Fusing microarray datasets using multivariate regression, with application to the cell-cycle in yeast

Walter R. Gilks*[1], Brian D. M. Tom[1] and Alvis Brazma[2]

[1] MRC Biostatistics Unit, Cambridge
[2] EMBL-EBI, Cambridge

1 Introduction

Microarrays have been widely used to compare gene expression in biological samples, but the resulting data suffer from many sources of variation. Therefore, to draw robust conclusions from microarray experiments, results from related experiments should be combined (or *fused*), where possible.

We describe the use of multivariate linear regression to fuse microarray datasets, taking due account differences in data quality between arrays. The aim of the analysis is to deliver a fused data set, which might then be subjected to further analyses by other formal or informal techniques. We apply our methodology to data from Rustici *et al.* (2004) on cell-cycle control in the fission yeast, *S. pombe*.

Our methods and application are described in more detail in Gilks, Tom and Brazma (2005).

2 Methods

Let D denote an $N \times m$ observed data matrix from N microarray hybridisations, each containing the same set of m gene probes. Typically, $m \gg N$. These N hybridisations may be from different laboratories and performed under different experimental conditions. However, we assume that they are related by a set of $n < N$ underlying biological conditions, for example n different cell types. In particular, we assume that the observed data are noisy observations of an unobserved, idealised, $n \times m$ data matrix C, which we aim to estimate from the data through the multivariate regression model:

$$\underset{N\times m}{D} \;=\; \underset{N\times n}{X}\;\underset{n\times m}{C} \;+\; \underset{N\times m}{\varepsilon}\;. \tag{2.1}$$

Here, X is an $N \times n$ design matrix, the construction of which will depend on the context. The matrix of regression parameters, C, must be estimated. Its estimate, $\hat{C}$, is the primary object of our analysis and will be interpreted as a fused data matrix. The $N \times m$ matrix ε represents unobserved noise.

We assume that residual error term ε has zero mean and its rows are uncorrelated. That is, we assume residual errors are uncorrelated *between* hybridisations. However, we can expect residual errors to be correlated *within* a hybridisation, due to complex unmodelled networks of gene interactions. Importantly, we also allow some hybridisations to be more noisy than others. Specifically, we assume for each hybridisation $h = 1, \ldots, N$:

$$\underset{m\times m}{\mathrm{Var}\,[\varepsilon_h]} \;=\; \omega_h \;\underset{m\times m}{\Sigma}\;, \tag{2.2}$$

where vector $\varepsilon_h^{\mathrm{T}}$ is the hth row of ε, where $^{\mathrm{T}}$ denotes matrix transposition; Σ represents an unknown $m \times m$ matrix reflecting a common within-experiment variance-covariance error structure; and ω_h is a scalar reflecting the relative noisiness of hybridisation h, such that $\sum_h \omega_h / N = 1$.

2.1 Estimation

We estimate fused data matrix C from D and X by Generalized Least Squares:

$$\hat{C} \;=\; \left(X^{\mathrm{T}}\Omega^{-1}X\right)^{-1} X^{\mathrm{T}}\Omega^{-1}D, \tag{2.3}$$

where Ω is a diagonal $N \times N$ matrix whose hth diagonal element is ω_h (see, for example, Mardia, Kent and Bibby (1979)). For the moment we assume the $\{\omega_h\}$ are known. We require that the $n \times n$ matrix $\left(X^{\mathrm{T}}\Omega^{-1}X\right)$ should have full rank n. Let $\hat{\varepsilon}$ denote the $N \times m$ matrix of estimated residuals

$$\hat{\varepsilon} = D - X\hat{C}. \tag{2.4}$$

Suppose now the variance modifiers ω_h are unknown. We propose an iterative approach to their estimation. Initially, each ω_h is set to unity. After each iteration i, ω_h is updated:

$$\omega_h^{(i+1)} \;=\; \frac{1}{s+1}\left(s + \frac{N\hat{\varepsilon}_h^{(i)^{\mathrm{T}}}\hat{\varepsilon}_h^{(i)}}{\sum_{k=1}^{N}\hat{\varepsilon}_k^{(i)^{\mathrm{T}}}\hat{\varepsilon}_k^{(i)}}\right)\;, \tag{2.5}$$

where $^{(i)}$ denotes quantities calculated at iteration i; $\hat{\varepsilon}_h^{(i)^{\mathrm{T}}}$ is the hth row of $\hat{\varepsilon}^{(i)}$; and s is a non-negative constant. The resulting $\Omega^{(i+1)}$ is substituted into (2.3) to calculate $\hat{C}^{(i+1)}$, which in turn is substituted into (2.4) to produce $\hat{\varepsilon}^{(i+1)}$. The whole process is iterated until convergence.

3 Application

We illustrate our multivariate regression approach using microarray time-course data from Rustici *et al.* (2004) on cell-cycle control in the fission yeast, *S. pombe*. These experiments employed a variety of techniques to synchronise cells to a common point in the cell cycle. Our primary aim is fuse these experiments to produce a single time-course dataset. However, the potential for each synchronisation method to synchronise to a different point in the cell cycle, and to differentially affect cell-cycle phase lengths, means that the data fusion is not straightforward. This motivates the use of our multivariate regression approach.

The normalised data from 9 experiments performed by Rustici *et al.* (2004) were downloaded from http://www.sanger.ac.uk/PostGenomics/S_pombe/. These experiments employed a variety of synchronisation techniques, and represent a total of 178 sample hybridisations. Our matrix D comprises the normalised data from these $N = 178$ hybridisations for the $m = 407$ genes identified by Rustici *et al.* (2004) as cell-cycle regulated.

Our analysis aims to fuse data at each of a number of points in the cell-cycle. However, we cannot simply assume that the kth hybridisation in each experiment relates to cells that have progressed identically through the cell cycle, even though a common sampling interval of 15 minutes was used throughout. We therefore adopted the following approach.

Each experiment was designed to collect data at 10 equally spaced time-points of the cell cycle. Accordingly, we aim to produce a fused 10×407 data matrix C, representing a canonical time-course experiment at 10 equally spaced times. We call these times *fusion times*. These

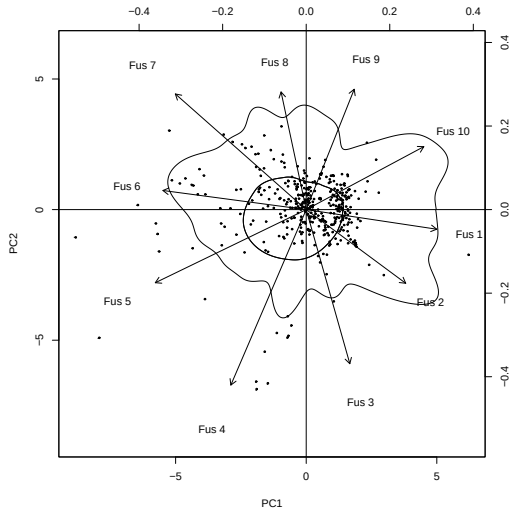

Figure 1: Peppered fried egg plot of $\hat{C}$.

fusion times would appear as equally spaced directions in the plane of the first two principle components of the data (the *principle plane*).

Our design matrix X will contain one row for each of the $N = 178$ hybridisations and one column for each of the $n = 10$ fusion times. We define each hth row of X to be a vector of weights reflecting the relevance of hybridisation h to each fusion time. Each hybridisation falls between two adjacent fusions times in the principle plane, and thus provides approximate information about gene-expression at these two times. The hth row of X is therefore constructed to assign non-zero weights only to the two adjacent fusion times for hybridisation h. Full details are given in Gilks, Tom and Brazma (2005).

3.1 Results

We obtained convergence of (2.3–2.5) within 10 iterations.

Fused data matrix $\hat{C}$ should not be thought of as an end in itself, but a starting point for other formal or informal analyses, depending on the investigators aims. We illustrate this in Figure 1 with a biplot of $\hat{C}$ (Gabriel, 1971). The biplot of a matrix plots the eigenvectors from the SVD of that matrix. It reveals relationships between the rows and columns of the matrix. In the present context, it helps to show how gene expression levels change across the cell cycle. Figure 1 also contains additional features.

We call Figure 1 a *peppered fried egg* plot. The arrows labelled `Fus 1` through `Fus 10` represent fusion times. The coordinates of the tip of each kth arrow are the kth elements of the first two left eigenvectors of $\hat{C}$, and can be read off from the upper and right axes. The length of the kth arrow indicates the variability in expression across the genes at fusion time k. Fusion times 4 and 7 appear most extreme in this regard. However, Figure 1 captures only 83% of the total variance in $\hat{C}$, the other 17% residing in the 8 dimensions orthogonal to the plane of the figure.

The dots (specks of pepper) in Figure 1 represent genes. The coordinates of each ith dot are the ith elements of the first two right eigenvectors of $\hat{C}$, and can be read off from the lower and left axes. The length of the radius to the ith dot indicates the degree of fluctuation in the expression of gene i across the cell-cycle.

The boundary of the 'egg yolk' describes the average radius of dots at each point of the cell

cycle (from a `loess` curve fitted through the radii of the dots). Thus it shows greater cell-cycle fluctuation in genes expressed around fusion times 4 and 5 than in those expressed around fusion times 9 and 10. The boundary of the 'egg white' describes the average density of dots at each point of the cell cycle (from a `loess` curve fitted through a histogram of the dots). Thus it shows that many of these genes are associated with fusion times 6 and 10 through 2.

As noted above, the first two SVD dimensions of $\hat{C}$ contain 83% of its variance, overall. Some genes lie in or close to the plane of Figure 1: these genes have simple sinusoidal cell-cycle dependence, being maximally upregulated at one point of the cell cycle, and maximally downregulated at the opposite point. However, genes which are switched on and off more than once in a cell-cycle will not lie close to this plane, and will tend to project onto the plane near the centre of Figure 1.

4 Discussion

We have shown how multivariate linear regression can be used to fuse microarray datasets. The regression should be seen as a half-way house, enabling further techniques to more easily and transparently focus on matters of interest.

An important aspect of our approach is that it explicitly and automatically takes account of differentials in experimental quality. Thus experiments, or hybridisations, which fail to cohere with the generality of results will be downweighted, and contribute relatively little to the results produced.

References

Gabriel, K. (1971). The biplot graphic display of matrices with application to principal components analysis. *Biometrika*, *58*, 453-467.

Gilks, W., Tom, B. and Brazma, A. (2005). Fusing microarray experiments with multivariate regression. *Bioinformatics*. To appear.

Mardia, K., Kent, J. and Bibby, J. (1979). *Multivariate Analysis*. Academic Press.

Rustici, G., Mata, J., Kivinen, K., Lio, P., Penkett, C., Burns, G., Hayles, J., Brazma, A., Nurse, P. and Bahler, J. (2004). Periodic gene expression program of the fission yeast cell cycle. *Nat. Genet.*, **36**, 809-817.