdot j}-\bar{y}_{\cdot j}\sim N(0,\sigma^2(n_x+n_y)/n_xn_y)$ and therefore, $\sum_{j=1}^m S_{\Delta,j}/\sigma^2\sim\chi^2_m$ and $\sum_{j=1}^m S_{x,j}+S_{y,j}/\sigma^2\sim\chi^2_{m(n-2)}$. The ratio of χ^2 distributions is F distributed, thus

$$\mathrm{BC}\approx nE\ln\left(1+\frac{F_{m,m(n-2)}}{n-2}\right). \tag{13}$$

The density of $F_{m,m(n-2)}$ (Hogg *et al.*, 2005, p. 185) is

$$f(x)=\frac{(n-2)^{m/2}x^{m/2-1}}{B(m,m(n-2))(1+x/(n-2))^{m(n-1)/2}}, \quad x>0 \tag{14}$$

and $B(x,y)$ is the Beta function. The expected value in Equation (13) can be found by applying a transformation theorem (Gradshteyn and Ryzhik, 2000; Hogg *et al.*, 2005), whereas subsequent approximations use the following expansion of the psi function, $\psi(z)=\ln(z)-1/2z+O(z^{-2})$ (Abramowitz and Stegun, 1965, p. 259),

$$\begin{aligned}\mathrm{BC} &\approx \frac{n}{B(m,m(n-2))}\int_0^\infty \frac{\ln(1+\frac{x}{n-2})v^{m/2}x^{m/2-1}}{(1+\frac{x}{n-2})^{m(n-1)/2}}\,dx\\ &= n\left[\psi\left(\frac{m(n-1)}{2}\right)-\psi\left(\frac{m(n-2)}{2}\right)\right] &(15)\\ &= n\left[\ln\left(\frac{n-1}{n-2}\right)+O(m^{-1}n^{-2})\right] &(16)\\ &\approx \frac{n}{n-2}. &(17)\end{aligned}$$

Given that $\psi(z+1)=\psi(z)+1/z$ (Abramowitz and Stegun, 1965, p. 260), it can be shown that Equation (16) is a lowerbound for Equation (15), which results in an equality if and only if $m=2$. Similarly, Equation (17) is an upperbound for Equation (15). Using these three approximations, we can order the three forms of approximately Bartlett-corrected likelihood ratio statistic:

$$\Lambda_{(17)}(x,y)\le\Lambda_{(15)}(x,y)\le\Lambda_{(16)}(x,y).$$

The corresponding p-values, therefore, have the reverse ordering $P_{(16)}\le P_{(15)}\le P_{(17)}$, making the simple $n/(n-2)$ correction the most conservative of the three test statistics.

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

		Target BC correction		Approximate BC correction		
n	m	$\alpha=4$ $\delta=0.25$	$\alpha=10^4$ $\delta=10^4$	Equation (15)	Equation (16)	Equation (17)
4	5	2.02	1.73	1.76	1.62	2.00
6	5	1.62	1.40	1.40	1.34	1.50
8	5	1.35	1.27	1.27	1.23	1.33
16	5	1.20	1.11	1.12	1.10	1.14
30	2	1.10	1.07	1.07	1.05	1.07
	5	1.10	1.06	1.06	1.05	1.07

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 $\alpha=4$, $\delta=0.25$ and $\alpha=10\,000,\delta=10\,000$, respectively. The approximation BC correction columns are the three derived approximations, where Equation (17) is the simplest and most conservative one.

Table 1 shows the comparison between the three forms of the Bartlett correction. The column with the Monte Carlo approximations should be seen as the (approximately) true values that the final three columns try to approximate. Importantly, the Bartlett correction that we propose is independent of α and δ, given that it is based on the asymptotic approximation (9), in particular, on large values of δ. Therefore, it is not very surprising that the first approximation (15) does an excellent job in capturing the Bartlett correction for $\delta=10\,000$. However, for large values of α and δ, each of the channels is forced to have a similar variance and simpler methods would be available. On the other hand, the approximation (15) is somewhat liberal for small values for α and δ. The somewhat more conservative and very simple upperbound (17) of the Bartlett Correction results in better agreement with Monte Carlo runs in that case. For small m, the latter lies closer to the upperbound and for large m closer to its lowerbound.

Therefore, we conclude that for cases in which there is some channel variance heterogeneity (i.e. small to moderate δ, relative to $\sqrt{\alpha}$), the simple Bartlett-Correction approximation, $n/n-2$, does do an excellent job as first-order correction of the likelihood-ratio statistic.

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 fMRI 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 leukaemia (cf Section 4.3.2) for the BC-LRT without and with 'dependence correction' (DC)

	m	LRT	m^*	LRT-DC
GO:0003697	19	0.0001	13	0.0007
GO:0004725	19	0.0001	14	0.0006
GO:0005096	27	0.0001	16	0.0019
GO:0005244	19	0.0003	12	0.0027
GO:0004842	20	0.0009	14	0.0040
GO:0005089	13	0.0012	9	0.0052
GO:0017017	6	0.0018	5	0.0036
GO:0004879	11	0.0022	8	0.0068
GO:0005544	9	0.0027	7	0.0064
GO:0004693	8	0.0030	6	0.0077
GO:0005164	8	0.0034	6	0.0086
GO:0005201	23	0.0041	15	0.0147
GO:0005001	7	0.0047	6	0.0075
GO:0003746	7	0.0049	6	0.0078
GO:0003735	17	0.0328	11	0.0631
GO:0030955	21	0.0917	14	0.1298
GO:0032395	17	0.1527	8	0.2149

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 α and δ 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 α and δ, 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 χ^2_1 variable under H_0, in fact $m \times \chi^2_1$ distributed, rather than a χ^2_m 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 $n = n_x + n_y < m$, 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

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

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
$$R = \mathrm{Cor}\left(\begin{pmatrix} x \\ y \end{pmatrix}\right).$$
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.
$$m^* = \min_k \left\{ k \mid \frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{m} \lambda_i} > 0.95 \right\}. \quad (18)$$
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 $P - \text{value} = P(\chi^2_{m^*} \geq \Lambda^*_{(17)})$.

Then it is clear that there are really only two independent channels in both cases. This is consistent with fact that in both cases $(\lambda_1 + \lambda_2)/(\lambda_1 + \ldots + \lambda_4) = 0.995 > 0.95$. 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 3.1 converges slower with a small number of channels, since the optimum is more likely obtained for a degenerate channel variance ($\delta, \alpha \to \infty$), 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. We consider two series simulations to test either aspect of the test. We would like to compare the method with possible alternatives.

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

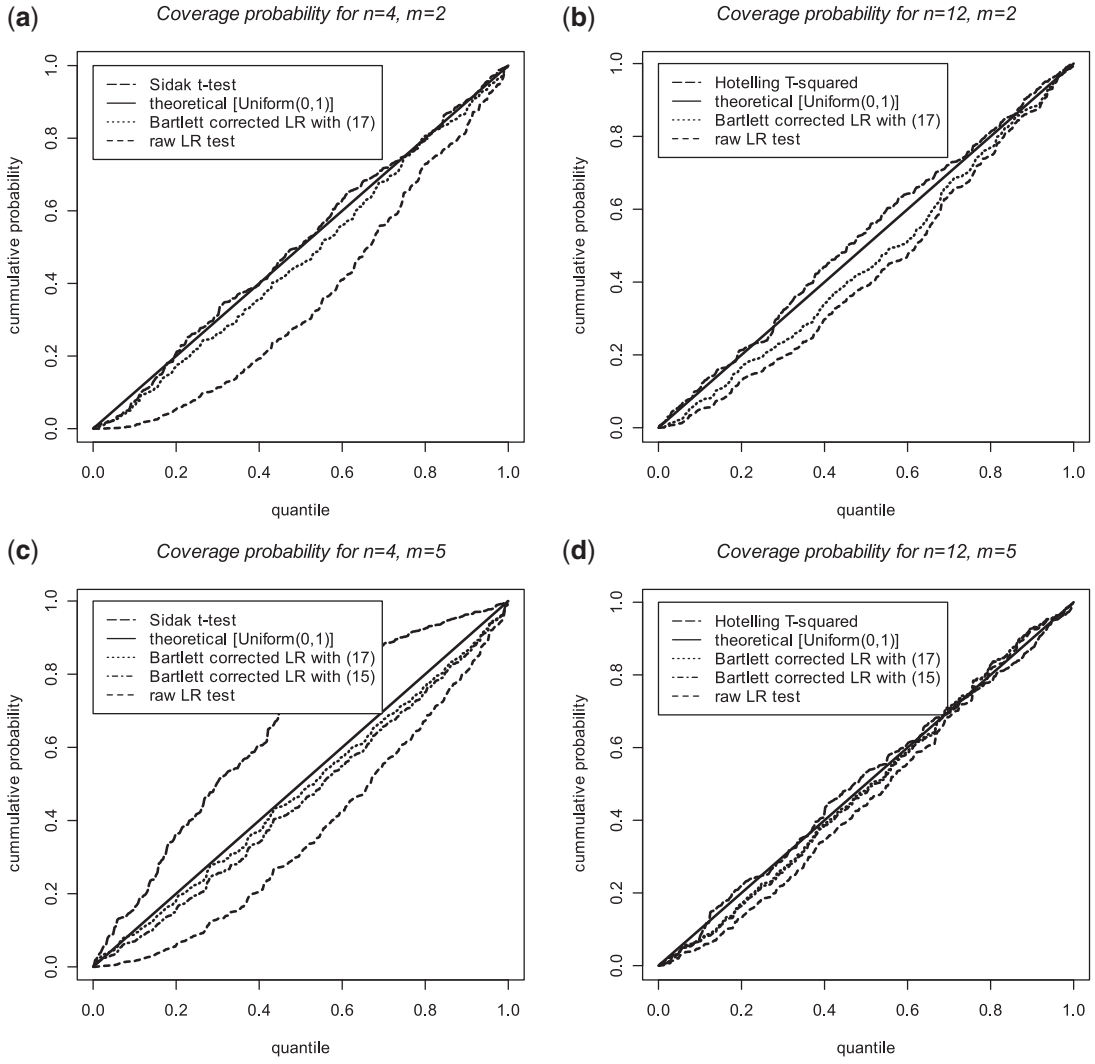

Fig. 2. Coverage probabilities of the BC-LRTs, in comparison with Šidák corrected univariate t-tests ($n=4, m=2$ and $n=4, m=5$) and the Hotelling T^2 alternative ($n=12, m=2$ and $n=12, m=5$) based on 300 simulations

cases, it is possible to test the composite null hypothesis through a series of univariate tests, joined through multiple testing correction, such as a step-down Šidák correction, i.e. $P_{\text{Šidák}} = 1-(1-P_{(1)})^m$, where the $P_{(1)}$ is the smallest observed P-value (Dudoit *et al.*, 2003). We would like to point out that, despite the small sample sizes, the t-tests are the optimal tests to use '*from a univariate point of view*'.

4.1.1 Is the method unbiased? In this section, we show that the method does not systemically 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 δ—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

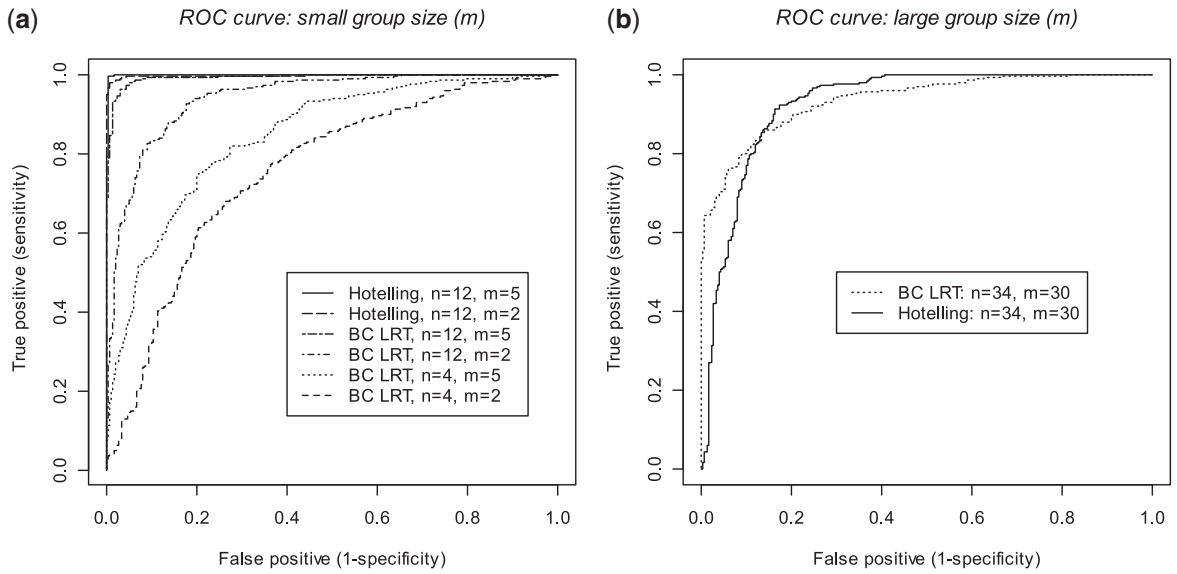

Fig. 3. Receiver operating characteristics of Hotelling T^2 versus Bartlett corrected LRT using factor $n/n-2$ for (**a**) various sample sizes n, various numbers of channels m, with effect size $\mu_x-\mu_y=0.5$ and variance heterogeneity given by $\alpha=3,\delta=1/3$; (**b**) for a larger number of channels $m=30$, with effect size $\mu_x-\mu_y=0.1$ and a larger level of variance homogeneity ($\alpha=12,\delta=4/3$). Both simulations are based on 600 hypotheses: 300 had effect size 0 and 300 had effect size as indicated above. For clarity, the legends are in the order of the lines in the graph

channels show activity in the presence of channel heterogeneity ($\alpha=3$ and $\delta=1/3$). We perform a total of 600 simulations, whereby half of the null hypotheses are true and the other half false with effect size $\mu_x-\mu_y=0.5$. We plot the receiver operating curve for the simulation in Figure 3a. From this we can clearly see that the BC-LRT has significant power. However, in cases where the Hotelling T^2 test can be used, i.e. if the number of observations are somewhat larger than than the number of channels, to wit $n>m+2$, then this test achieves a higher power. This is due to the fact that in these simulation there are so few channels m: it is impossible to 'borrow' much 'strength' across the channels present.

A second simulation shows the dependence of the power of the likelihood ratio test on the variance heterogeneity and the numbers of channels present. We look at $m=30$ channels with the same average variance as above, but with two levels of variance heterogeneity, i.e. $\alpha=3,\delta=1/3$ and $\alpha=12,\delta=4/3$. From Figure 3b, we can see how in these cases the likelihood ratio test beats the Hotelling T^2 test. Moreover, the more variance homogeneity, the larger the power of the test. The power of Hotelling T^2 test does not depend on the variance heterogeneity and only a single line has been plotted. As before, the simulation is based on 600 iterations, whereby 300 under H_0 and 300 under H_1 (here, effect size is taken $\mu_x-\mu_y=0.1$).

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 (Manly, 1985) 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 χ^2 distributed with $2m$ 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 T^2 approach 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 (Goeman *et al.*, 2004).

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 (Hanahan and Weinberg, 2011). 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

GO term	m	Šidák	Global	Fisher	LRT
Cystatin	8	0.0002	0.0285	0.0181	<0.0001
Fas	12	0.0740	0.0285	0.0027	0.0001
Gap junction protein	8	0.0156	0.0285	0.0108	0.0006
Mitochondrial	48	0.6542	0.1142	0.0047	0.0008
Keratin	18	0.0342	0.0571	0.0032	0.0011
Proteasome	32	0.4701	0.2000	0.0017	0.0011
Fibulin	10	0.1830	0.0857	0.0053	0.0014
Cyclin	76	0.4047	0.2000	0.2090	0.0030
Claudin	6	0.0252	0.0285	0.0119	0.0031
VEGF	12	0.7665	0.4857	0.0274	0.0078
Cell division	30	0.7432	0.1714	0.0231	0.0154
Helicase	10	0.1532	0.1142	0.0578	0.0160
Polymerase	46	0.0642	0.1714	0.2610	0.0198
Laminin	14	0.2014	0.0571	0.3032	0.0274
IGF	32	0.0881	0.2000	0.7113	0.0399
Spectrin	8	0.3134	0.2285	0.1327	0.0325
Translocase	10	0.3108	0.1428	0.0925	0.0416

Comparison of several group testing procedures: (i) Šidák test, (ii) Fisher combined probability method based on parallel univariate tests, (iii) Goeman's Global test and (iii) 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 m 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 (9216) of gene expressions per slide means that we can use a flexible slide correction model ($df \approx 20$) and still obtain an effective number of observations $8-4\times 20/4\times 9216=7.998$ close to 8. This correction has been applied before analyzing the data. We focus on 94 groups of genes identified by various GO terms. The group or channel size, m, varies from 4 to 178. It is clear that in most of these cases it is impossible to apply a Hotelling T^2 test as m exceeds n. We report the top 17 of the 94 GO terms in Table 3.

From Table 3, we can see that only 3 of the 17 GO terms were declared differentially expressed by all four methods, i.e. the Šidák adjusted t-tests, Global test (Goeman *et al.*, 2004), Fisher's combined probability test (Manly, 1985) and our likelihood ratio test with the most conservative Bartlett correction. These GO terms are cystatin, gap junction protein (GJP) and claudin. Below, we refer to GO terms as negative or positive DE, i.e. expression levels for the cancerous condition are, respectively, mainly less or more than for normal. Only three GO terms are negative DE: cystatin, GJP and laminin. Cystatin (also known as stefin) genes are responsible for inhibiting cysteine proteases, i.e. enzymes that degrade proteins by hydrolysis and are typically down regulated in skin cancer (Keppler, 2006), concurring with our finding. GJP is a family of membrane proteins (connexin genes) that enables trafficking, through channels, of small ions and metabolites between neighbouring cells. Important for cell–cell communication and coordination, GJPs maintain tissue homeostasis. Cancer is a genetic disease involving aberrant cell behaviour so that DE of GJP, that results in a loss of coordination and homeostasis, is consistent with disruption by tumour activity. Human skin consists of a thin epidermal layer attached by a basement membrane (BM) to a dermal layer of fibroblast-dominated connective tissue (Sorrell and Caplan, 2004). Laminin is in important constituent of BM and the direction of DE is tissue- and tumour-specific, e.g. a gene for laminin 332 is positive and negative DE for squamous and basal cell carcinomas, respectively (Marinkovich, 2007). Claudin, a family of cell-membrane tight junction proteins, is also tissue-specific (Singh *et al.*, 2010). Laminin GO is detected by our LRT-BC at the 5% level whereas at best it is marginal by Fisher's, Global or Šidák tests. A roughly similar pattern of detection occurs for spectrin, 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 mitochondria, 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. Tumoru'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 keratinoctes (Lewis *et al.*, 2009) and the reduction of IGF-2 signal in tumoarigenic 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.*, 2005). 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 (Chen *et al.*, 2010; Green, 2010; Hanahan and Weinberg, 2011; Peter *et al.*, 2007). Insensitivity to death signalling is yet another hallmark of cancer. Three tests detect Fas DE, whereas Šidák 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 leukemia (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

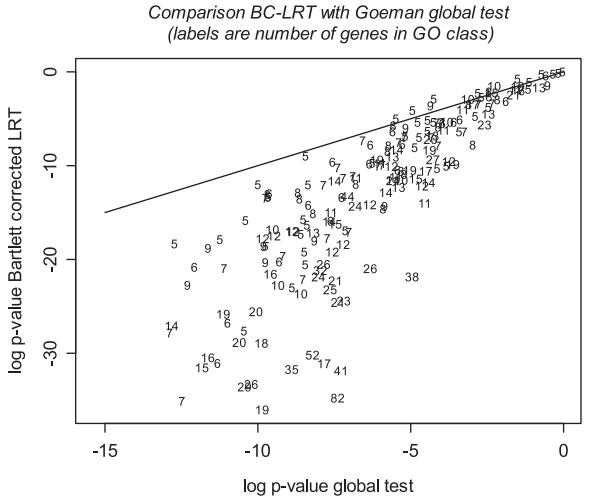

Fig. 4. Comparison of the log-transformed *P*-values, calculated on 235 GO classes, of our BC-LRT versus Goeman's Global test. The numbers in the graph represents the number of genes in each GO class

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).

Conflict of Interest: none declared.

REFERENCES

Abramowitz,M. and Stegun,I.A. (1965) *Handbook of Mathematical Functions*. Dover Publications

Al-Shahrour,F. *et al.* (2005) Discovering molecular functions significantly related to phenotypes by combining gene expression data and biological information. *Bioinformatics*, **21**, 2988–2993.

Ashburner,M. *et al.* (2000) Gene ontology: tool for the unification of biology. *Nat. genet.*, **25**, 25–29.

Breitling,R. *et al.* (2004) Iterative group analysis (iGA): a simple tool to enhance sensivity and faciltate interpretation of microarray experiments. *BMC Bioinformatics*, **5**.

Chen,L. *et al.* (2010) Cd95 promotes tumour growth. *Nature*, **465**, 492–496.

Dudoit,S. *et al.* (2002) Comparison of discrimination methods for the classification of tumors using gene expression data. *J. Am. Stat. Assoc.*, **97**, 77–87.

Dudoit,S. *et al.* (2003) Multiple hypothesis testing in microarray experiments. *Stat. Sci.*, **18**, 71–103.

Goeman,J.J. *et al.* (2004) A global test for groups of genes: testing association with clinical outcome. *Bioinformatics*, **20**, 93–99.

Golub,T. *et al.* (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531.

Gradshteyn,I.S. and Ryzhik,I.M. (2000) *Table of Integrals, Series, and Products*. Academic Press, San Diego, USA.

Green,D. (2010) Cancer: a wolf in wolf's clothing. *Nature*, **465**, 433–433.

Hanahan,D. and Weinberg,R. (2011) Hallmarks of cancer: the next generation. *Cell*, **144**, 646–674.

Hogg,R.V. *et al.* (2005) *Introduction to Mathematical Statistics*. Pearson Prentice Hall, Upper Saddle River, New Jersey, USA.

Hummel,M. *et al.* (2008) Globalancova: exploration and assessment of gene group effects. *Bioinformatics*, **24**, 78.

Ideker,T. *et al.* (2000) Testing for differentially-expressed genes by maximum-likelihood analysis of microarray data. *J. Computa. Biol.*, **7**, 805–817.

Ishii,G. *et al.* (2005) *In vivo* and *in vitro* characterization of human fibroblasts recruited selectively into human cancer stroma. *Int. J. Cancer*, **117**, 212–220.

Keppler,D. (2006) Towards novel anti-cancer strategies based on cystatin function. *Cancer Lett.*, **235**, 159–176.

Kerr,M. K. *et al.* (2000) Analysis of variance for gene expression microarray data. *J. Computa. Biol.*, **7**, 819–837.

Kong,S. *et al.* (2006) A multivariate approach for integrating genome-wide expression data and biological knowledge. *Bioinformatics*, **22**, 2373.

Lewis,D. *et al.* (2009) The igf-1/igf-1r signaling axis in the skin: a new role for the dermis in aging-associated skin cancer. *Oncogene*, **29**, 1475–1485.

Manly,B. (1985) *The Statistics of Natural Selection on Animal Populations*. Chapman and Hall, New York, USA.

Mansmann,U. and Meister,R. (2005) Goeman's global test versus an ancova approach. *Methods Inf. Med.*, **44**, 449–453.

Marinkovich,M. (2007) Laminin 332 in squamous-cell carcinoma. *Nat. Rev. Cancer*, **7**, 370–380.

Martin,D. *et al.* (2004) Gotoolbox: functional analysis of gene datasets based on gene ontology. *Genome Biol.*, **5**, R101.

Peter,M. *et al.* (2007) The cd95 receptor: apoptosis revisited. *Cell*, **129**, 447–450.

Press,W. H. *et al.* (1992) *Numerical Recipes in FORTRAN*. Cambridge University Press, Cambridge, UK.

Prokhorov,A. (2001) *Encyclopaedia of Mathematics*, Hotelling T2-distribution. Springer.

Rosa,G. *et al.* (2005) Reassessing design and analysis of two-colour microarray experiments using mixed effects models. *Comparat. Funct. Genom.*, **6**, 123–131.

Singh,A. *et al.* (2010) Claudin family of proteins and cancer: an overview. *J. Oncol.*, **2010**, 11.

Sorrell,J. and Caplan,A. (2004) Fibroblast heterogeneity: more than skin deep. *J. Cell Sci.*, **117**, 667.

Wit,E. and McClure,J. (2004) *Statistics for Microarrays: Design, Analysis and Inference*. John Wiley and Sons, Chichester, UK.

Wolfinger,R. *et al.* (2001) Asessing gene significance from cdna microarray expression data via mixed models. *J. Compuat. Biol.*, **8**, 625–637.