Biclustering Models for Structured Microarray Data

Heather L. Turner, Trevor C. Bailey, Wojtek J. Krzanowski, and Cheryl A. Hemingway

Abstract—Microarrays have become a standard tool for investigating gene function and more complex microarray experiments are increasingly being conducted. For example, an experiment may involve samples from several groups or may investigate changes in gene expression over time for several subjects, leading to large three-way data sets. In response to this increase in data complexity, we propose some extensions to the plaid model, a biclustering method developed for the analysis of gene expression data. This model-based method lends itself to the incorporation of any additional structure such as external grouping or repeated measures. We describe how the extended models may be fitted and illustrate their use on real data.

Index Terms—Biclustering, two-way clustering, overlapping clustering, partial supervision, repeated measures, three-way data.

1 Introduction

There has been considerable recent interest in the analysis of microarray data. A typical microarray experiment will investigate thousands of genes, recording their expression level over tens of samples. Genes with similar expression patterns over the samples are said to be coexpressed, which may indicate a common function. Likewise, samples with similar expression profiles may have attributes in common, for example they may be samples from patients with the same disease. With the aim of identifying such groups and samples, clustering has a natural role in the exploratory analysis of microarray data.

A number of scenarios that occur in microarray experiments are not catered for by all clustering techniques and this should be taken into account when selecting a method for analysis. First, a gene may be involved in more than one biological process and may exhibit an expression profile that is a result of the regulatory effect of each process. If there are other genes that are involved in some subset of these processes, the structure should be represented by overlapping clusters. Second, a group of genes may be coexpressed under limited conditions. In this case, the structure should be represented by a two-way cluster or *bicluster*, a group of genes and an associated group of samples over which the genes are coexpressed. Finally, genes may not be related to the subject of the investigation and exhibit near-constant expression profiles. Rather than filter out these "uninteresting" genes on the basis of some ad hoc criteria, gene selection should be an integral part of the clustering process, so that genes that do not exhibit interesting patterns are left unclustered. All these scenarios have their equivalent in terms of samples, for example, a cluster of samples may only be distinguished by a cluster of genes.

Several clustering methods have been developed in recent years that cater to one or more of these scenarios. These include gene-shaving [8], context-specific Bayesian clustering [2], EMMIX-GENE [14], interrelated two-way clustering [25], simultaneous clustering [17], coupled two-way clustering [7], rich probabilistic models [22], double conjugated clustering [4], SAMBA [24], order preserving submatrix clustering [3], biclustering [6], [23], and the plaid model [12]. The plaid model is one method that accommodates all the scenarios described earlier and is particularly attractive as it uses continuous gene expression levels and estimates the "usual" expression level for each gene (in the context of the data set), so that biclusters of an unusual expression pattern can be discovered. Furthermore, as a model-based clustering method, the plaid model can be naturally extended to appropriately analyze structured microarray experiments which are the focus of interest in this paper.

First, we consider microarray experiments for which an a priori group structure is available for the genes or samples. In this case, we partially supervise the plaid model algorithm to favor biclusters that correspond to the external grouping, so that biclusters can be interpreted as features relating to one or more a priori groups. We compare the results of a partially supervised analysis to the results of an unsupervised analysis for an experiment investigating forms of tuberculosis.

Second, we consider microarray experiments in which the expression levels of a set of genes are measured over time for several samples. For this type of data, we extend the plaid model so that instead of clustering single expression levels, whole time series of expression levels are clustered. This allows complete three-way microarray data sets to be analyzed, obviating the need for collapsing such data sets to a two-way data structure, which in some cases can lead to a substantial loss of information. We

- *H. Turner, T.C. Bailey, and W.J. Krzanowski are with the Department of Mathematical Sciences, University of Exeter, Laver Building, North Park Rd., Exeter, Devon, EX4 4QE, UK.*
  E-mail: {heather.l.turner, t.c.bailey, w.j.krzanowski}@exeter.ac.uk.
- *C.A. Hemingway is with the Department of Academic Paediatrics, Imperial College Medical School, Norfolk Place, London, W2 1PG, UK.*
  E-mail: c.hemingway@imperial.ac.uk.

Manuscript received 8 Nov. 2004; revised 28 Apr. 2005; accepted 21 June 2005; published online 1 Nov. 2005.
For information on obtaining reprints of this article, please send e-mail to: tcbb@computer.org, and reference IEEECS Log Number TCBB-0186-1104.

illustrate the extension for repeated measures on a set of genes over a set of samples with data from an experiment investigating genetic susceptibility to tuberculosis.

The use of partial supervision and the extension to three-way data relate to other methods for microarray analysis that use these techniques, in particular, algorithms that identify interesting biclusters on the basis of predefined gene groups [16], [10], [21] and gene clustering of three-way data [18]. We compare our approach to these methods in the discussion (Section 5).

As a basis for our development of plaid model clustering, we use the algorithm introduced by Turner et al. [26], which was shown to have advantages over the original algorithm proposed by Lazzeroni and Owen [12]. In the next section, we review the plaid model and the algorithm introduced by Turner et al. [26]. In doing so, we propose some additional variations of the algorithm which can enhance the interpretability of results.

2 The Plaid Model

The plaid model consists of a series of additive *layers* intended to capture the underlying structure in a set of gene expression data that can be represented in the form of a matrix, with expression levels Y_{ij} for the ith gene in the jth sample, $i = 1, \dots, n; j = 1, \dots, p$. The model includes a *background layer* containing all the genes and samples, to account for global effects in the data. Any subsequent layers represent additional effects corresponding to biclusters of the genes and samples that exhibit a strong pattern not explained by the background layer.

In the plaid model Y_{ij} is modeled by

$$\begin{aligned} Y_{ij} &= \Theta_{ij0} + \sum_{k=1}^{K} \rho_{ik}\kappa_{jk}\Theta_{ijk} + \epsilon_{ij} \\ &= (\mu_0 + \alpha_{i0} + \beta_{j0}) + \sum_{k=1}^{K} (\mu_k + \alpha_{ik} + \beta_{jk})\rho_{ik}\kappa_{jk} + \epsilon_{ij}, \end{aligned}$$

where k is a layer index starting at zero for the background layer running to K, the number of biclusters; Θ_{ijk} is the model for layer k; ρ_{ik} is a binary cluster membership parameter defined for $k \geq 1$ and equal to one if the ith gene is in the kth bicluster, zero otherwise; κ_{jk} similarly indicates cluster membership for the jth sample, and ϵ_{ij} is the residual error. Here, the layer model Θ_{ijk} is defined as the sum of the mean effect μ_k, the gene effects α_{ik}, and the sample effects β_{jk}. Thus, the full model is similar to the model used in two-way analysis of variance, except that the two-way interaction between genes and samples is replaced by cluster effects, cluster by gene effects and cluster by sample effects. In this way, the plaid model seeks to decompose the gene by sample interaction effect into additive layers that are more useful for interpretation.

As set out in the introduction, we prefer to use the algorithm proposed by Turner et al. [26] which uses binary least squares to fit the cluster membership parameters, unlike the original algorithm [12] which indirectly optimizes these parameters by relaxing the binary constraints at certain stages of the fitting process. Turner et al. [26] demonstrated that using the binary least squares method reduces the level of false structure incorporated in the plaid model biclusters. Their algorithm, which we shall call Algorithm 1, fits the background layer first of all, then searches for one bicluster at a time as outlined in Fig. 1. Bicluster-specific layers are added to the plaid model until a prespecified number is reached or no more significant layers can be found, as determined by a permutation test. In considering a single layer, we shall drop the layer index k for simplicity, as in Fig. 1.

To initialize the algorithm, starting values for the cluster membership parameters are derived from one-way k-means clusters. A k-means algorithm with k = 2 is used to cluster the genes and the samples independently, then the cluster with fewer members from each result is taken to form the starting bicluster.

The algorithm finds a bicluster of genes and samples for which the layer model fits better than the null model. Any genes and samples which do not fit the layer sufficiently well are usually pruned out of the bicluster by adjusting the cluster membership parameters as follows:

$$\tilde{\rho}_i = \begin{cases} 1 & \text{if } \hat{\rho}_i = 1 \text{ and } \sum_{j:\hat{\kappa}_j=1}(\hat{Z}_{ij} - \hat{\kappa}_j(\hat{\mu} + \hat{\alpha}_i + \hat{\beta}_j))^2 < \\ & \qquad (1 - \tau_1)\sum_{j:\hat{\kappa}_j=1} \hat{Z}_{ij}^2, \\ 0 & \text{otherwise,} \end{cases}$$

$$\tilde{\kappa}_j = \begin{cases} 1 & \text{if } \hat{\kappa}_j = 1 \text{ and } \sum_{i:\hat{\rho}_i=1}(\hat{Z}_{ij} - \hat{\rho}_i(\hat{\mu} + \hat{\alpha}_i + \hat{\beta}_j))^2 < \\ & \qquad (1 - \tau_2)\sum_{i:\hat{\rho}_i=1} \hat{Z}_{ij}^2, \\ 0 & \text{otherwise,} \end{cases}$$

where $\tau_1, \tau_2 \in (0, 1)$ specify for genes and samples, respectively, the minimum proportional reduction in residual sum of squares required for cluster membership. Thus, τ_1 and τ_2 may be viewed as the minimum desired R^2 for the genes and samples.

The genes are pruned first and then the samples. If all the genes or all the samples are pruned out, the algorithm is terminated. The values of τ_1 and τ_2 are usually chosen in the range [0.5, 0.7] to ensure the layer model is an important component of the corresponding expression levels, but also to allow for overlapping biclusters and random error.

Turner et al. only prune the genes and samples once (Step 8, Fig. 1). However, the pruning criteria will only be met for genes and samples simultaneously if this pruning step is repeated until a stable bicluster is obtained. Since we prefer smaller, tighter clusters and the additional computation required is light, we consider it worthwhile to prune until convergence.

After a bicluster has been added to the model, the layer effects for all layers in the current model are usually reestimated in the light of this additional structure. This is particularly important for the background layer as this represents the base expression level for all genes and samples and a good estimate will make it easier to identify a further bicluster, if it exists, in the next round of iterations. The back fitting is carried out sequentially, fitting each layer to the residuals from the model excluding that layer. The complete process may be repeated to improve the fit of the model, but the number of rounds of back fitting is usually kept low to achieve a sensible trade-off between the accuracy of layer effects and computational burden.

```
1. Compute $\hat{Z}_{ij}$: matrix of residuals from model so far
2. Compute starting values $\hat{\rho}_i^0$ and $\hat{\kappa}_j^0$ using one-way k-means clusters
3. Set $s = 1$
4. Update layer effects using $Z^*$: submatrix of $\hat{Z}_{ij}$ indicated by $\hat{\rho}_i^{s-1}$ and $\hat{\kappa}_j^{s-1}$
       $\hat{\mu}^s = \bar{Z}^*_{..}$
       $\hat{\alpha}_i^s = \begin{cases} \bar{Z}^*_{i.} - \hat{\mu}^s & \forall\ i : \hat{\rho}_i^{s-1} = 1 \\ 0 & \text{otherwise.} \end{cases}$
       $\hat{\beta}_j^s = \begin{cases} \bar{Z}^*_{.j} - \hat{\mu}^s & \forall\ j : \hat{\kappa}_j^{s-1} = 1 \\ 0 & \text{otherwise.} \end{cases}$
5. Update cluster membership parameters
       $\hat{\rho}_i^s = \begin{cases} 1 & \sum_j [Z_{ij} - \hat{\kappa}_j^{s-1}(\hat{\mu}^s + \hat{\alpha}_i^s + \hat{\beta}_j^s)]^2 < \sum_j Z_{ij}^2 \\ 0 & \text{otherwise.} \end{cases}$
       $\hat{\kappa}_j^s = \begin{cases} 1 & \sum_i [Z_{ij} - \hat{\rho}_i^{s-1}(\hat{\mu}^s + \hat{\alpha}_i^s + \hat{\beta}_j^s)]^2 < \sum_i Z_{ij}^2 \\ 0 & \text{otherwise.} \end{cases}$
6. Repeat steps 4 and 5 for $s = 2 \dots S$ iterations
7. Compute $\hat{\mu}^{S+1}$, $\hat{\alpha}_i^{S+1}$ and $\hat{\beta}_j^{S+1}$ as in step 4
8. Prune bicluster to remove ill-fitting genes and samples; update layer effects again
9. Calculate layer sum of squares
       $LSS = \sum_{i,j} (\hat{\mu} + \hat{\alpha}_i + \hat{\beta}_j)\hat{\rho}_i \hat{\kappa}_j$
10. Permute $\hat{Z}_{ij}$ and follow steps 2 to 9; repeat $T$ times
11. Accept bicluster if LSS is greater than LSS for all permuted runs, otherwise stop
12. Refit all layers in the model $R$ times, then search for next layer
```

Fig. 1. Outline of Algorithm 1: the plaid model algorithm used by Turner et al. [26]. The layer index k is dropped for simplicity.

In Sections 3 and 4, we shall introduce extensions of the above plaid model that cater to microarray experiments that are more structured than the basic gene by sample matrix considered so far. Before moving on to these extensions, we propose two variations of Algorithm 1, which together tend to produce more concise and interpretable results.

The first variation is a modification of the pruning method. It is reasonable to suppose that genes are more likely to fit the profile of a bicluster with a low number of samples than the profile of a bicluster with a higher number of samples. To take this into account, the sums of squares used in the pruning step can be adjusted for the associated degrees of freedom. This will introduce a bias against genes for which the number of parameters in the layer model is high relative to the number of corresponding expression levels, as it would be in a bicluster with a low number of samples. The gene pruning is adjusted as follows:

$$\tilde{\rho}_i = \begin{cases} 1 & \text{if } \hat{\rho}_i = 1 \text{ and } \dfrac{\sum_{j:\hat{\kappa}_j=1} (\hat{Z}_{ij} - \hat{\kappa}_j(\hat{\mu} + \hat{\alpha}_i + \hat{\beta}_j))^2}{df_{res}} < \\ & \qquad (1 - \tau_1) \dfrac{\sum_{j:\hat{\kappa}_j=1} \hat{Z}_{ij}^2}{df_{tot}}, \\ 0 & \text{otherwise,} \end{cases}$$

where the total degrees of freedom, df_{tot}, is $\left(\sum_i \hat{\rho}_i\right)\left(\sum_j \hat{\kappa}_j\right)$ and the residual degrees of freedom, df_{res} is

$$df_{tot} - \left(\sum_i \hat{\rho}_i + \sum_j \hat{\kappa}_j - 1\right).$$

Note these are not the degrees of freedom for the sum of squares in the pruning criteria, but the degrees of freedom for the equivalent sum of squares relating to the whole bicluster, which leads to the same logical comparison and simplifies implementation. τ_1 and τ_2 are now interpreted as the minimum *adjusted* R^2 desired for the genes and samples. If the residual degrees of freedom is zero, the bicluster cannot be pruned and we suggest that the bicluster should be rejected in this case.

Sample pruning can be adjusted in the same way to deal appropriately with biclusters with a low number of genes. When using this variation of the pruning method, the layer effects should be updated after pruning either genes or samples, so that the degrees of freedom in the pruning step reflect the degrees of freedom in estimating the layer effects.

The second variation concerns the interpretability of biclusters. Under the plaid model, it is possible that a bicluster may contain genes with a similar pattern of *change* in expression level (i.e., similar sample effects), but completely different expression profiles in terms of up and down-regulation (allowed for by gene effects). For example, up-regulated genes may appear in a generally down-regulated bicluster. This is undesirable since one would expect biclustered genes to share the same broad features and for these features to dominate any gene or sample-specific parameters.

The Plaid™ software [15], based on the original plaid model algorithm proposed by Lazzeroni and Owen [12] has "unisign" options to favor consistent clusters. In particular, the "unisignrow" option prefers genes for which $\mu + \alpha_i$ is the same sign as μ and the "unisigncol" option prefers samples for which $\mu + \beta_j$ is the same sign as μ. These options are described as a preference that, even if met, can be violated by back fitting or bicluster pruning. Therefore, we presume that some weighting is applied during the stages of the fitting process in which the binary constraints on the cluster membership parameters are relaxed, which would not transfer to the binary least squares algorithm used here.

For the binary least squares algorithm, we prefer to address the problem of inconsistent biclusters by the use of *search models*, simplified models that represent the features of a bicluster that are considered to be most important. We propose that the search model is used within the layer iterations (Steps 4 and 5 of Fig. 1), then the full plaid model is fitted before pruning the bicluster and back fitting. It is not appropriate to use a simplified model throughout the algorithm as we expect to see gene and sample effects in practice and do not wish to treat these effects in the same way as pure error.

For example, to search for biclusters that may be described as simply up-regulated or down-regulated, the suitable search model would be $Z_{ij} = \rho_i \kappa_j \mu + \epsilon_{ij}$. By using the mean effect to search for genes and samples to add to the bicluster, the expression levels corresponding to the bicluster will generally be closer to the cluster mean than to the null model. For this reason, when gene and sample effects are added at the end, it is unlikely that the fitted values will be inconsistent with the global cluster effect. Although this is not guaranteed, we have always found the biclusters to remain consistent throughout pruning and back fitting, when using a mean effect search model on two-way data (we discuss the application to three-way data in Section 4).

Using this kind of search model can lead to instability in the first couple of layer iterations, due to the usual imbalance in the dimensions of the data (the number of genes being much greater than the number of samples). We have found this can be avoided by updating the cluster membership parameters in series, that is, using the updated $\hat{\rho}_i$ to update the $\hat{\kappa}_j$.

3 Grouped Data

We now turn to the first extension of the plaid model analysis. There is often more information available on the genes and samples in a microarray experiment than simply the gene expression data. For example, the samples may belong to certain treatment groups, or a functional classification of the genes may be known. Biclusters may be expected to correspond to these groups, in which case it may be useful to employ this information in the clustering process.

The plaid model can be fully supervised by clustering complete a priori groups instead of individual genes or samples. However, we would like to allow for misclassification, experimental error, and the presence of biclusters in the data that are unrelated to the external grouping. Therefore, we do not consider a fully supervised approach to be appropriate. Partial supervision is preferable and can be implemented by using a supervised model to start the search for a layer, reverting to the unsupervised model after a set number of iterations.

If searching on the basis of the full plaid model, the supervised layer model would be

$$Z_{g(i)h(j)} = \nu_g \omega_h (\mu + \alpha_{g(i)} + \beta_{h(j)}) + \epsilon_{g(i)h(j)}$$

in which ν_g is equal to one if gene group g is in the layer, zero otherwise; ω_h is equal to one if sample group h is in the layer, zero otherwise; and individual genes and samples are now indexed by $g(i)$, the ith gene in group g and $h(j)$ the jth sample in group h, respectively.

The group-level cluster membership parameters are updated using binary least squares, as follows:

$$\hat{\nu}_g = \begin{cases} 1 & \text{if } \sum_{g(i),h,h(j)} [Z_{g(i)h(j)} - \omega_h(\mu + \alpha_{g(i)} + \beta_{h(j)})]^2 < \\ & \qquad \sum_{g(i),h,h(j)} Z^2_{g(i)h(j)}, \\ 0 & \text{otherwise,} \end{cases}$$

$$\hat{\omega}_h = \begin{cases} 1 & \text{if } \sum_{h(j),g,g(i)} [Z_{g(i)h(j)} - \nu_g(\mu + \alpha_{g(i)} + \beta_{h(j)})]^2 < \\ & \qquad \sum_{h(j),g,g(i)} Z^2_{g(i)h(j)}, \\ 0 & \text{otherwise.} \end{cases}$$

Starting values for the group-level cluster membership parameters can either be found by "averaging" or "conversion." The first method averages the expression levels within each group so that the groups can be treated as if they were individual genes or samples in an unsupervised analysis.

In the conversion method, starting values are found on a gene and sample level, then converted to group-level starting values by taking the majority vote within each group. If the proportion of genes or samples selected in this way exceeds 0.5, then the complementary set of genes or samples is taken to be consistent with the notion that a bicluster represents an *unusual* pattern in the context of the data. If the converted group-level starting values for genes or samples are all zero, then the preconversion starting values are used.

We have described partial supervision for the general case of two-way supervision, but, of course, the model may be supervised in one dimension only if required. Since the supervised model effectively treat groups as individuals, there can be no overlap in group membership. However, an overlapping group structure can still be represented by considering the overlap as a separate group. Genes or samples for which the classification is unknown can be considered as groups of size one.

We illustrate the effect of partial supervision on data taken from a study on a range of human diseases. The subset of arrays that we consider are the first batch of arrays in the experiment, one array for each of 19 patients with some form of tuberculosis (TB). The patients may be classified into three disease groups: pulmonary TB (seven patients), pulmonary TB with complications (five patients), and TB meningitis (seven patients). Blood samples were taken from each patient at presentation and the total RNA extracted, amplified, and hybridized against a common reference (Stratagene®).

For the pulmonary TB patients, blood samples were taken premedication; however, this was not possible for the TB meningitis patients as this disease is far more serious. Therefore, the arrays may also be grouped as before or after medication.

Plate effects were removed from the background-corrected signal intensities using the robust median correction available in the LIMMA R package [19]. The log-ratio (base 2) of the sample intensity to reference intensity was then calculated. Print-tip (spatial) effects were removed from these log-ratios using an intensity-dependent loess normalization [28]. Then, finally, a robust scale correction was applied to log-ratios on each array, so that the arrays were comparable [28].

The genes were then filtered to remove those genes with more than 20 percent of the expression levels missing in any one class. This left 28,339 genes in the data set. The remaining missing values were interpolated with the row mean plus the column mean minus the global mean.

To begin with, we analyzed the data using an unsupervised plaid model algorithm (Algorithm 2, Fig. 2), implementing some of the modifications described in previous sections. In particular, Algorithm 2 is an example of using a mean effect search model.

The results from the unsupervised plaid model are summarized Table 1. For each layer in the model, Table 1 shows the number of genes and samples in the layer, the degrees of freedom associated with the layer effects, the sum of squares, and the mean square. The mean of the fitted layer, $\hat{\mu}$, is also given for the bicluster layers. Since a mean effect search model was used, the layer means summarize the main feature of the biclusters. So, the first layer represents a bicluster of 397 genes up-regulated over seven of the samples (layer mean 0.76) and the second layer represents a bicluster of 203 genes down-regulated over 6 of the samples (layer mean -0.54). Also shown in Table 1 is the model adjusted R^2 value of 0.891, showing that the fitted plaid model explains a high proportion of the data variance.

The layer sum of squares gives an indication of the importance of each cluster since it measures the variation uniquely characterized by the layer model. Since the layers are found sequentially, the layers would ideally be found in order of sum of squares, so that the most important effects are added to the model first. This is the case here, as Table 1 shows the layer sum of squares for the first cluster is 1,923 and the layer sum of squares for the second cluster is 408. The first cluster also has a higher mean square than the second cluster (4.77 compared to 1.96), showing that the larger sum of squares corresponding to the first cluster is not simply due to a greater number of genes and samples.

Fig. 3 shows the gene and sample-centered expression levels for the biclustered genes across all samples. The first bicluster contains four of the seven samples from pulmonary TB patients and three of the five samples from pulmonary TB patients with complications. The second bicluster contains a subset of the samples in the first bicluster, with one fewer sample from the pulmonary TB with complications group. Therefore, neither bicluster corresponds well to the known grouping of the patients, even if the pulmonary TB groups are considered as a whole.

In Fig. 3, the expression levels outside of the biclusters appear similar in magnitude to the expression levels inside the biclusters. This is a consequence of using the gene and sample-centered expression levels: Since the data contains biclusters, the gene and sample means are not the best estimators of the "typical" expression level for each gene and sample. In the plaid model, the background layer models the typical expression level in the context of the experiment and the biclusters represent departures from this. If the residuals from the background layer had been used in Fig. 3, most of the data values outside of the biclusters would be close to zero, emphasising the abnormality of the bicluster. However, the background layer is clearly model-dependent; therefore, we have used gene and sample-centered expression levels to allow comparison with later results.

We analyzed the TB variants data set again using a partially supervised version of Algorithm 2. The samples were labeled as pulmonary TB (P), pulmonary TB with complications (P+), or TB meningitis (M) and a group-level cluster membership parameter was used instead of a sample cluster membership parameter in the first five iterations of each layer. After these five iterations, the unsupervised model was used as in Algorithm 2. The algorithm was not allowed to terminate until at least one unsupervised iteration had been conducted, so that convergence was reached on the basis of individual samples. Starting values were obtained by the conversion method described earlier.

Table 2 summarizes the results from the partially supervised analysis. This plaid model has three biclusters a down-regulated bicluster with 872 genes and 10 samples a down-regulated bicluster with 200 genes and six samples

```
1. Compute $\hat{Z}_{ij}$: matrix of residuals from model so far
2. Cluster genes and samples using 5 k-means iterations with $k = 2$ and let $\hat{\rho}_i^0$ and $\hat{\kappa}_j^0$ indicate
   the smaller cluster in each case
3. Set $s = 1$
4. Update layer mean
       $\hat{\mu}^s = \frac{\sum_{i,j} \hat{\rho}_i^{s-1} \hat{\kappa}_j^{s-1} Z_{ij}}{(\sum_i \hat{\rho}_i^{s-1})(\sum_j \hat{\kappa}_j^{s-1})}$
5. Update cluster membership parameters
       $\hat{\rho}_i^s = \begin{cases} 1 & \sum_j [Z_{ij} - \hat{\kappa}_j^{s-1} \hat{\mu}^s]^2 < \sum_j Z_{ij}^2 \\ 0 & \text{otherwise.} \end{cases}$
       $\hat{\kappa}_j^s = \begin{cases} 1 & \sum_i [Z_{ij} - \hat{\rho}_i^s \hat{\mu}^s]^2 < \sum_i Z_{ij}^2 \\ 0 & \text{otherwise.} \end{cases}$
6. Repeat steps 4 and 5 for $s = 2, \ldots$ until convergence at iteration $S$
7. Compute $\hat{\mu}^{S+1}$, $\hat{\alpha}_i^{S+1}$ and $\hat{\beta}_j^{S+1}$ as Algorithm 1, step 4
8. Prune genes with $\tau_1 = 0.7$, adjusting for degrees of freedom; update layer effects
9. Prune samples with $\tau_2 = 0.7$, adjusting for degrees of freedom; update layer effects
10. Repeat steps 8 and 9 until convergence
11. Calculate layer sum of squares as Algorithm 1, step 9
12. Permute $\hat{Z}_{ij}$ and follow steps 2 to 11; repeat 3 times
13. Accept bicluster if LSS is greater than LSS for all permuted runs, otherwise stop
14. Refit all layers in the model twice; search for next layer
```

Fig. 2. Outline of Algorithm 2. the unsupervised plaid model algorithm used to analyze the TB variants data.

and an up-regulated bicluster with 694 genes and six samples. The gene and sample-centered expression levels for the biclustered samples are shown in Fig. 4.

The first bicluster from the partially supervised analysis contains genes that have a similar profile over *all* the samples to the genes in the first bicluster from the unsupervised analysis. However, since the partially supervised analysis favors biclusters corresponding to complete groups, the profile is interpreted as a down-regulation occurring mainly in the TB meningitis samples. With the emphasis on similarity within the TB meningitis samples, a larger group of genes is discovered: 872 genes in the first bicluster of the partially supervised plaid model compared to 397 genes in the first bicluster of the unsupervised plaid model. The two biclusters have 220 genes in common; these are identified as group A in Fig. 4.

The second bicluster from the partially supervised analysis, which contains four of the seven samples from

TABLE 1
Results from the Unsupervised Analysis of the TB Variants Data

Layer	Genes	Samples	Df	SS	MS	Mean
0	28339	19	28357	569285.62	20.08	n/a
1	397	7	403	1922.61	4.77	0.76
2	203	6	208	408.10	1.96	-0.54

Model $R^2_{adj} = 0.891$

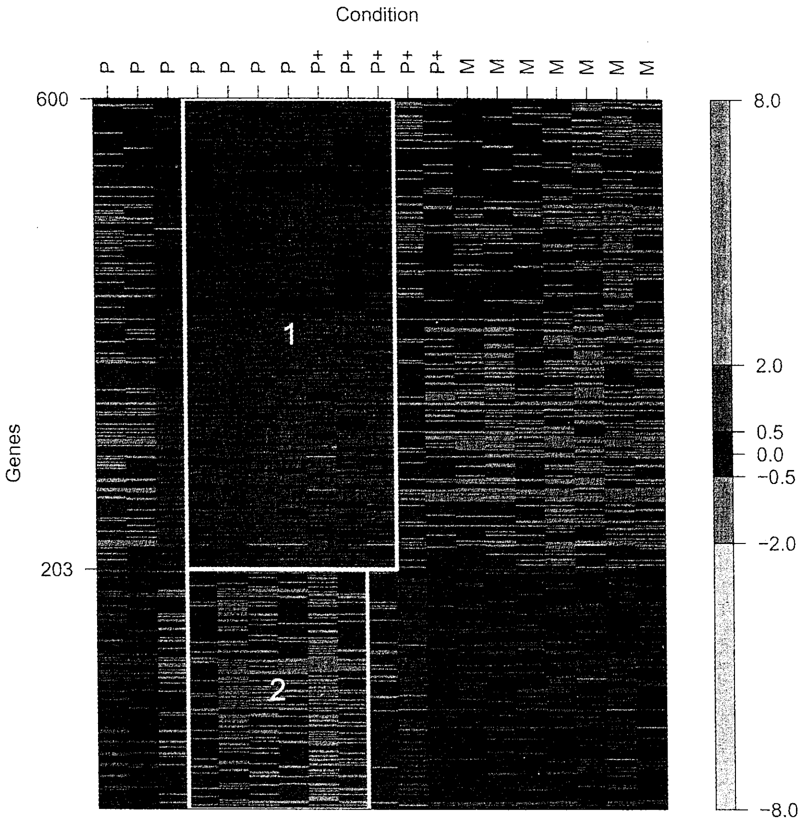


Fig. 3. Gene and sample-centered expression levels for genes biclustered in the unsupervised analysis of the TB variants data where P denotes pulmonary TB, P+ denotes pulmonary TB with complications, and M denotes TB meningitis.

pulmonary TB patients and two of the five samples from pulmonary TB patients with complications, is virtually identical to the second bicluster from the unsupervised analysis. In fact, all 200 genes in the partially supervised bicluster, identified as group B in Fig. 4, are also in the unsupervised bicluster, meaning that these two biclusters differ by only three genes. This shows that the partial supervision does not prevent the discovery of biclusters that do not closely correspond to the a priori structure and indeed the performance of the algorithm is equivalent to the unsupervised analysis in this case.

The third bicluster in the partially supervised model represents a feature of the data that was not discovered in the unsupervised analysis. It identifies a group of genes that are down-regulated in five out of the seven TB meningitis samples and one sample from the pulmonary TB with complications group. In this case, the search for a bicluster has clearly been assisted by the use of grouping information. The third bicluster may be considered as important as the second bicluster in this analysis since it has a comparable mean square (1.87 compared to 2.06). Therefore, the additional information revealed by the use of partial supervision is nontrivial.

The adjusted R^2 of the partially supervised model is 0.894 (Table 2), only slightly higher than the adjusted R^2 of the unsupervised model which was 0.891. However, changes in the bicluster structure are unlikely to make a large difference to the model adjusted R^2 when the background layer already accounts for so much of the variation: The background-only model has an R^2 of 0.889.

The advantage of the partially supervised model can be seen in the improved correspondence between the biclusters and the sample groups. This correspondence can be quantified by calculating the average purity and efficiency of the biclusters [7], where purity and efficiency are defined as follows:

TABLE 2
Results from the Partially Supervised Analysis of TB Variants Data

Layer	Genes	Samples	Df	SS	MS	Mean
0	28339	19	28357	572730.94	20.2	n/a
1	872	10	881	4572.97	5.19	-0.66
2	200	6	205	423.05	2.06	-0.55
3	694	6	699	1309.71	1.87	0.52

Model $R^2_{adj} = 0.894$

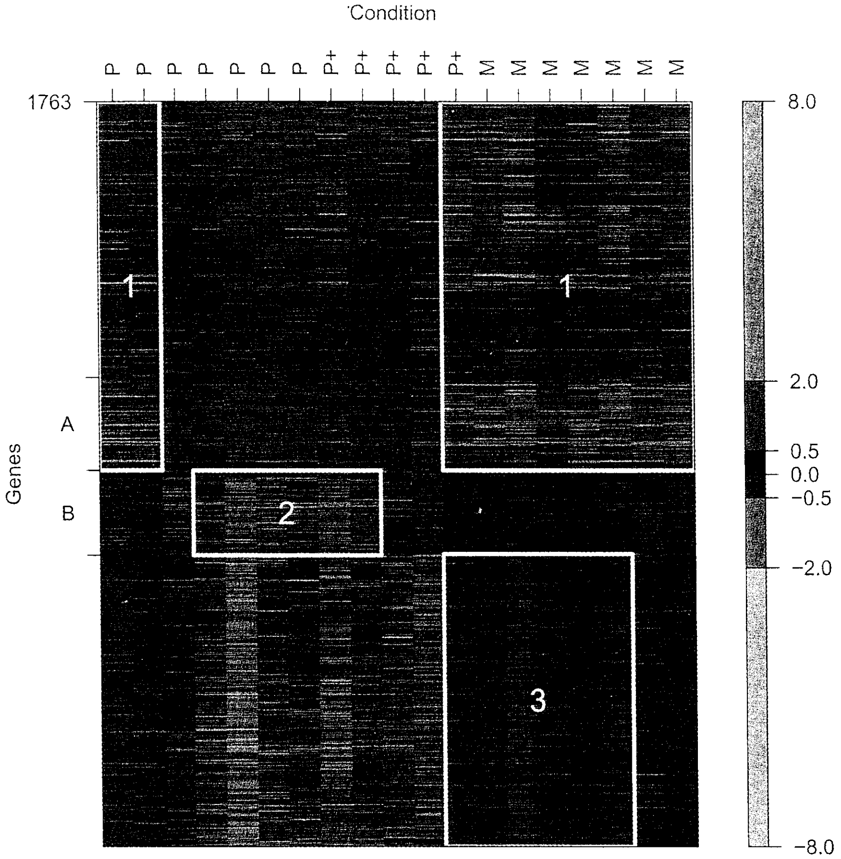


Fig. 4. Gene and sample-centered expression levels for genes biclustered in the partially supervised analysis of the TB variants data where P denotes pulmonary TB, P+ denotes pulmonary TB with complications, and M denotes TB meningitis.

$$\text{purity} = \frac{|c \cap g|}{|c|}$$

$$\text{efficiency} = \frac{|c \cap g|}{|g|}$$

in which c is the bicluster for which the measure is being obtained and g is the dominant group in that cluster. For the unsupervised model, the average purity and efficiency are 0.667 and 0.571, respectively, while the equivalent values for the partially supervised model are 0.833 and 0.714, showing a marked improvement.

4 Repeated Measures Data

A natural extension of the typical genes by sample microarray experiment is to measure expression levels over time for each gene in each sample. This poses a problem for analysis as most existing clustering methods require a two-way data structure. Clearly, a data set with repeated measures can be arranged in a two-way format, as a genes by sample matrix. However, biclustering the data in this format would not be appropriate as the resultant biclusters could contain arrays from noncontiguous time points or arrays from different samples at different time points, which may be difficult to interpret.

An alternative way of obtaining a two-way data structure is to collapse the data set by averaging over the samples. However, this is only reasonable if the samples belong to a single group or if the relationship between multiple sample groups can be summarized by a single contrast of interest, such as the difference in gene expression level between the first two sample groups and the last three sample groups. Otherwise, averaging over the samples results in a loss of information. In this case, the data might be analyzed one group, or one time point at a time. Segregating the analysis in this way makes it difficult to get the full picture and, again, important features may be missed.

The plaid model in the form described so far also requires data to be in a two-way format. However, the plaid model can be extended to handle repeated measures on genes and samples. The concept of a bicluster still applies in the sense that a bicluster is a subset of genes with a similar expression over a subset of samples, but now it is not individual expression levels that are clustered but complete expression profiles over time. In terms of the model, all that is required is the addition of parameters in each layer to account for changes in expression level over time. The simplest way to achieve this is to add time main effects to the model for each layer.

In detail, the expression level, Y_{ijt}, $i = 1, \ldots, n; j = 1, \ldots, p, t = 1, \ldots, q$ of the ith gene in the jth sample at time t is estimated by the following general linear model:

$$Y_{ijt} = \mu_0 + \alpha_{i0} + \beta_{j0} + \tau_{t0} + \sum_{k=1}^{K} \rho_{ik}\kappa_{jk}(\mu_k + \alpha_{ik} + \beta_{jk} + \tau_{tk}) + \epsilon_{ij}. \tag{1}$$

This model is also an extension of the generalized INDCLUS model [5], which may be expressed as

$$Y_{ijt} = \tau_{t0} + \sum_{k=1}^{K} \rho_{ik}\kappa_{jk}(\tau_{tk}) + \epsilon_{ij}.$$

As in Algorithm 1 (Fig. 1), the layer effects in the repeated measures plaid model may be estimated by ordinary least squares and the membership parameters in the model by binary least squares. When considering a single layer, we shall again omit the layer index for simplicity.

The k-means method for obtaining starting values for the cluster membership parameters needs to be adapted for three-way data. We propose two approaches along the lines of those suggested for the supervised two-way model. The equivalent conversion method is to find starting values at each time point and establish a final set of starting values by majority vote. The equivalent averaging method is to average over time points and proceed as for two-way data.

In repeated measures data, the focus is on changes in expression level over time. Therefore, a sensible search model to use for the analysis of repeated measures over genes and samples would be

$$Z_{ijt} = \rho_i\kappa_j(\mu + \tau_t) + \epsilon_{ijt}$$

as opposed to the mean only model recommended for the analysis of two-way data. For repeated measures data, the gene and sample effects in the full layer model are expected to be small enough not to change the overall profile in terms of up and down-regulation.

The permutation test (Steps 10 and 11, Fig. 1) also needs to be adapted for repeated measures data. It is no longer suitable to permute the data over all dimensions simultaneously since the data is ordered in one dimension. Rather, the data should be permuted within each time point, to give a random profile over time for each gene and sample pair. The candidate bicluster is then compared to one that may be found in a set of random expression profiles over time.

We shall illustrate the potential of the repeated measures plaid model using data from an ongoing experiment into human genetic susceptibility to tuberculosis (TB). The available data consists of 64 arrays, comprising eight experimental runs of eight time points. Five of the runs use blood samples infected with BCG-*lux*, a luciferase transformed version of the TB bacillus that is used for vaccination, which can stimulate an immune response but does not cause TB. Three of these BCG-*lux* positive runs use blood samples taken from individuals classified as resistant to TB and the remaining two use blood samples taken from individuals classified as susceptible. The other three runs in the data set are BCG-*lux* negative controls for two of the three resistant individuals and one of the two susceptible individuals. Control runs were not carried out for the remaining individuals due to financial constraints.

On each array, the amplified RNA was compared to a standard reference. Replicated baseline arrays were available for each time series, so that the expression levels could be made relative to their preinfection values. All the arrays were normalized following the procedure, discussed in Section 3, that was used for the arrays in the TB variants data set. Genes with more than 20 percent of data missing across all 64 arrays were removed from the data set. This left 30,897 genes and the remaining missing values were interpolated as before.

The classification of individuals was made on the basis of a BCG-*lux* growth assay [11], which assesses the ability of an individual to restrict growth of BCG-*lux* via a light ratio. One of the individuals classified as resistant by the growth assay was previously thought to be susceptible on the basis of reaction to a tuberculin skin test. Although the growth assay takes precedence, due to this contradictory result and due to the variability inherent in bioassays, we chose not to supervise our analysis on the basis of the classification of individuals. As the classification of individuals and the experimental treatment were considered to be equally important, it was not appropriate to supervise the analysis by treatment alone and, therefore, an unsupervised analysis was used. Fig. 5 outlines the method used (Algorithm 3), which takes the same approach as the unsupervised analysis of the TB variants data, adapted for the repeated measures model using a search model with mean and time effects only.

The results of the analysis of the TB susceptibility data are given in Table 3. Bicluster layers are now summarized by the truncated mean profile over time, to capture the main features of the biclustered time series. A fuller picture is given in Fig. 6, which shows the expression levels over time for the biclustered genes and samples after the fitted background layer has been subtracted.

There are three biclusters in the fitted plaid model. The first bicluster contains 1,244 genes and a BCG-*lux* positive run from the susceptible individual for which no control run was conducted. The genes in this bicluster are down-regulated over the first three time points (2 to 8 hours) and possibly down-regulated at 48 hours. The second bicluster contains 1,164 genes and both BCG-*lux* positive and BCG-*lux* negative runs from the other susceptible individual. In this bicluster, genes are down-regulated over the three time points from 4 to 12 hours. A similar pattern is seen in the third bicluster, which contains the same runs and a different set of 21 genes, except that the down-regulation from 4 to 12 hours is not as great.

Nearly 50 percent of the genes in the first bicluster are also in the second bicluster. So, it is possible that an early down-regulation in these genes is characteristic of susceptible individuals, but there is some variability between individuals as to when this down-regulation occurs. The plaid model has not identified a similar pattern for these genes in the resistant individual for which the classification from the BCG-*lux* growth assay contradicted prior belief, nor is such a pattern evident in the data. Therefore, the gene expression data appears to be consistent with the results from the growth assay.

The relationship between the first two biclusters caused some difficulty in the analysis. In particular, if the averaging method was used to find starting values, then depending on the seed used to initiate the k-means clustering, the analysis would usually identify a bicluster of genes that are down-regulated from 2 to 8 hours in the BCG-*lux* positive run for the first susceptible individual and

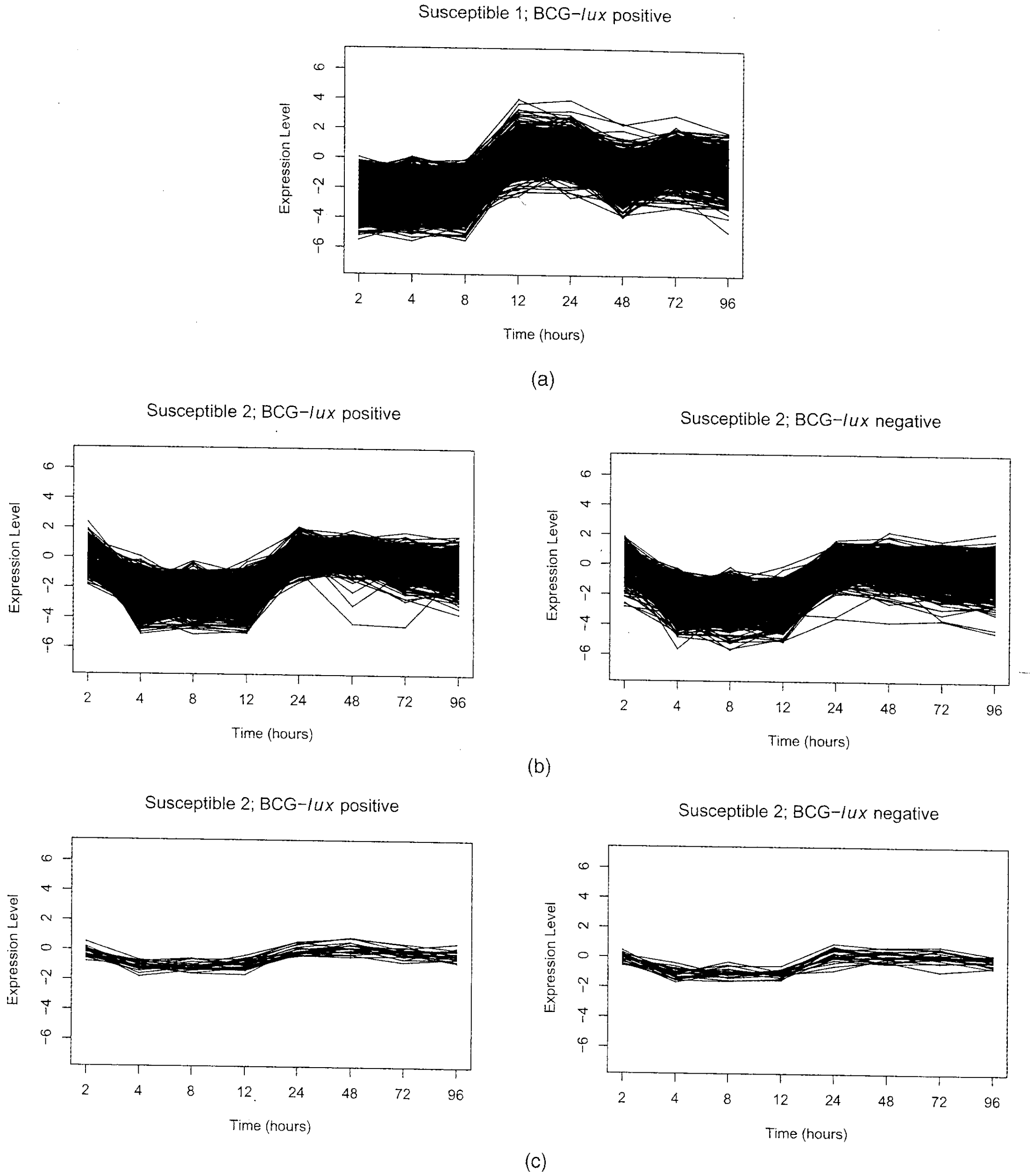
```
1. Compute $\hat{Z}_{ij}$: matrix of residuals from model so far
2. For each $t$ find $\hat{\rho}^0_{it}$ and $\hat{\kappa}^0_{jt}$ as in Algorithm 2, step 2; convert to $\hat{\rho}^0_i$ and $\hat{\kappa}^0_j$ by majority vote
3. Set $s = 1$
4. Update layer effects
       $\hat{\mu}^s = \frac{\sum_{i,j,t} \hat{\rho}_i^{s-1} \hat{\kappa}_j^{s-1} Z_{ijt}}{q(\sum_i \hat{\rho}_i^{s-1})(\sum_j \hat{\kappa}_j^{s-1})}$        $\hat{\tau}_t^s = \frac{\sum_{i,j} \hat{\rho}_i^{s-1} \hat{\kappa}_j^{s-1} Z_{ijt}}{(\sum_i \hat{\rho}_i^{s-1})(\sum_j \hat{\kappa}_j^{s-1})} - \hat{\mu}^s$
5. Update the cluster membership parameters
       $\hat{\rho}_i^s = \begin{cases} 1 & \sum_{j,t}[Z_{ijt} - \hat{\kappa}_j^{s-1}(\hat{\mu}^s + \hat{\tau}_t^s)]^2 < \sum_{j,t} Z_{ijt}^2 \\ 0 & \text{otherwise.} \end{cases}$
       $\hat{\kappa}_j^s = \begin{cases} 1 & \sum_{i,t}[Z_{ijt} - \hat{\rho}_i^s(\hat{\mu}^s + \hat{\tau}_t^s)]^2 < \sum_{i,t} Z_{ijt}^2 \\ 0 & \text{otherwise.} \end{cases}$
6. Repeat steps 4 and 5 for $s = 2, \ldots$ until convergence at iteration $S$
7. Compute $\hat{\mu}^{S+1}$, $\hat{\tau}_t^{S+1}$ and
       $\hat{\alpha}_i^{S+1} = \frac{\sum_{j,t} \hat{\rho}_i^S \hat{\kappa}_j^S Z_{ijt}}{q \sum_j \hat{\kappa}_j^S} - \hat{\mu}^S$        $\hat{\beta}_j^{S+1} = \frac{\sum_{i,t} \hat{\rho}_i^S \hat{\kappa}_j^S Z_{ijt}}{q \sum_i \hat{\rho}_i^S} - \hat{\mu}^S$
8. Prune genes as below, updating layer effects afterwards
       $\tilde{\rho}_i = \begin{cases} 1 & \text{if } \hat{\rho}_i = 1 \text{ and } \frac{\sum_{j,t}(\hat{Z}_{ijt} - \hat{\kappa}_j(\hat{\mu} + \hat{\alpha}_i + \hat{\beta}_j + \hat{\tau}_t))^2}{df_{res}} < (1 - 0.7)\frac{\sum_j \hat{Z}_{ij}^2}{df_{tot}}, \\ 0 & \text{otherwise.} \end{cases}$
   using $df_{tot} = q(\sum_i \rho_i)(\sum_j \kappa_j)$ and $df_{res} = df_{tot} - (\sum_i \rho_i + \sum_j \kappa_j + \sum_t \tau_t - 2)$
9. Similarly prune samples with $\tau_2 = 0.7$, then update layer effects
10. Repeat steps 8 and 9 until convergence
11. Calculate layer sum of squares as in Algorithm 1, step 9
12. Permute $\hat{Z}_{ij}$ and follow steps 2 to 11; repeat 3 times
13. Accept bicluster if LSS is greater than LSS for all permuted runs, otherwise stop
14. Refit all layers in the model twice; search for next layer
```

Fig. 5. Outline of Algorithm 3: the plaid model algorithm used to analyze the TB susceptibility data. The layer index k has been dropped for simplicity.

TABLE 3
Results from the Analysis of the TB Susceptibility Data

Layer	Genes	Samples	Df	SS	MS	Truncated Mean Profile
0	30897	8	30911	271518.04	8.78	n/a
1	1244	1	1251	24868.01	19.88	-2 -2 -2 0 0 -1 0 0
2	1164	2	1172	41974.73	35.81	0 -2 -2 -2 0 0 0 0
3	21	2	29	165.81	5.72	0 -1 -1 -1 0 0 0 0

Model $R^2 = 0.352$


Susceptible 1; BCG-*lux* positive

(a)

Susceptible 2; BCG-*lux* positive

Susceptible 2; BCG-*lux* negative

(b)

Susceptible 2; BCG-*lux* positive

Susceptible 2; BCG-*lux* negative

(c)

Fig. 6. Biclustered profiles from the analysis of the TB susceptibility data after the fitted background layer has been subtracted.

down-regulated from 4 to 12 hours in both BCG-*lux* positive and BCG-*lux* negative runs for the second susceptible individual. However, the layer model would not fit very well since the main effects model (1) assumes that the time effects are the same across all the biclustered genes and runs, which does not allow for a time shift between individuals. Therefore, most of the genes would be pruned out and the analysis would not go on to identify the two profiles separately. The only way this could be achieved was to use the conversion method to single out the first susceptible individual as in the analysis presented here. A side effect of choosing this method is that the layer sum of squares for the first bicluster is much lower than the layer sum of squares for the second bicluster (24,868 compared to 41,975). A more satisfactory approach might be to include gene by time interaction effects in the model so that a time shift between individuals could be allowed within a layer.

One reason why the first susceptible individual may be singled out by the conversion method is that there are more features of the data associated with this individual than just down-regulation over the first three time points for a subset of the genes. This is shown by the first bicluster in Fig. 6, which includes at least two different types of gene. All the genes are down-regulated over the first three time points, but, for some genes, this is the main feature, with a smaller down-regulation at 48 hours, while, for other genes, the initial down-regulation is marginal compared to an up-regulation from 12 to 24 hours. These two types of profiles are shown more clearly in Fig. 7, which displays the lower 3 percent and upper 3 percent of genes in the first bicluster as ordered by gene effect. These two features cover the

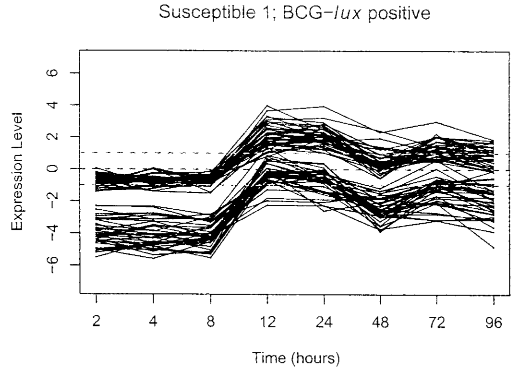


Fig. 7. Gene, sample, and time-centered expression levels for genes in the lower 3 percent or the upper 3 percent of the first bicluster from the analysis of the TB susceptibility data as ordered by fitted gene effect.

majority of time points between them, but are opposite in sign, which might explain how the first susceptible individual could be singled out by the conversion method, yet if selected by the averaging method is always selected with both samples from the second susceptible individual as the initial down-regulation is the dominating feature.

The presence of two types of profiles in the first bicluster also shows that the use of a search model has not been completely effective. Ideally, we would like to obtain biclusters of genes that are consistent in sign across all time points, not just those relevant to the main feature of the bicluster (such as the first three time points in the first bicluster here). Inconsistency does not appear to be a problem in the second and third biclusters, which suggests that the inconsistency in the first bicluster is due to the search model accommodating two features of the data which ought to be modeled separately because they do not relate to the same set of genes. This situation might be avoided if the bicluster was pruned on the basis of the search model. Clearly, the same level of fit could not be expected as for the model with gene and sample effects, but using say $\tau_1 = \tau_2 = 0.5$ would suggest the majority of the variation is explained by the mean profile over time. Alternatively, two stages of pruning might be considered: first at a low value of τ_1 and τ_2 to ensure consistency in sign, then at a higher value of τ_1 and τ_2 after gene and sample effects have been added to ensure that the final model fits well. In either case, gene and sample effects should still be included in the final layer model to avoid adding noise to the data when residuals are taken to find the next layer.

The proportion of variance explained by the plaid model in this analysis is only 0.352 (Table 3), which is far less than in the analysis of two-way data (e.g., $R^2_{adj} = 0.891$ for the unsupervised model, Table 1). This suggests that the main effects model used here may be oversimplistic. When analyzing two-way data with a mean-only search model, biclusters represent blocks of up-regulation and down-regulation in the gene expression matrix for which a two-way main effects model is more than adequate. When analyzing three-way data, on the other hand, biclustered profiles may be characterized by periods of down-regulation, periods of up-regulation, and periods of expression at the genes' usual level. These changes can be roughly approximated by time main effects, but clearly this approximation is insufficient. In addition, a fixed gene or sample effect may be insufficient to describe the gene or sample-specific up-regulation and down-regulation in an expression profile. For example, if a bicluster is characterized by a period of up-regulation and a period of down-regulation, a gene or sample which is up-regulated more than the bicluster average in the first period is not necessarily down-regulated to a lesser degree in the second period. Thus, a more flexible layer model may be necessary to effectively bicluster three-way data.

5 Discussion

We have shown that the previously proposed plaid model can be naturally extended to incorporate external grouping information or to bicluster profiles of repeated measures. Since the parameter updates are simple to compute, the plaid model therefore provides a flexible framework for biclustering large, structured microarray data sets, as exemplified in this paper.

The analysis of the TB variants data showed that using sample grouping to partially supervise the plaid model algorithm can help discover biclusters that closely correspond to the known structure of the data. Related research suggests that gene supervision could also be effective. Owen et al. [16] use a group of genes known to have a closely related function to search for a bicluster ("cassette") of genes with similar expression patterns. Starting with the input set of genes, their "gene recommender" algorithm identifies a subset of the samples over which the query genes are coexpressed, then searches for genes with similar expression profiles over these samples. This is similar to using a single gene group to specify the starting values for a single layer of the plaid model. Ihmel et al. [10] take a similar approach to Owen et al, but go further in the application of their "signature algorithm," using a diverse collection of input gene sets defined by common sequence, common function or one-way clusters. This leads to a large number of retrieved biclusters ("transcription modules") from which those passing an evaluation of reliability are selected. This approach is closer to partially supervising the plaid model by an external grouping defined over all the genes.

The algorithms of Owen et al. and Ihmel et al. are only supervised by gene groups. The GeneXPress tool of Segal et al. [21] is a related method that also takes sample groups into account. Starting with a diverse collection of input gene sets in the manner of Ihmel et al., the GeneXPress tool identifies the samples for each gene set in which a significant fraction of the gene set are coordinately expressed. Gene sets with similar expression signatures are merged and inconsistent genes are pruned out. Finally, the sample groups are identified in which the merged gene sets are significantly up-regulated or down-regulated to produce the final bicluster ("module"). Due to the manner in which genes are selected for the bicluster and the fact that sample groups are treated as fixed, this method is more strongly supervised than the two-way partially supervised approach proposed in this paper.

Although partial supervision can incorporate external grouping on the genes or samples, there may be further types of auxiliary data that could be relevant to the characterization of biclusters such as the survival rate of patients. Such information is not so easy to incorporate into the plaid model. It may be possible to add certain parameters or covariates into the layer model or, similar to supervised gene shaving [8], optimize the layer model and a model of a secondary response simultaneously. However, the algorithm would need to be tailored to the particular application and a more general approach such as the rich probabilistic models used by Segal et al. [22] may be more appropriate.

Although several model-based clustering methods have been developed for the analysis of time-course gene expression data [1], [20], [13], [27], [9], little work has been conducted in the area of clustering three-way data with one dimension over time. Recently, Qin and Self [18] proposed a model for such data, though they only illustrate their method for the special case of two-way gene expression data over time. For three-way data, they propose a linear mixed effects model with a fixed cluster effect curve and random gene and sample effect curves modeled using a spline basis. Their approach is designed for one-way clustering of the genes. As far as we are aware, biclustering times series over genes and samples has not been attempted before. The analysis of the TB susceptibility data suggested that more complex models than the main effects model (1) may be necessary to bicluster such data effectively and the model proposed by Qin and Self suggests an alternative. The use of a spline basis would be far more computationally intensive, however, and such an approach may only be feasible for a few hundred genes, rather than the thousands analyzed here.

Along with the two extensions of plaid model analysis discussed above, we have also suggested some variations of the algorithm proposed by Turner et al. [26] that may enhance its performance. These include modifications to bicluster pruning and the concept of search models to produce biclusters that have expression profiles of consistent sign. While search models appear to work satisfactorily for two-way data, they were not completely effective in the analysis of the three-way TB susceptibility data. This issue might be addressed by requiring a stronger level of fit to the search model, as suggested in Section 4.

The algorithms presented in this paper were implemented in R [19] and the code is provided as supplementary material online.

Acknowledgments

The microarray data described in this paper was generated as part of a collaboration between Imperial College, London, and the Relman Laboratory, Stanford University, that was funded by the Wellcome Trust.

References

[1] Z. Bar-Joseph, G. Gerber, D.K. Gifford, T.S. Jaakkola, and I. Simon, "A New Approach to Analyzing Gene Expression Time Series Data," *Proc. Sixth Ann. Int'l Conf. Computational Biology (RECOMB-02)*, pp. 39-48, 2002.

[2] Y. Barash and N. Friedman, "Context-Specific Bayesian Clustering for Gene Expression Data," *J. Computational Biology*, vol. 9, no. 2, pp. 169-191, 2002.

[3] A. Ben-Dor, B. Chor, R. Karp, and Z. Yakhini, "Discovering Local Structure in Gene Expression Data: The Order-Preserving Submatrix Problem," *Proc. Sixth Ann. Int'l Conf. Computational Biology (RECOMB-02)*, pp. 49-57, 2002.

[4] S. Busygin, G. Jacobsen, and E. Krämer, "Double Conjugated Clustering Applied To Leukemia Microarray Data," *Proc. Second SIAM ICDM, Workshop Clustering High Dimensional Data*, 2002.

[5] A. Chaturvedi and J.D. Carroll, "An Alternating Combinatorial Optimization Approach to Fitting the INDCLUS and Generalized INDCLUS Models," *J. Classification*, vol. 11, no. 2, pp. 155-170, 1994

[6] Y. Cheng and G.M. Church, "Biclustering of Expression Data," *Proc. Eighth Int'l Conf. Intelligent Systems for Molecular Biology (ISMB-2000)*, vol. 8, pp. 93-103, 2000.

[7] G. Getz, E. Levine, and E. Domany, "Coupled Two-Way Clustering Analysis of Gene Microarray Data," *Proc. Nat'l Academy of Science USA*, vol. 97, no. 22, pp. 12079-12084, 2000.

[8] T. Hastie, R. Tibshirani, M.B. Eisen, A. Alizadeh, R. Levy, L. Staudt, W.C. Chan, D. Botstein, and P. Brown, "'Gene Shaving' as a Method for Identifying Distinct Sets of Genes with Similar Expression Patterns," *Genome Biology*, vol. 1, no. 2, pp. 0003.1-0003.21, 2000.

[9] N.A. Heard, C.C. Holmes, and D.A. Stephens, "A Quantitative Study of Gene Regulation Involved in the Immune Response of Anopheline Mosquitos: An Application of Bayesian Hierarchical Clustering of Curves," *J. Am. Statistical Assoc.*, to appear.

[10] J. Ihmels, G. Friedlander, S. Bergmann, O. Sarig, Y. Ziv, and N. Barkai, "Revealing Modular Organization in the Yeast Transcriptional Network," *Natural Genetics*, vol. 31, no. 4, pp. 370-377, 2002.

[11] B. Kampmann, P.Ó. Gaora, V.A. Snewin, M.-P. Gares, D.B. Young, and M. Levin, "Evaluation of Human Antimycobacterial Immunity Using Recombinant Reporter Mycobacteria," *J. Infectious Diseases*, vol. 182, no. 3, pp. 895-901, 2000.

[12] L. Lazzeroni and A. Owen, "Plaid Models for Gene Expression Data," *Statistical Sinica*, vol. 12, no. 1, pp. 61-86, 2002.

[13] Y. Luan and H. Li, "Clustering of Time-Course Gene Expression Data Using a Mixed-Effects Model with B-Splines," *Bioinformatics*, vol. 19, no. 4, pp. 474-482, 2003.

[14] G.J. McLachlan, R.W. Bean, and D. Peel, "A Mixture Model-Based Approach to the Clustering of Microarray Expression Data," *Bioinformatics*, vol. 18, no. 3, pp. 413-422, 2002.

[15] A.B. Owen Plaid™ Software, http://www-stat.stanford.edu/~owen/clickwrap/plaid.html, 2005.

[16] A.B. Owen, J. Stuart, K. Mach, A.M. Villeneuve, and S. Kim, "A Gene Recommender Algorithm to Identify Coexpressed Genes in C. Elegans," *Genome Research*, vol. 13, no. 8, pp. 1828-1837, 2003.

[17] K.S. Pollard and M.J. van der Laan, "Statistical Inference for Simultaneous Clustering of Gene Expression Data," *Math. Bioscience*, vol. 176, no. 1, pp. 99-121, 2002.

[18] L.-X. Qin and S.G. Self, "The Clustering of Regression Models Method with Applications in Gene Expression Data," Technical Report 239, UW Biostatistics, Univ. Washington, http://www.bepress.com/uwbiostat/paper239, 2005.

[19] R Development Core Team, "R: A Language and Environment for Statistical Computing," R Foundation for Statistical Computing, Vienna, Austria, 2005.

[20] M.F. Ramoni, P. Sebastiani, and I.S. Kohane, "Cluster Analysis of Gene Expression Dynamics," *Proc. Nat'l Academy of Sciences USA*, vol. 99, no. 14, pp. 9121-9126, 2002.

[21] E. Segal, N. Friedman, D. Koller, and A. Regev, "A Module Map Showing Conditional Activity of Expression Modules in Cancer," *Natural Genetics*, vol. 36, no. 10, pp. 1090-1098, 2004.

[22] E. Segal, B. Taskar, A. Gasch, N. Friedman, and D. Koller, "Rich Probabilistic Models for Gene Expression," *Bioinformatics*, vol. 17, no. 90001, pp. S243-S252, 2001.

[23] Q. Sheng, Y. Moreau, and B. De Moor, "Biclustering Microarray Data by Gibbs Sampling," *Bioinformatics*, vol. 19, no. 2, pp. ii196-ii205, 2003.

[24] A. Tanay, R. Sharan, and R. Shamir, "Discovering Statistically Significant Biclusters in Gene Expression Data," *Bioinformatics*, vol. 18, no. 90001, pp. S136-S144, 2002.

[25] C. Tang, L. Zhang, A. Zhang, and M. Ramanathan, "Interrelated Two-Way Clustering: An Unsupervised Spproach for Gene Expression Data Analysis," *Proc. Second Ann. IEEE Int'l Symp. Bioinformatics and Bioeng. (BIBE 2001)*, pp. 41-48, 2001.

[26] H. Turner, T. Bailey, and W. Krzanowski, "Improved Biclustering of Microarray Data Demonstrated Through Systematic Performance Tests," *Computer Statistics Data Analysis,* vol. 48, no. 2, pp. 235-254, 2005.

[27] J.C. Wakefield, C. Zhou, and S.G. Self, "Modeling Gene Expression Data over Time: Curve Clustering with Informative Prior Distributions," *Bayesian Statistics 7, Proc. Seventh Valencia Int'l Meeting,* pp. 721-732, 2003.

[28] Y.H. Yang, S. Dudoit, P. Luu, D.M. Lin, V. Peng, J. Ngai, and T.P. Speed, "Normalization for cDNA Microarray Data: A Robust Composite Method Addressing Single and Multiple Slide Systematic Variation," *Nucleic Acids Research,* vol. 30, no. 4, p. e15, 2002.

Heather L. Turner received the BSc degree in applied statistics from The University of Reading (2000). As part of this degree, she worked for a year at the Institute for Arable Crop Research in Bristol. She was recently awarded the PhD degree for her work on biclustering microarray data at the University of Exeter. She is currently a research assistant at the University of Warwick, where she works on statistical methods for social science research.

Trevor C. Bailey received the MSc degree in mathematical statistics from Imperial College, London (1976), and the PhD degree from Exeter University (1981). He is currently a senior lecturer in statistics at Exeter University, having formerly lectured at the Australian Graduate School of Management, University of New South Wales, Australia. He is a chartered statistician and fellow of the Royal Statistical Society. His current research interests are in spatial statistics (particularly health applications and multivariate spatial methods) and applied statistical modeling more generally. He has published more than 60 articles in academic journals and conference proceedings and a book on spatial analysis.

Wojtek J. Krzanowski received the BSc degree in mathematics from Leeds University in 1967, a diploma in mathematical statistics from Cambridge University in 1968, and the PhD degree in applied statistics from Reading University in 1974. Following three-year appointments as a scientific officer at Rothamsted Experimental Station and senior research fellow at the RAF Institute of Aviation Medicine, he was successively a lecturer, senior lecturer, and reader in applied statistics at Reading University between 1974 and 1990. Since 1990, he has been a professor of statistics at Exeter University. His interests are in multivariate analysis, statistical modeling, classification, and computational methods. He has published five books, more than 20 contributions to books, and about 90 articles in scientific journals. He is a former joint editor of the *Journal of the Royal Statistical Society, Series C,* and has served on the editorial board of the *Journal of Classification* since its inception in 1984.

Cheryl A. Hemingway is a paediatric neurologist currently working as a clinical research fellow at Imperial College, London. She is employed through a Wellcome-Trust-Burroughs-Wellcome initiative, on a collaborative project involving Imperial College, London, the University of Cape Town, South Africa, and Stanford University. She is using microarray and other novel technologies to investigate the host response in children with central nervous system infections, such as TB meningitis and meningococcal meningitis.