The Canadian Journal of Statistics
Vol. xx, No. yy, 2015, Pages 1–22
La revue canadienne de statistique

A Simple Model-based Approach to Variable Selection in Classification and Clustering

Vahid Partovi Nia[1]* and Anthony C. Davison[2]

[1]*Corresponding author: GERAD Research Center and Department of Mathematical and Industrial Engineering, Polytechnique Montréal, 2900 Edouard-Montpetit, J3T 1J4 Montréal, Canada,* e-mail: vahid.partovinia@polymtl.ca
[2]*École Polytechnique Fédérale de Lausanne, EPFL-FSB-MATHAA-STAT, Station 8, 1015 Lausanne, Switzerland,* e-mail: anthony.davison@epfl.ch

Key words and phrases: Classification; clustering; high-dimensional data; hierarchical partitioning; Laplace distribution; mixture model; variable selection.

MSC 2010: Primary 62H30; secondary 62F15

Abstract: Clustering and classification of replicated data is often performed using classical techniques that inappropriately treat the data as unreplicated, or by complex modern ones that are computationally demanding. In this paper we introduce a simple approach based on a 'spike-and-slab' mixture model that is fast, automatic, allows classification, clustering and variable selection in a single framework, and can handle replicated or unreplicated data. Simulation shows that our approach compares well with other recently proposed methods. The ideas are illustrated by application to microarray and metabolomic data. *The Canadian Journal of Statistics* xx: 1–22; 2015 © 2015 Statistical Society of Canada

Résumé: Insérer votre résumé ici. We will supply a French abstract for those authors who can't prepare it themselves. *La revue canadienne de statistique* xx: 1–22; 2015 © 2015 Société statistique du Canada

1. INTRODUCTION

Modern problems in the biosciences often involve data on many variables, measured on fewer experimental units. As an example below we discuss metabolite fingerprinting, which involves the spectra of the total composition of metabolites, based on experimental techniques such as time-of-flight mass spectrometry, infrared spectrometry, or gas chromatography-mass spectrometry (Gohlke and McLafferty, 1993). Once metabolite profile data have been obtained, it may be desired to group different profiles and to classify new ones, and to say which metabolites are key to doing so. The number of profiles, the sample size, is typically smaller than the number of metabolites, or variables. In many cases the data are replicated, i.e., several observations are taken on the same sample under the same or similar experimental conditions.

Dimension reduction is an essential element of the analysis of modern biological data. One approach to this is projection to fewer dimensions, using techniques such as principal component analysis (Yeung and Ruzzo, 2001), independent component analysis (Scholz *et al.*, 2004), covariance reparametrisation of Gaussian mixtures (Bergé *et al.*, 2012) or projection pursuit (Friedman, 1987). It can be hard to interpret the results of these procedures, however, and they may obscure the clustering (Chang, 1983). For this reason variable selection is generally preferred. Below we describe simple fast variable selection procedures for classification and clustering with replicated data, which can also be applied in non-replicated cases.

* *Author to whom correspondence may be addressed.*
E-mail: vahid.partovinia@polymtl.ca

FIGURE 1: Subset of metabolite data and illustration of mixture model. Upper panel: profile for $T = 3$ plants (*WsWT*, *isa2*, *dpe2*) each with $R = 4$ replicates measured on $V = 15$ metabolites (*maltose.MX1*, ..., *aspartic.3*). Lower panel: ideal (solid) and realized (dashed) profiles and data (grey lines) for three classes, each with a single type, generated under the mixture model (1). Squares at the foot show variables active under the model of §2.3.

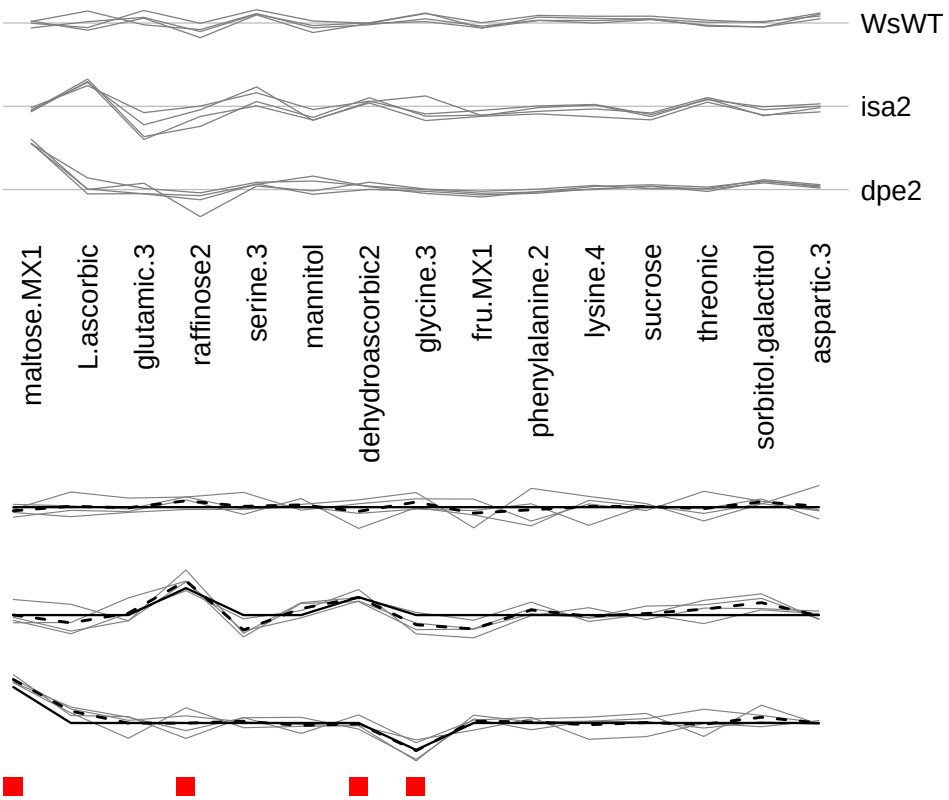


Figure 1 illustrates the type of data we consider. Its top part shows measurements on 15 metabolites obtained from three plants, each with four replicate measurements; see §4.2 for more details. Without loss of generality, the data have been centered so that the metabolite-wise averages equal zero. There is clear systematic variation of certain metabolites, but determination of the most important among them requires a model, illustrated in the lower part of the figure, where just four metabolites (shown by the squares) actively contribute to systematic variation in the profiles (the solid black lines). This systematic variation is obscured by two further layers of variability, one yielding the dashed profile, and another leading to the grey lines that represent the observed data. Thus the data are regarded as stemming from a mixture of discrete and continuous components, a so-called spike and slab model (Mitchell and Beauchamp, 1988).

Tadesse *et al.* (2005) and Kim *et al.* (2006), among others, have described fully Bayesian mixture models for clustering and variable selection that are fitted using Markov chain Monte Carlo simulation. Such models provide a coherent inferential framework, but the parameters of prior distributions must be chosen, auxiliary elements such as proposal distributions may need to be tuned and the convergence of a Markov chain to a complex distribution on a high-dimensional space must be checked. Such convergence may be difficult to ascertain, and the algorithms are sufficiently complex that a major effort is required to implement them; very often they cannot be systematically compared with other approaches, because applying them to many simulated

datasets is computationally infeasible.

In contrast, this paper describes a simple fast automatic approach. We aim to provide a 'mixture model-lite', intermediate between overly simple and computationally complex procedures, that mitigates some of the drawbacks of the former without paying the full tariff of the latter. Our approach may be regarded as empirical Bayesian: the prior parameters may be estimated by maximum likelihood. Despite its simplicity, in similar settings our procedure compares well with the algorithms of Kim *et al.* (2006), Tadesse *et al.* (2005) and Witten and Tibshirani (2010).

We do not attempt to survey the vast literatures on clustering, classification and variable selection, but rather give some key recent references. Classical clustering techniques (Kaufman and Rousseeuw, 1990) are based on measures of the similarity of different observations, and their outcomes are often represented graphically by a dendrogram. Modern clustering techniques (Everitt *et al.*, 2011) are mainly based on mixture models (e.g., McLachlan and Peel, 2000; Ghahramani and Beal, 2000; Fraley and Raftery, 2002; Heller and Ghahramani, 2005; Heard *et al.*, 2006; Booth *et al.*, 2008). Classification is a long-standing statistical problem with many solutions that are well described by Hastie *et al.* (2009). The literature on variable selection is huge; see Claeskens and Hjort (2008) for an overview of classical techniques, and Pan and Shen (2007), Wang and Zhu (2008) and Guo *et al.* (2010) for work closer to that described here. Bayesian variable selection and approximations thereto have been discussed by many authors, including George and McCulloch (1997), Raftery and Dean (2006), Tadesse *et al.* (2005), and Kim *et al.* (2006).

2. MIXTURE MODELS

2.1. Basic model

We suppose that measurements are available on a number of replicates of different types, and that these types are themselves grouped into classes. For example, a type might represent a plant, with replicates representing different leaves of that plant, and the goal would be to cluster the plants into disjoint classes on the basis of variables measured on the leaves, while assigning importances to the variables. In mathematical terms, we suppose that there are C classes, that class $c \in \{1, \ldots, C\}$ consists of T_c types, that there are R_{ct} replicates of the tth type, and that V variables are measured on each replicate, for $t \in \{1, \ldots, T_c\}$. The total number of types is $T = \sum_{c=1}^{C} T_c$, the total number of replicates is $\sum_{c=1}^{C} \sum_{t=1}^{T_c} R_{ct}$, and the total number of measurements is $V \sum_{c=1}^{C} \sum_{t=1}^{T_c} R_{ct}$. Often V is tens of thousands or more, whereas T is at very most a few hundreds. If there is no replication, then $R_{ct} = 1$ $(t = 1, \ldots, T_c; c = 1, \ldots, C)$. The result of each measurement is a scalar y_{vctr}, which we assume may be expressed as

$$y_{vctr} = \mu + \gamma_{vc}\theta_{vc} + \eta_{vct} + \varepsilon_{vctr}, \qquad (1)$$
$$v = 1, \ldots, V, c = 1, \ldots, C, t = 1, \ldots, T_c, r = 1, \ldots, R_{ct},$$

where θ_{vc}, η_{vct} and ε_{vctr} are independent continuous random variables with zero means, and γ_{vc} is a Bernoulli variable satisfying $\Pr(\gamma_{vc} = 1) = p$. In equation (1), μ represents an overall value for all the variables and types. If $\gamma_{vc} = 1$ then the corresponding variable-class combination is said to be active, and in an ideal setting its mean would be $\mu + \theta_{vc}$. If $\gamma_{vc} = 0$, then the combination is inactive and in an ideal setting its mean would be μ. No realizable setting is ideal, however, and additional variation between types, perhaps due to varying experimental conditions, is represented by the variables η_{vct}, leading to a mean $\mu + \theta_{vc} + \eta_{vct}$ for the tth type and variable-class combination (v, c). Further variability between replicates is due to measurement error, ε_{vctr}.

The lower part of Figure 1 illustrates the model. The addition of η_{vct} to the solid ideal profile $\mu + \gamma_{vc}\theta_{vc}$ $(v = 1, \ldots, V)$ corresponding to class c yields the dashed line of the realized profile

for the tth type, $\mu + \gamma_{vc}\theta_{vc} + \eta_{vct}$ $(v = 1, \ldots, V)$, which is further obscured by adding the measurement errors ε_{vctr}. If the continuous random variables are independent and Gaussian, with $\theta_{vc} \sim N(0, \sigma_\theta^2)$, $\eta_{vct} \sim N(0, \sigma_\eta^2)$ and $\varepsilon_{vctr} \sim N(0, \sigma^2)$, where $\sigma^2, \sigma_\theta^2 > 0$, $\sigma_\eta^2 \geq 0$, the γ_{vc} are independent Bernoulli variables with success probability $p \in (0, 1)$, and $-\infty < \mu < \infty$, then (1) is a variant of the classical mixed effects model in which a random component disappears if $\gamma_{vc} = 0$.

With this choice of Gaussian variables, one can readily compute the marginal density of the data for a specified variable-class combination. With a slight abuse of notation, let y with fewer indices denote a vector of measured quantities—for example, y_v denotes the data available for variable v, y_c denotes the data in class c, and y_{vc} denotes the data available for the combination of the vth variable and cth class—and let f denote a generic probability density. Then a calculation given in the Appendix shows that the joint density of y_{vc} may be written as

$$f(y_{vc}) = p f_1(y_{vc}) + (1 - p) \prod_{t=1}^{T_c} f_0(y_{vct}), \tag{2}$$

where

$$\begin{aligned} f_0(y_{vct}) = {} & (2\pi)^{-R_{ct}/2} \sigma^{1-R_{ct}} (R_{ct}\sigma_\eta^2 + \sigma^2)^{-1/2} \\ & \times \exp\left\{ -\frac{1}{2\sigma^2} \left(\sum_{r=1}^{R_{ct}} y_{vctr}^2 - R_{ct}\overline{y}_{vct}^2 \right) - \frac{(\overline{y}_{vct} - \mu)^2}{2(\sigma_\eta^2 + \sigma^2 R_{ct})} \right\}, \end{aligned}$$

in which $\overline{y}_{vct} = R_{ct}^{-1} \sum_{r=1}^{R_{ct}} y_{vctr}$, and $f_0(\cdot)$ and $f_1(\cdot)$ are the joint densities when the variable-class combination (v, c) is respectively active and inactive. The joint density $f_1(\cdot)$ when the combination is active depends on the distribution of θ_{vc}. If θ_{vc} is Gaussian, then $f_1(\cdot)$ corresponds to a $\sum_{t=1}^{T_c} R_{ct}$-dimensional multivariate Gaussian variable, with mean $\mu\mathbf{1}$, where $\mathbf{1}$ is a column vector of ones, and covariance matrix $\boldsymbol{\Sigma}$ having $\sigma^2 + \sigma_\eta^2 + \sigma_\theta^2$ on the main diagonal, and off-diagonal elements equal to $\sigma_\eta^2 + \sigma_\theta^2$ for observations from the same type and to σ_θ^2 for those from different types.

2.2. Asymmetric Laplace effects

If it is preferred to use an asymmetric density for the variable-class combinations, or a density with heavier tails, one can give the θ_{vc} an asymmetric Laplace distribution (Bhowmick *et al.*, 2006) by taking $-X_L$ or X_R with probabilities 1/2, where X_L and X_R are independent exponential random variables with rates $\sigma_{\theta_L}^{-1}, \sigma_{\theta_R}^{-1} > 0$. This yields an asymmetric distribution with median zero and variance $\sigma_{\theta_L}^2 + \sigma_{\theta_R}^2$; the usual Laplace distribution appears when $\sigma_{\theta_L} = \sigma_{\theta_R}$. In the asymmetric case the mean is non-zero, and so minor changes to the description of the model (1) are needed. Under this model the marginal density of the data for a variable-class combination with $\gamma_{vc} = 0$ is unchanged, but

$$f_1(y_{vc}) = k_0(k_L I_L + k_R I_R), \tag{3}$$

where

$$k_0 = (2\pi\sigma^2)^{-\sum_{t=1}^{T_c} R_{ct}/2}(2\pi\sigma_\eta^2)^{-T_c/2}(2\pi\sigma_\eta^2/T_c)^{1/2} \times (2\pi)^{T_c/2}|\mathbf{A}|^{-1/2}\exp\left\{-\frac{1}{2\sigma^2}\sum_{r=1}^{R_{ct}}\sum_{t=1}^{T_c} y_{vctr}^2\right\},$$

$$k_L = (2\sigma_{\theta_L})^{-1}\exp\left(\frac{\sigma_\eta^2}{2T_c\sigma_{\theta_L}^2} - \frac{\mu}{\sigma_{\theta_L}}\right), \tag{4}$$

$$I_L = \exp\left(\frac{1}{2}\mathbf{b}_L'\mathbf{A}^{-1}\mathbf{b}_L\right)\Phi\left(\frac{c_L + \mathbf{d}_L'\mathbf{A}^{-1}\mathbf{b}_L}{\sqrt{1+\mathbf{d}_L'\mathbf{A}^{-1}\mathbf{d}_L}}\right),$$

$$k_R = (2\sigma_{\theta_R})^{-1}\exp\left(\frac{\sigma_\eta^2}{2T_c\sigma_{\theta_R}^2} + \frac{\mu}{\sigma_{\theta_R}}\right), \tag{5}$$

$$I_R = \exp\left(\frac{1}{2}\mathbf{b}_R'\mathbf{A}^{-1}\mathbf{b}_R\right)\Phi\left(\frac{c_R + \mathbf{d}_R'\mathbf{A}^{-1}\mathbf{b}_R}{\sqrt{1+\mathbf{d}_R'\mathbf{A}^{-1}\mathbf{d}_R}}\right).$$

Here Φ denotes the standard Gaussian distribution function, $|\mathbf{A}|$ denotes the determinant of the $T_c \times T_c$ symmetric positive definite matrix $\mathbf{A} = (R_{ct}\sigma^{-2} + \sigma_\eta^{-2})\mathbf{I} - (T_c^{-1}\sigma_\eta^{-2})\mathbf{1}\mathbf{1}'$, where $\mathbf{I}$ is an identity matrix, $\mathbf{b}_L = \left(R_{ct}\overline{y}_{vct}\sigma^{-2} + T_c^{-1}\sigma_{\theta_L}^{-1}\right)\mathbf{1}$, $\mathbf{b}_R = \left(R_{ct}\overline{y}_{vct}\sigma^{-2} - T_c^{-1}\sigma_{\theta_R}^{-1}\right)\mathbf{1}$, $\mathbf{d}_L = -T_c^{-\frac{1}{2}}\sigma_\eta^{-1}\mathbf{1}$ and $\mathbf{d}_R = T_c^{-\frac{1}{2}}\sigma_\eta^{-1}\mathbf{1}$ are $T_c \times 1$ vectors, and $c_L = \{\mu - \sigma_\eta^2/(T_c\sigma_{\theta_L})\}/(\sigma_\eta^2/T_c)^{1/2}$ and $c_R = \{-\mu - \sigma_\eta^2/(T_c\sigma_{\theta_L})\}/(\sigma_\eta^2/T_c)^{1/2}$ are constants.

2.3. Variable selection model

The model (1) treats all variable-class combinations as independent. A natural generalization is to add indicators that determine whether each variable is active, thereby yielding

$$y_{vctr} = \mu + \delta_v\gamma_{vc}\theta_{vc} + \eta_{vct} + \varepsilon_{vctr}, \tag{6}$$
$$v = 1,\ldots,V, c = 1,\ldots,C, t = 1,\ldots,T_c, r = 1,\ldots,R_{ct},$$

where the δ_v are independent Bernoulli variables with probability q. Thus q is the proportion of active variables, and p is the proportion of active classes, given that a variable is active. Under this model the joint density of the data y_v for variable v is $f(y_v;\varphi) = qf(y_v \mid \delta_v = 1) + (1-q)f(y_v \mid \delta_v = 0)$, where $f(y_v \mid \delta_v = 1) = \prod_{c=1}^{C} f(y_{vc})$ is defined in (2) and $f(y_v \mid \delta_v = 0) = \prod_{c=1}^{C}\prod_{t=1}^{T_c} f_0(y_{vct})$. A similar model can be constructed using an asymmetric Laplace density for the θ_{vc}.

3. INFERENCE

3.1. Introduction

Ready computation of the marginal density of the data, (2), has useful consequences; for example, we shall see in Section 3.3 that it allows a fast algorithm for agglomerative clustering. Moreover, the parameters $\varphi = (\mu, \sigma^2, \sigma_\eta^2, \sigma_\theta^2, p)$ of the prior density may be estimated by maxi-

mizing the log likelihood

$$\ell(\varphi) = \sum_{v=1}^{V} \sum_{c=1}^{C} \log f(y_{vc}; \varphi). \tag{7}$$

In the unreplicated case, $R_{ct} \equiv 1$, only $\sigma^2 + \sigma_\eta^2$ is estimable, but by fixing $\sigma_\eta^2 = 0$ we can estimate the other parameters. Maximum likelihood estimation can likewise be performed for the parameters of the other models, based on (3) and on the corresponding marginal densities for those in §2.3.

Such estimators are not robust to all features of real data, so are best regarded as indicating a likely range for the parameters, rather than as optimal values for use at all costs. In particular, the data may contain little information about the probabilities p and, if present, q, and in practice we find it helpful to tune the model by varying p and q while estimating the other parameters.

Bayes factors provide insight when assessing the importance of variables and variable-class combinations. Under model (6), the Bayes factor for variable δ_v is defined as $B_v = f(y_v \mid \delta_v = 1)/f(y_v \mid \delta_v = 0)$, while that for γ_{vc} is $B_{vc} = f(y_{vc} \mid \delta_v = 1, \gamma_{vc} = 1)/f(y_{vc} \mid \delta_v = 1, \gamma_{vc} = 0)$.

3.2. Classification

Suppose that data $y_1, \ldots, y_C$ are available from C distinct classes, with y_c representing the T_c types known to belong to class c, and that a new and independent dataset y^* must be classified to one of these classes, or declared to be of a previously unseen class. We define the multinomial variable U taking values in $\{1, \ldots, C, C+1\}$, where $U = u$ will denote that y^* should be classified to the class u, and class $C+1$ allows y^* to arise from an as-yet unobserved class. The probability density for y_c under (6) is

$$f(y_c; \varphi) = \prod_{v=1}^{V} \left[q \left\{ p f_1(y_{vc}) + (1-p) \prod_{t=1}^{T_c} f_0(y_{vct}) \right\} + (1-q) \prod_{t=1}^{T_c} f_0(y_{vct}) \right], \quad c = 1, \ldots, C,$$

where φ represents the vector of parameters, and if $U = c \in \{1, \ldots, C\}$, then

$$f(y^*, y_c; \varphi) = \prod_{v=1}^{V} \; q \left\{ p f_1(y_v^*, y_{vc}) + (1-p) f_0(y_v^*) \prod_{t=1}^{T_c} f_0(y_{vct}) \right\} + (1-q) f_0(y_v^*) \prod_{t=1}^{T_c} f_0(y_{vct}),$$

in which $f_1(y_v^*, y_{vc})$ is the joint density of y_{vc} and the data on the vth variable from the unknown type, y_v^*, treated as a single group of observations with the same θ_{vc} but a potentially different η_{vct}. If $U = C+1$, then we may write formally

$$f(y^*, y_{C+1}; \varphi) \equiv f(y^*; \varphi) = \prod_{v=1}^{V} \left[q \left\{ p f_1(y_v^*) + (1-p) f_0(y_v^*) \right\} + (1-q) f_0(y_v^*) \right].$$

As the type data are independent conditional on the model parameters, we have

$$\Pr(U = u \mid y^*, y_1, \ldots y_C; \varphi) = \frac{\Pr(U = u) f(y^*, y_u; \varphi) \prod_{c \neq u} f(y_c; \varphi)}{\sum_{c'=1}^{C+1} \Pr(U = c') f(y^*, y_{c'}; \varphi) \prod_{c \neq c'} f(y_c; \varphi)},$$

for $u \in \{1, \ldots, C, C+1\}$, thus yielding the posterior classification for y^*.

When the probability $p = 0$, no variable-class combination is active and every type is independent a posteriori; then the prior and posterior distributions for U are the same. Similarly, when $q = 0$ no variable-class combination is allowed and types are independent a posteriori. In our experience, higher values of p for a fixed positive q lead to more certain classifications, whereas small values of p lead to just a few active variable-class combinations and reduce over-fitting.

3.3. Agglomerative hierarchical clustering

Agglomerative hierarchical clustering is the sequential partitioning of types, initially assuming that all types are separate clusters, and then successively merging the two closest types until finally there is a single cluster. This requires a metric to measure the distance between two clusters, which a probability model provides through the change in posterior when they are merged. However a prior distribution on clusterings is required for Bayesian inference on them.

Consider a partition $\mathcal{C}$ of T types partitioned into $|\mathcal{C}| = C \in \{1, \ldots, T\}$ blocks, with T_1 types in cluster 1, T_2 types in cluster 2, and so forth. We assume prior exchangeability in the grouping of types, and hence need only specify a prior for the number of blocks in the partition and for their sizes. Heard *et al.* (2006) suggest a uniform discrete prior $\Pr(C) = 1/T$ $(C = 1, \ldots, T)$, for the number of distinct clusters of the partition, and the uniform multinomial-Dirichlet prior for the cluster sizes $T_1, \ldots, T_C$ given C, thereby yielding

$$\Pr(\mathcal{C}) \propto \frac{(C-1)!T_1! \ldots T_C!}{T(T+C-1)!}, \quad |\mathcal{C}| = C, \sum_{c=1}^{C} T_c = T. \tag{8}$$

Although equation (8) allows empty clusters to appear, this causes no difficulties for hierarchical clustering because dropping empty clusters always makes a partition more probable.

In our algorithm, every type is initially regarded as a separate cluster, so the initial partition has T blocks, each with one type. At each step every possible merger of pairs of blocks is considered, and the merger that maximizes the posterior probability of the resulting partition $\mathcal{C}'$ is applied. Suppose that the current partition is $\mathcal{C}$, and that the data for types comprising its $T + 1 - C$ blocks are denoted $\mathcal{Y}_1, \ldots, \mathcal{Y}_{T+1-C}$, containing $T_1, \ldots, T_C$ types respectively. If a proposed partition $\mathcal{C}'$ merges blocks $\mathcal{Y}_i$ and $\mathcal{Y}_j$ of $\mathcal{C}$ to form a new block whose data are denoted $\mathcal{Y}_{ij}$, then since the only change between $\mathcal{C}$ and $\mathcal{C}'$ concerns $\mathcal{Y}_i$ and $\mathcal{Y}_j$, the ratio of posterior probabilities for $\mathcal{C}$ and $\mathcal{C}'$ is

$$\frac{\Pr(\mathcal{C}')}{\Pr(\mathcal{C})} \frac{\prod_{c \in \mathcal{C}'} f(\mathcal{Y}_c; \varphi)}{\prod_{c \in \mathcal{C}} f(\mathcal{Y}_c; \varphi)} = \frac{(T+C-1)(T_i+T_j)!}{(C-1)T_i!T_j!} \frac{f(\mathcal{Y}_{ij}; \varphi)}{f(\mathcal{Y}_i; \varphi) f(\mathcal{Y}_j; \varphi)}, \tag{9}$$

where $f(\mathcal{Y}_i; \varphi)$ denotes the marginal density of the data for the types in block $\mathcal{Y}_i$. The new partition $\mathcal{C}'$ is chosen to maximize (9) over all possible pairs of blocks of $\mathcal{C}$.

When p or q equals zero, the posterior probability of any partition $\mathcal{C}$ equals the prior probability (8), and the most probable prior clustering is a single cluster containing all the types. If $pq > 0$, then usually the largest ratio (9) at each step of the algorithm exceeds unity up to a certain number of mergers, and then takes values less than unity. The log ratio provides a natural scale for comparison of different partitions, and so provides lengths for the arms of the dendrogram

corresponding to the successive partitions chosen by the algorithm. A monotone height function is needed to draw a dendrogram and can be obtained using a signed difference of marginal log posteriors. Let $\hat{\ell}_{\mathcal{C}}$ denote the log marginal posterior for partition $\mathcal{C}$, and let $\hat{\ell}'_{\mathcal{C}} = \hat{\ell}_{\mathcal{C}} - \max_{\mathcal{C}} \hat{\ell}_{\mathcal{C}}$; thus $\hat{\ell}'_{\hat{\mathcal{C}}} = 0$ for the optimal clustering $\hat{\mathcal{C}}$ found by the agglomerative algorithm. When drawing a dendrogram we take the length between two successive partitions $\mathcal{C} \subset \mathcal{C}'$ to be $|\hat{\ell}'_{\mathcal{C}} - \hat{\ell}'_{\mathcal{C}'}|$.

We perform parameter estimation for clustering at the first step of the algorithm, assuming that every type is a different cluster, and leaving the estimates unchanged during the agglomeration. Apart from the one-time optimization needed to provide initial estimates of φ before starting the hierarchical clustering, the number of computations is of order $O(VT^3)$. In many applications $V \gg T$, so the algorithm is rapid; see §§5, 6.

Like many other procedures for hierarchical clustering, our approach does not attempt an exhaustive search of all possible dendrograms. This is a disadvantage relative to more complex procedures, but if desired, the stability of the resulting optimal cluster may be explored using Markov chain Monte Carlo methods, such as the spilt-merge algorithm of Booth *et al.* (2008) or the delayed sampling algorithm of Green and Mira (2001).

4. DATA EXAMPLES

4.1. Microarray data

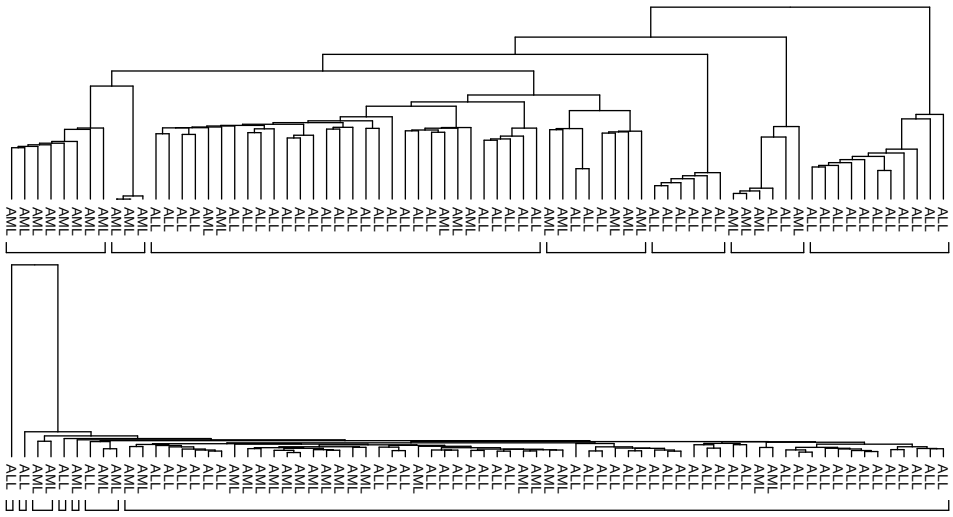
We first apply our approach to the Golub *et al.* (1999) leukaemia data. Patients had either acute lymphoblastic leukaemia (ALL) or acute myeloid leukaemia (AML). Affymetrix arrays were used to collect measurements for 7129 genes over 47 ALL tissues and 25 AML tissues. The data were processed in several stages to remove apparently aberrant values, finally giving 2030 log-expression ratios that are available in the supplementary materials of McNicholas and Murphy (2010) and have been analysed several times (e.g., Dudoit *et al.*, 2002; McLachlan *et al.*, 2002; Kim *et al.*, 2006) as a benchmark for classification and clustering.

In order to compare our approach with other recent proposals, we applied the `R` package `sparcl`, which embodies the framework for sparse feature selection in clustering proposed by Witten and Tibshirani (2010), under which an objective function that uses the elements of a weighted dissimilarity matrix is optimized under simultaneous L_1 and L_2 constraints on the weights. Using a sufficiently constrained L_1 penalty will provide variable selection by setting certain of the weights to zero; variables with large weights contribute strongly to the clustering. This approach does not incorporate an automated way to cut the dendrogram, so the number of clusters must be set manually.

Kim *et al.* (2006)'s Bayesian analysis using Dirichlet mixtures suggest a clustering of the Golub data into seven groups. In order that our model for these unreplicated data be identifiable, we set $\sigma^2_\eta = 0$, and estimated the other parameters by maximizing the log likelihood (7), giving $\hat{\sigma}^2 = 0.87\ (0.01), \hat{\sigma}^2_\theta = 3.31\ (1.45), \hat{\mu} = -0.01\ (0.00), \hat{p} = 0.04\ (0.03), \hat{q} = 0.93\ (0.11)$. Since this value of q is high, and the profile log likelihood shows that it is poorly-determined, we experimented by varying q and estimating the other parameters; the results below are with $q = 0.06$, which gives $\hat{\sigma}^2 = 0.93$, $\hat{\sigma}^2_\theta = 8.14$, $\hat{\mu} = -0.01$ and $\hat{p} = 0.04$. The top panel of Figure 2 shows the hierarchical tree built using our method with these parameter values, and the bottom panel shows the hierarchical tree built using `sparcl` with its default options. In order to compare the methods, both trees are cut at seven groups. With this choice, `sparcl` gives one very large cluster containing a mixture of ALL and AML tissues.

The Rand index (Rand, 1971; Lau and Green, 2007) can be used to compare clustering performance. Suppose the data grouping is coded in the label vector $\mathbf{d} = (d_1, \ldots, d_T)$, whose elements are integers $1, 2, \ldots$. Data belonging to same cluster have the same integer in $\mathbf{d}$, so the number of groups $C = \max(\mathbf{d})$. If the labels $d_1, \ldots, d_T$ allocating types to clusters are estimated by

FIGURE 2: Dendrograms for the Golub data using our method (top panel) and `sparcl` (bottom panel). Both dendrograms are cut at seven groups.



$\hat{d}_1, \ldots, \hat{d}_T$, then the Rand index may be written as

$$\text{RI} = \frac{2}{T(T-1)} \sum_{t=2}^{T} \sum_{t'<t} \left(\mathbb{I}_{\{d_t = d_{t'}, \hat{d}_t \neq \hat{d}_{t'}\}} + \mathbb{I}_{\{d_t \neq d_{t'}, \hat{d}_t = \hat{d}_{t'}\}} \right), \tag{10}$$

where $\mathbb{I}_{\{\cdot\}}$ denotes the indicator function; small values of RI are preferable. This index was 0.43 for our method, and was 0.45 for `sparcl`, confirming that our method clusters the data somewhat better, with this number of groups.

Figure 3 shows the dendrogram for our method, but for the maximum a posteriori height. This has 25 groups, just three of which contain both AML and ALL tissue types. This is a large number of clusters, but the image plot of the gene expression data shows the presence of clear sub-groups within the ALL and AML tissue types, which are found by our method. The corresponding Rand index equals 0.505, but this is considerably better than a chance finding, as shown in Figure 4, which compares the values of RI when the dendrograms in Figure 2 are cut at heights yielding from 2 to 25 clusters. Our approach produces lower values of RI for a wide range of cluster sizes, and when compared to results from data in which the sample labels were permuted, it does much better than chance over most of the range, unlike `sparcl`.

We also applied `HDclassif` (Bergé *et al.*, 2012), a recent partitioning method that clusters high-dimensional data through projection in a subspace, but does not implement variable selection or provide a dendrogram. With its default settings `HDclassif` found a single cluster containing all the tissues.

4.2. Metabolomic data

These data consist of 14 profiles each comprising 43 metabolites, all but one profile being replicated four times; the overall number of profiles is 55. The goal of the study from which the data were drawn was to use gas chromatography-mass spectrometry spectra for different plant

FIGURE 3: Analysis of Golub data. Left: dendrogram for clustering using our method (left side), cut at the maximum a posteriori point (dashed grey line), which gives the 25 clusters shown at the right. The image plot of the log expression values (center) is shown for the 305 genes with $\log B_v > 5$.

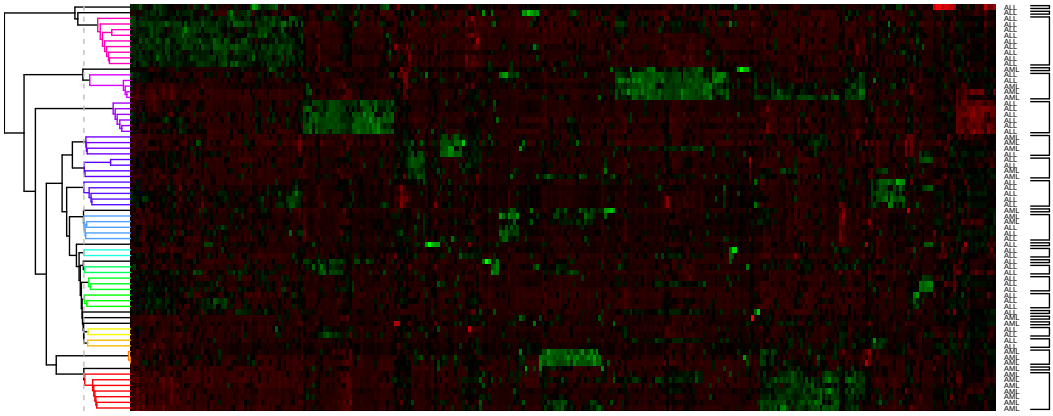
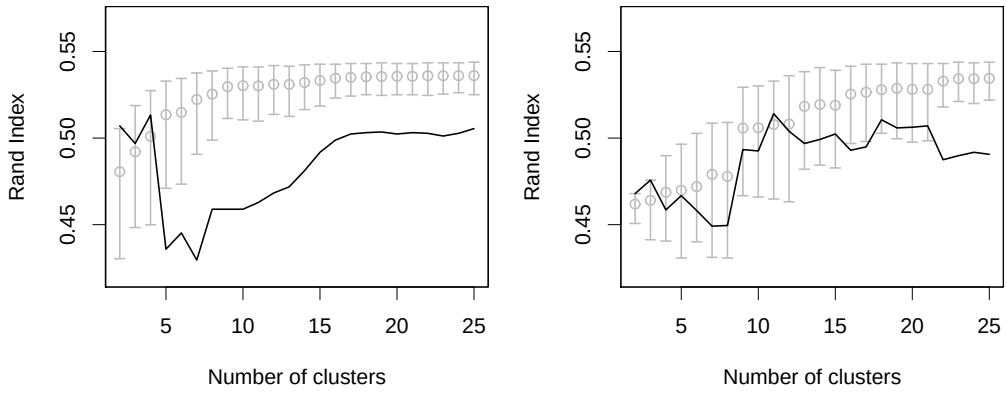


FIGURE 4: Rand index RI for clusterings for the Golub data found using our method (left) and `sparcl` (right). The black lines show how RI depends on the cuts in the dendrograms in Figure 2. The grey circles and whiskers show the average and the upper and lower 2.5% quantiles for RI for 1000 datasets in which the sample labels were randomly permuted.



phenotypes to classify forward genetic mutants of *Arabidopsis thaliana*. The data involve three wildtype plants, *WsWT*, *RLDWT*, and *ColWT*; two mutants defective in starch biosynthesis, *pgm* and *isa2*; four mutants defective in starch degradation, *sex1*, *sex4*, *mex1*, and *dpe2*; a mutant that accumulates starch as a pleiotropic effect, *tpt*; and four unknown mutants, *d172*, *d263*, *ke103* and *sex3*. There are only three replicates of the wildtype *ColWT*. The idea was to use the classification of the four unknown mutants to indicate what avenues should be explored first when seeking to characterize them, and which metabolites are important for this task. The raw data were first preprocessed and 43 reliably detected metabolites were selected from the many available; then the data were rescaled to allow for experimental variation between different runs, as assessed by the inclusion of the same wild types in each run; see Messerli *et al.* (2007). The data analysed below are the log profiles.

	WsWT	*RLDWT*	*tpt*	*pgm*	*sex4*	*mex1*	*dpe2*	New
ColWT	34.04	29.35	20.69	0.44	6.57	0	0	8.92
isa2	0.28	0.14	0.02	5.36	63.72	0	0	30.48
sex1	2.53	0.53	1.97	93.04	1.72	0	0	0.21
d172	0.44	1.06	0.17	0.88	97.33	0	0	0.13
d263	1.56	3.88	0.73	5.12	87.88	0	0	0.83
ke103	25.41	6	44.86	20.31	1.57	0	0	1.84
sex3	16.24	64.7	7.96	0.02	2.08	0	0	9.00

TABLE 1: Posterior classification percentages for the asymmetric Laplace variable selection model, assuming a uniform classification prior. The maximum a posteriori percentages are boxed.

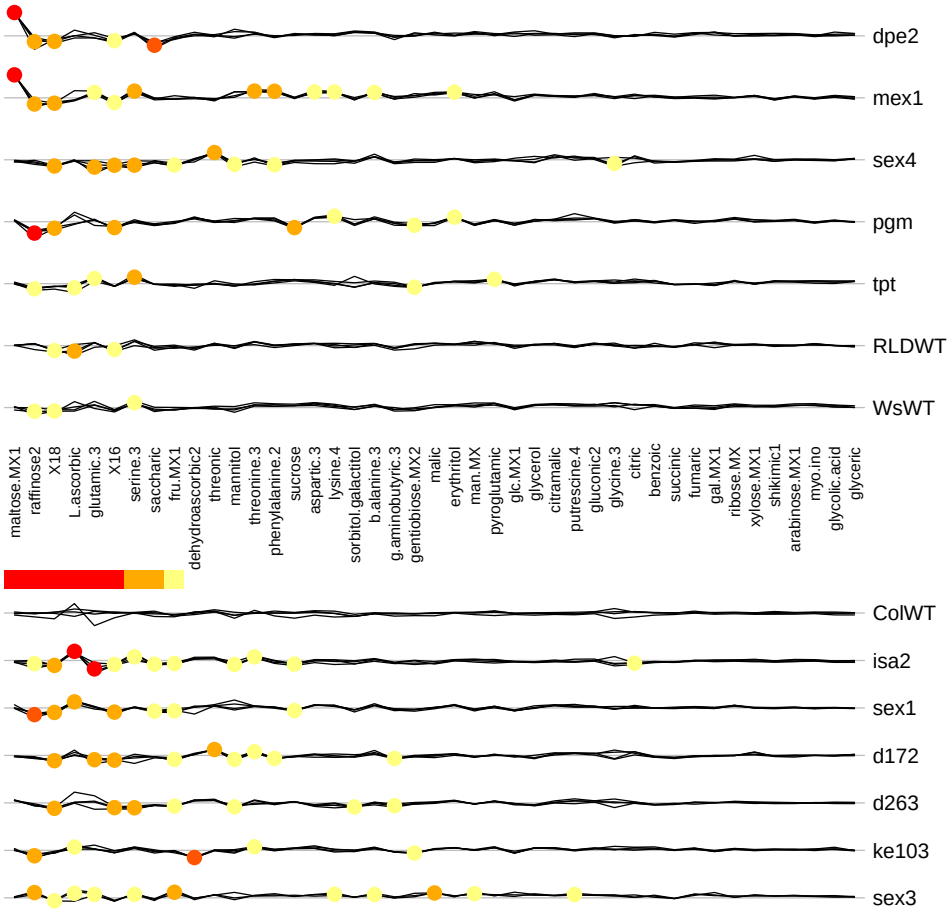
The parameter estimates and standard errors for the Gaussian model without variable selection obtained by maximizing (7), were $\hat{\mu} = 0.083$ (0.028), $\hat{\sigma}^2 = 0.159$ (0.005), $\hat{\sigma}^2_\eta = 0.373$ (0.032), $\hat{\sigma}^2_\theta = 5.155$ (2.773) and $\hat{p} = 0.034$ (0.019). However the standard error for p is of doubtful value, because the profile log likelihood is not quadratic: the 95% confidence interval for p based on the profile likelihood, $(0.015, 0.064)$, is very different from what would be obtained from a normal approximation to the distribution of $\hat{p}$. The estimates and standard errors for the corresponding asymmetric Laplace model were $\hat{\mu} = 0.085$ (0.028), $\hat{\sigma}^2 = 0.159$ (0.005), $\hat{\sigma}^2_\eta = 0.350$ (0.043), $\hat{\sigma}^2_{\theta_L} = 0.983$ (0.778), $\hat{\sigma}^2_{\theta_R} = 1.547$ (2.361) and $\hat{p} = 0.078$ (0.071), with 95% profile likelihood confidence interval $(0.038, 0.134)$ for p; similar comments apply as for the Gaussian model. The asymmetric Laplace model gives a smaller variance, $\sigma^2_\theta = \sigma^2_{\theta_L} + \sigma^2_{\theta_R}$, and a larger p, than the Gaussian model. The maximized values of the log likelihood for the Gaussian, symmetric Laplace and asymmetric Laplace models are $\hat{\ell}_{\mathrm{G}} = -1938.98$, $\hat{\ell}_{\mathrm{SL}} = -1938.24$ and $\hat{\ell}_{\mathrm{AL}} = -1938.11$, respectively: as judged by AIC, the symmetric Laplace model is best, but the differences are so small that the Gaussian model could also be chosen.

In order to find those metabolites important in classifying the different plants, we apply the Laplace variable selection model. To obtain results readily comparable with those for the models without variable selection, we fixed σ^2, σ^2_η, $\sigma^2_{\theta_L}$, $\sigma^2_{\theta_R}$ and μ, which have the same interpretations in both models, to the estimates above, and with these fixed we found $\hat{p} = 0.83$ and $\hat{q} = 0.183$. Figure 5 shows the data, with metabolites sorted according to the Bayes factors B_v computed using the model (6). Six metabolites have $\log B_v > 5$, namely *maltose.MX1*, *raffinose2*, *X18*, *L.ascorbic*, *glumatic.3*, and *X16*. A similar result is obtained using the Gaussian model.

Table 1 shows the posterior classification percentages using the Laplace variable selection model, applied both to the unknown types and to some known ones used as references. The classification probability for the reference wild type *ColWT* is spread between the other wild types *WsWT* and *RLDWT*, and also *tpt*; classification of *ColWT* to *tpt* is not surprising, because Figure 5 shows that the wild types and *tpt* have similar profiles. The reference plant *isa2* defective in starch biosynthesis is close to *sex4* but may also be a previously-unobserved class. The reference *sex1* defective in starch degradation is classified to *pgm* with high probability, and likewise the unknown *d172* and *d263* are classified with *sex4*. The unknown mutant *ke103* is classified to *tpt*, but may also be from *pgm* or *WsWT*. The unknown mutant *sex3* is closest to *RLDWT*.

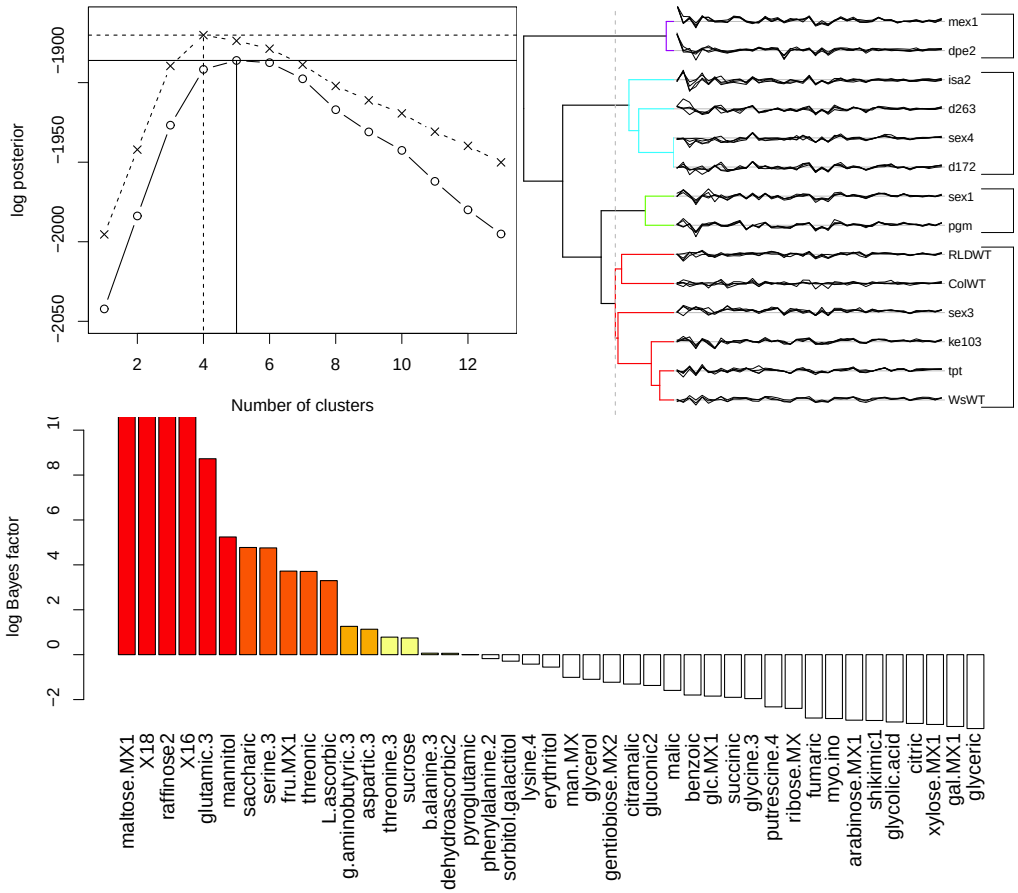
The right panel of Figure 6 displays the dendrogram produced using the the asymmetric Laplace model with the data. The vertical line cutting the dendrogram shows the optimal partition into four components. The Gaussian model yields a similar dendrogram, but with five components. Both models successfully clustered wild types differently from the known mutants *pgm*, *isa2*, *sex1*, *sex4*, *dpe2* and *mex1*. The log posterior plot in the top left panel of Figure 6

FIGURE 5: Metabolomic profiles, with variable and variable-class Bayes factors for the asymmetric Laplace variable selection model. The profiles in the upper part of the figure correspond to known types, and those in the lower part to the classifying types. In log Bayes factor calculation for classification, every known class is treated as a separate cluster. The metabolites, given in the middle of the figure, are sorted from left to right according to the Bayes factors B_v, shown by the horizontal heat bar. The scale for the Bayes factors is: clearly important (red, $\log B_v > 5$), important (dark orange, $3 < \log B_v \leq 5$), somewhat important (light orange, $1 < \log B_v \leq 3$), and negligible (yellow, $0 < \log B_v \leq 1$); negative values of $\log B_v$ are not shown. Blobs correspond to the Bayes factor B_{vc} and use the same heat scale.



confirms that for the Gaussian and the asymmetric model the marginal posteriors for the four-, five- and six-cluster groupings are close, whereas the asymmetric Laplace model is superior. The bottom panel shows the log Bayes factors B_v of the asymmetric Laplace model. Clustering uses the grouping with the highest posterior is used to compute the log Bayes factors, and yields more variables with $\log B_v > 0$. The variables with $\log B_v > 5$ are the same as those in the classification reported in Figure 5, except for *mannitol* which was unimportant ($\log B_v < 0$) in classification, but appears important ($\log B_v > 5$) for clustering.

FIGURE 6: Agglomerative clustering for the metabolomic data. Top left panel: Log marginal posterior as a function of the number of clusters when performing agglomerative clustering, for the Gaussian (circle) and asymmetric Laplace (crosses) models. The best clusterings found are shown by the vertical lines, and the values of the marginal log posterior at the optimal point are shown by the horizontal lines. Top right panel: Data, with dendrogram obtained using the asymmetric Laplace model, with the optimal clustering given by the vertical line cutting the dendrogram, and by the grouping at the right of the panel. Bottom panel: Log Bayes factor B_v for metabolites calculated using the asymmetric Laplace clustering of the top right panel, for the color scale of the bar chart see caption to Figure 5.



5. SIMULATION STUDIES

5.1. First comparison

Our first set of simulation experiments compares our approach with `sparcl` (Witten and Tibshirani, 2010) and `HDclassif` (Bergé *et al.*, 2012), for data generated from our assumed model. `sparcl` is a hierarchical method for clustering high-dimensional data, and like our technique it produces a dendrogram and selects variables. `HDclassif` performs partitioning through projection into a subspace; it neither implements variable selection nor provides a dendrogram.

Our procedures are fully automatic, but neither `sparcl` nor `HDclassif` provides an estimated number of clusters. We therefore used the true number of clusters to provide a grouping estimation of simulated data using `sparcl` and `HDclassif`, and used a variant of our procedure with the true number of clusters; even if this number is known, the estimated groupings may be incorrect. Since `sparcl` and `HDclassif` do not handle replicated data, we treat the data as unreplicated for all the methods. We only describe results for our Gaussian model, as those for the asymmetric Laplace model were similar.

We generated datasets under two configurations suggested by §4.2, with data simulated from our Gaussian model using $V = 50$ variables measured on $T = 10$ types, each replicated $R_{ct} = 4$ times. The number of clusters was chosen uniformly in the range $2, \ldots, 5$ and types were allocated using the uniform multinomial-Dirichlet law. The model parameters were $\mu = 0$ and $p = q = 0.5$. The variables in (6) had Gaussian distributions with $\sigma^2 = 1$, $\sigma_\eta^2 = 4$, once with small clustering signal $\sigma_\theta^2 = 9$, and once with large clustering signal $\sigma_\theta^2 = 36$.

We simulated 1000 datasets, and estimated the parameters of the Gaussian model for each of them. We also used oracle versions of our model. In one version we fixed the parameters to their true values and cut the dendrogram at the correct number of clusters $\tilde{G}_+$; even if the true number of clusters is used, the corresponding clustering may be incorrect.

Another variant of our method, denoted $\tilde{G}$, assumes known parameters but estimates the number of clusters. Comparison of $\tilde{G}$ and $\tilde{G}_+$ shows how much the quality of clustering changes when the number of clusters is unknown. In another version, we considered that there is knowledge about the proportion of effective variables, q, which we set to its true value 0.5 while estimating the other parameters. This gives us two further variants of our procedure, one with the correct number of clusters inserted in the procedure, $\check{G}_+$, and one with the number of clusters estimated by the maximum a posteriori value, $\check{G}$. Comparing $\tilde{G}$ with $\check{G}$ and $\tilde{G}_+$ with $\check{G}_+$ shows how much clustering improves when the proportion of effective variables is known. We also included the fully automatic version of our proposed method, denoted by G, which estimates both the parameters and the number of clusters. Comparing G with $\tilde{G}$, and G_+ with $\tilde{G}_+$, shows the effect of parameter estimation on our procedure.

As a benchmark we also report agglomerative clustering with average linkage over the activated variables, with its tree cut at the correct number of clusters, E_+^*. Comparison of E_+^* with the other techniques subscripted by $+$ shows how knowing the true active variables can improve clustering accuracy.

In the tables we use the subscript $+$ if the number of clusters is set to its true value, a superscript $*$ if the procedure is implemented using only the true active variables, $\tilde{}$ if the parameters for our procedures are set to their true values, and $\check{}$ for our procedures with parameter q set to its true value but the others estimated. We use the Rand index (10) as a clustering loss function, and compare variable selection properties using the false positive and false negative rates, and their sum, the total error. Note that E_+^* uses the true variables and `HDclassif` does not implement variable selection.

The empirical losses are shown in Table 2. None of the techniques beats E_+^* in terms of clustering loss, since E_+^* knows the true clustering variables but the others do not. Clustering performance depends heavily on knowledge of the clustering variables. If the number of clusters (but not their composition) is known, then estimating all the parameters increases the loss for our method from 0.35 to 0.43 for $\sigma_\theta^2 = 36$, whereas not knowing the number of clusters increases the loss from 0.35 to 16.48. A similar pattern is observed for small signal to noise ratio, $\sigma_\theta^2 = 9$. Thus, the major increase comes from estimating the number of clusters rather than from estimating the parameters. Fixing parameters to their true values, as for $\tilde{G}$ and $\check{G}$, helps in finding the correct grouping, but is less efficient for finding the correct clustering variables. `sparcl`, implemented as the oracular S_+ which knows the true number of clusters, has lower losses than

TABLE 2: Empirical clustering loss and false positive and false negative probabilities ($\times 100$) for data generated under the Gaussian model. We use a subscript $+$ if the number of clusters is set to its true value, a superscript $*$ if the procedure is implemented using only the true active variables, $\tilde{}$ if the parameters for our procedures are set to their true values, and $\check{}$ if the proportion of active variables q is set to its true value. The clustering procedures used are: as a benchmark, agglomerative clustering using Euclidean distance on the true active variables and the dendrogram cut at the true number of clusters (E^*_+); variants of our Gaussian method ($\tilde{G}_+$, $\tilde{G}$, $\check{G}_+$, $\check{G}$, G_+ and G), `sparcl` (S_+) and `HDclassif` (H_+) with their default settings and the number of clusters set to the true value. The largest standard errors for the differences of two clustering losses, two false positive and two false negative sums are 0.06, 0.08 and 0.08, respectively.

		Procedure								
		E^*_+	$\tilde{G}_+$	$\tilde{G}$	$\check{G}_+$	$\check{G}$	G_+	G	S_+	H_+
$\sigma^2_\theta = 36$	Clustering Loss	0.04	0.35	16.48	0.43	15.83	0.43	15.96	0.92	1.90
	False Positive (%)	—	68	68	53	53	47	47	45	—
	False Negative (%)	—	8	8	13	13	15	15	27	—
	Sum	—	76	76	66	66	62	62	72	—
$\sigma^2_\theta = 9$	Clustering Loss	4.71	12.89	22.28	11.94	21.18	13.07	21.51	9.20	122.5
	False Positive (%)	—	85	85	82	82	68	68	33	—
	False Negative(%)	—	6	6	7	7	13	13	46	—
	Sum	—	91	91	89	89	81	81	79	—

does G; the comparable oracle method G_+ based on our model has a lower loss for $\sigma^2_\theta = 36$ and a similar loss for $\sigma^2_\theta = 9$. `HDclassif` is less efficient than S_+ and G_+ even though it is applied as H_+, with the true number of clusters known.

False negative and positive percentages averaged over the 1000 simulated datasets show that our method works better if the signal is considerable, $\sigma^2_\theta = 36$, while `sparcl` has larger false negative rates than the variants of our procedure. This perhaps explains why S_+ can have smaller clustering loss than G_+: our procedure does not perform hard selection of variables, whereas `sparcl` guards against including noise variables, and therefore has a larger false negative rate. Table 3 shows that a good clustering method may be poor at selecting active variables. It appears that clustering with a small subset of clustering variables, i.e., clustering with a large false negative selection, can be more efficient than a method that weights clustering variables more appropriately.

5.2. Second comparison

For a second comparison, we consider high-dimensional data with just a few important variables. Data are simulated under the setting of Kim *et al.* (2006), in which a latent binary vector indicates the discriminating variables and a Dirichlet process mixture defines the cluster structure. These variables and the clusters are sought using a Metropolis–Hastings algorithm involving split-merge moves, and whose performance depends on parameters that must be specified by the user. These authors tested their algorithm using simulated Gaussian data with 15 independent profiles comprising 1000 variables. The profiles are split into four clusters by only 20 of these variables, through the expression

$$y_{vt} \sim \mathbb{I}_{\{1\le t\le 4\}} N(\mu_1, \sigma^2_1) + \mathbb{I}_{\{5\le t\le 7\}} N(\mu_2, \sigma^2_2) + \mathbb{I}_{\{8\le t\le 13\}} N(\mu_3, \sigma^2_3) + \mathbb{I}_{\{14\le t\le 15\}} N(\mu_4, \sigma^2_4), \quad (11)$$

for $v = 1, \ldots, 20$, where $\mathbb{I}_{\{\cdot\}}$ denotes the indicator function. The means and variances μ_i and σ_i^2 are chosen uniformly from $[-5, 5]$ and $[0.01, 1]$, respectively. The remaining noise variables y_{vt} $(v = 21, \ldots, 1000; t = 1, \ldots, 15)$ are generated as independent standard normal variates.

After $100,000$ iterations of their Metropolis–Hastings algorithm, the last $60,000$ of which were used for inference, Kim *et al.* (2006) were able to identify the correct clustering and 17 of the variables that led to it, though this latter number varied somewhat with the parameter values used. Tadesse *et al.* (2005) constructed a reversible jump algorithm for a related model and tested it on several datasets close to that of Kim *et al.* (2006), showing rather better ability to identify all 20 active variables.

We generated 1000 datasets from this model, and applied different variants of our Gaussian procedure. As the data are not simulated from our model there are no 'correct' parameter values, so we both used the 'vanilla' version of our procedure, which estimates the parameters by maximum likelihood, and also tried tuning it by choosing the parameters, as we now explain. The parameter σ^2 is the variance of data in noise clusters, so we set $\sigma^2 = 1$. The mean of clusters is generated according to $U[-5, 5]$, and we set $\sigma_\theta^2 = 8.3$ equal to the variance of the $U[-5, 5]$ distribution. The experimental noise variance is set as $\sigma_\eta^2 = 0$ to make the model identifiable for these unreplicated data. The overall mean is set to to $\mu = 0$, and the proportion of active variables is set to $q = 0.02 = 20/1000$. The proportion of the active cluster-type combinations for active variables $p \approx 1$, since all clusters centers differ from μ for activated variables. We attempted to compare the performance of our procedures with `sparcl` and with `HDclassif`, but the latter frequently stopped our simulations due to convergence problems, so finally we had to exclude it. It appears that `HDclassif` is difficult to apply when the proportion of clustering variables is tiny, as it is in this simulation.

Table 3 shows that the correct cluster structures for all 1000 simulated datasets are found using our method when the parameters are manually tuned, and that the oracle version of `sparcl`, with the correct number of clusters known, was able to recover the clustering structure in all cases. When the parameters are estimated, however, our method G may not find the correct clustering, though it finds the correct variables. The table shows that inserting more information about the parameters of our model improves the clustering, but can undermine variable selection.

All variants of our procedures, except those with manually tuned parameters, selected the correct 20 active variables, yielding 0% false positives and 0% false negatives, whereas `sparcl` may incorrectly drop active variables. It thus appears that our very simple approach is at least competitive with that of Kim *et al.* (2006) in terms of accuracy and variable selection and much faster and more straightforward to use in practice. If there is no information about the number of clusters, our method is applicable, but `sparcl` is not.

The time for a single clustering using our approach on a laptop was 5s, of which around 4s were needed for the parameter estimation, and the only parameter that had to be fixed, $\sigma_\eta^2 = 0$, was required for the model to be identifiable.

6. DISCUSSION

Our approach entails Bayesian variable selection (George and McCulloch, 1997) adapted for classification and clustering, with covariates taken to be independent variates, and is related to contributions of Heard *et al.* (2006) and Tadesse *et al.* (2005). It may be adapted for any Bayesian model with a closed form marginal density and provides supervised, semi-supervised, and unsupervised clustering combined with variable selection. Simulations suggest that it has a similar performance to the sophisticated Bayesian model proposed by Kim *et al.* (2006), despite not requiring the user to run a Markov chain algorithm. Although direct comparison with recently proposed methods like `sparcl` is not possible, since such methods typically do not provide an

TABLE 3: Empirical clustering loss and false positive and false negative probabilities ($\times 100$) for data generated with 1000 variables and 15 clustering types using (11). V variables are active, and values for the others are simulated independently from a standard normal distribution. The largest standard error is 0.2 for the difference of two clustering losses and that for the difference of the two false positive or false negative sums is 0.28; see the caption to Table 2 for more details, though here $\sim$ indicates that the parameters have been tuned as explained in the text, rather than set to their correct values.

		Procedure							
		E_+^*	$\tilde{G}_+$	$\tilde{G}$	$\check{G}_+$	$\check{G}$	G_+	G	S_+
$V = 20$	Clustering Loss	0.00	0.00	0.00	0.25	9.71	0.25	9.71	0.00
	False Positive (%)	—	3	3	0	0	0	0	0
	False Negative (%)	—	0	0	0	0	0	0	26
	Sum	—	3	3	0	0	0	0	26
$V = 10$	Clustering Loss	0.00	0.00	0.01	2.94	10.86	2.94	10.86	0.00
	False Positive (%)	—	3	3	0	0	0	0	1
	False Negative (%)	—	0	0	0	0	0	0	8
	Sum	—	3	3	0	0	0	0	9

automatic way to choose the number of groups, oracle versions seem to be comparable, with our method typically having a slight edge, and it provides automatic clustering. Our method is much faster: clustering of the metabolite data took 0.1s and parameter estimation took about 0.06s on an ordinary laptop. Its computational complexity can be reduced from $O(VT^3)$ to $O(VT^2 \log T)$, where V and T are the numbers of variables and of types, if the marginal posterior has the Lance–Williams (1967) property, but this seems worthwhile only if $T > 100$.

Our approach ignores correlations between variables. At first sight this seems unwise, but previous authors have also found that ignoring correlations in high-dimensional data can yield good classifiers. Empirical experience supporting this is described by Hand (2006) and Hand and Yu (2001), and some theoretical explanation is provided by Bickel and Levina (2004) and Hall *et al.* (2005). Work not reported here supports this: we found that our procedure performs reasonably well on simulated data with correlated variables, and that if cluster centres are sufficiently separated, correlation has little effect on clustering performance.

The mixing distribution in the proposed models has little effect on the performance of the clustering algorithm, but parameter estimation may be awkward for the asymmetric Laplace model. Usually large values of p and q result in more clusters. If the estimates of p or q seem unreasonably large, setting the parameter $0 < q \leq 1$ to some reasonable value and treating p or the signal to noise ratio σ_θ^2/σ^2 as a tuning parameter may be appropriate.

An R-package embodying our approach, *bclust*, is available through the R-CRAN repository (`http://cran.r-project.org`); for details see Partovi Nia and Davison (2012).

ACKNOWLEDGEMENTS

We thank the editor, associate editor and reviewers for constructive comments that have greatly improved the paper, and Gaelle Messerli and Sam Zeeman for bringing the problem to our attention. The work was supported by the Swiss National Science Foundation and the Natural Sciences and Engineering Research Council of Canada.

APPENDIX

The marginal density can be calculated using a hierarchical representation of equation (5) of the paper, i.e.,

$$\begin{aligned} y_{vctr} \mid \eta_{vct} &\overset{\text{iid}}{\sim} N(\eta_{vct}, \sigma^2), \\ \eta_{vct} \mid \theta_{vc} &\overset{\text{iid}}{\sim} N(\theta_{vc}, \sigma_\eta^2), \\ \theta_{vc} \mid \gamma_{vc} &\overset{\text{iid}}{\sim} N(\mu, \gamma_{vc}\sigma_\theta^2), \\ \gamma_{vc} &\overset{\text{iid}}{\sim} B(\delta_v p), \\ \delta_v &\overset{\text{iid}}{\sim} B(q). \end{aligned}$$

We note that η_{vct}, θ_{vc}, and γ_{vc} in this model differ from those of equation (1), but this does not affect the result, because they are integrated out.

Joint density

Since the models impose independent variables $f(y) = \prod_{v=1}^{V} f(y_v)$, by conditioning on δ_v we can write

$$f(y) = \prod_{v=1}^{V} \left\{ q f(y_v \mid \delta_v = 1) + (1-q) f(y_v \mid \delta_v = 0) \right\}, \tag{1}$$

but when $\delta_v = 0$, no variable-class combination is active, yielding

$$f(y_v \mid \delta_v = 0) = \prod_{c=1}^{C} \prod_{t=1}^{T_c} f_0(y_{vct}),$$

where $f_0(y_{vct}) = f(y_{vct} \mid \delta_v = 0) = f(y_{vct} \mid \delta_v = 1, \gamma_{vc} = 0)$. For active variables, however, only data in different classes are independent, that is $f(y_v \mid \delta_v = 1) = \prod_{c=1}^{C} f(y_{vc} \mid \delta_v = 1)$. By summing over values of the Bernoulli variable γ_{vc} we may write

$$f(y_{vc} \mid \delta_v = 1) = p f_1(y_{vc}) + (1-p) \prod_{t=1}^{T_c} f_0(y_{vct}),$$

where $f_1(y_{vc}) = f(y_{vc} \mid \delta_v = 1, \gamma_{vc} = 1)$ corresponds to a density with an active variable-class (cluster) combination, sharing the same θ_{vc}, but involving types with different values of η_{vct} $(t = 1, \ldots, T_c)$. The density $f(y_{vc} \mid \delta_v = 1, \gamma_{vc} = 0)$ equals $\prod_{t=1}^{T_c} f_0(y_{vct})$, because when the variable-class (cluster) combination is inactive, the types inside the class are independent.

Calculation of $f_0(\cdot)$

The density $f_0(\cdot)$ does not depend on the effects θ_{vc}, so under both Gaussian and asymmetric Laplace models it is

$$\begin{aligned} f_0(y_{vct}) = f(y_{vct} \mid \delta_v = 1, \gamma_{vc} = 0) &= \int_{-\infty}^{\infty} \prod_{r=1}^{R_{ct}} f(y_{vctr} \mid \eta_{vct}) f(\eta_{vct}) d\eta_{vct} \\ &= (2\pi\sigma^2)^{-R_{ct}/2} (2\pi\sigma_\eta^2)^{-1/2} \\ &\quad \times \int_{-\infty}^{\infty} \exp\left[-\frac{1}{2\sigma^2} \left\{ \sum_{r=1}^{R_{ct}} (y_{vctr} - \eta_{vct})^2 \right\} - \frac{1}{2\sigma_\eta^2} (\eta_{vct} - \mu)^2 \right] d\eta_{vct}, \end{aligned}$$

which reduces to equation (3) of the paper on completing the square in the exponent and simplifying.

Calculation of $f_1(\cdot)$ for the Gaussian model

When the variable-class (cluster) combination is active, i.e., $f_1(y_{vc}) = f(y_{vc} \mid \delta_v = 1, \gamma_{vc} = 1)$, the Gaussian model is a variance components model. Letting η_{vc} be a vector of length T_c with elements η_{vct} and $\mathbf{Z}$ be a design matrix with $\sum_{t=1}^{T_c} R_{ct}$ rows and T_c columns, we may re-express the reduced model as

$$y_{vc} \mid \eta_{vc} \sim N_{\sum_{t=1}^{T_c} R_{ct}} \left(\mu + \mathbf{Z}\eta_{vc}, \sigma^2 \mathbf{I} \right), \quad \eta_{vc} \sim N_{T_c}(\mathbf{0}, \mathbf{\Omega}),$$

where N_d represents a d-variate Gaussian distribution. The covariance matrix $\mathbf{\Omega}_{T_c \times T_c}$ has diagonal elements $\sigma_\eta^2 + \sigma_\theta^2$ and off-diagonal elements σ_θ^2, obtained after integration over a univariate θ_{vc}. Using standard mixed effects calculations (McCulloch and Searle, 2001, p. 159) the marginalized model over the vector η_{vc} is

$$y_{vc} \sim N_{\sum_{t=1}^{T_c} R_{ct}}(\mu \mathbf{1}, \mathbf{\Sigma} = \sigma^2 \mathbf{I} + \mathbf{Z}\mathbf{\Omega}\mathbf{Z}'),$$

where $\mathbf{\Sigma}$ has diagonal elements $\sigma^2 + \sigma_\eta^2 + \sigma_\theta^2$ and off-diagonal elements $\sigma_\eta^2 + \sigma_\theta^2$ for observations of the same type and σ_θ^2 for observations from different types.

Calculation of $f_1(\cdot)$ for the asymmetric Laplace model

If η_{vc} denotes a vector of length T_c with elements η_{vct}, then the required density $f(y_{vc} \mid \delta_v = 1, \gamma_{vc} = 1)$ is

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \left\{ \prod_{r=1}^{R_{ct}} f(y_{vctr} \mid \eta_{vct}) \right\} f(\eta_{vct} \mid \delta_v = 1, \gamma_{vc} = 1) d\eta_{vc}.$$

We first calculate $f(\eta_{vct} \mid \delta_v = 1, \gamma_{vc} = 1)$, which equals

$$\begin{aligned} &(2\pi\sigma_\eta^2)^{-T_c/2} (2\sigma_{\theta_L})^{-1} \int_{-\infty}^{\mu} \exp\left\{ -\frac{1}{2\sigma_\eta^2} \sum_{t=1}^{T_c} (\eta_{vct} - \theta_{vc})^2 + \frac{\theta_{vc} - \mu}{\sigma_{\theta_L}} \right\} d\theta_{vc} \\ &+ (2\pi\sigma_\eta^2)^{-T_c/2} (2\sigma_{\theta_R})^{-1} \int_{\mu}^{+\infty} \exp\left\{ -\frac{1}{2\sigma_\eta^2} \sum_{t=1}^{T_c} (\eta_{vct} - \theta_{vc})^2 + \frac{\mu - \theta_{vc}}{\sigma_{\theta_R}} \right\} d\theta_{vc}. \end{aligned}$$

If we write $\overline{\eta}_{vc} = T_c^{-1} \sum_{t=1}^{T_c} \eta_{vct}$, the first integral may be expressed as

$$(2\pi\sigma_\eta^2/T_c)^{1/2} \exp\left\{ -\frac{1}{2\sigma_\eta^2} \sum_{t=1}^{T_c} \eta_{vct}^2 - \frac{\mu}{\sigma_{\theta_L}} + \frac{T_c}{2\sigma_\eta^2} \left(\overline{\eta}_{vc} + \frac{\sigma_\eta^2}{T_c \sigma_{\theta_L}} \right)^2 \right\} \Phi\left\{ \frac{\mu - \overline{\eta}_{vc} - \sigma_\eta^2/(T_c\sigma_{\theta_L})}{\sqrt{\sigma_\eta^2/T_c}} \right\},$$

and the second integral as

$$(2\pi\sigma_\eta^2/T_c)^{1/2} \exp\left\{ -\frac{1}{2\sigma_\eta^2} \sum_{t=1}^{T_c} \eta_{vct}^2 + \frac{\mu}{\sigma_{\theta_R}} + \frac{T_c}{2\sigma_\eta^2} \left(\overline{\eta}_{vc} - \frac{\sigma_\eta^2}{T_c \sigma_{\theta_R}} \right)^2 \right\} \Phi\left\{ \frac{\overline{\eta}_{vc} - \mu - \sigma_\eta^2/(T_c\sigma_{\theta_R})}{\sqrt{\sigma_\eta^2/T_c}} \right\}.$$

Hence

$$f(y_{vct} \mid \delta_v = 1, \gamma_{vc} = 1) = (2\pi\sigma^2)^{-\sum_{t=1}^{T_c} R_{ct}/2} \exp\left(-\frac{1}{2\sigma^2} \sum_{t=1}^{T_c} \sum_{r=1}^{R_{ct}} y_{vctr}^2 \right) (k_L J_L + k_R J_R),$$

where k_L and k_R are defined in equation (4) of the paper, the term J_L can be written using the positive definite matrix $\mathbf{A}$, the vectors $\mathbf{b}_L$, $\mathbf{d}_L$ and the constant c_L defined in §2.2 as

$$\begin{aligned} J_L &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \exp\left\{ \eta_{vc}' \mathbf{A} \eta_{vc} - 2\mathbf{b}_L' \eta_{vc} \right\} \Phi\left(c_L + \mathbf{d}_L' \eta_{vc} \right) d\eta_{vc} \\ &= (2\pi)^{T_c/2} |\mathbf{A}|^{-1/2} \exp\left(\frac{1}{2} \mathbf{b}_L' \mathbf{A}^{-1} \mathbf{b}_L \right) \Phi\left(\frac{c_L + \mathbf{d}_L' \mathbf{A}^{-1} \mathbf{b}_L}{\sqrt{1 + \mathbf{d}_L' \mathbf{A}^{-1} \mathbf{d}_L}} \right), \end{aligned}$$

and

$$J_R = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \exp\left\{ \eta_{vc}' \mathbf{A} \eta_{vc} - 2\mathbf{b}_R' \eta_{vc} \right\} \Phi\left(c_R + \mathbf{d}_R' \eta_{vc} \right) d\eta_{vc}$$

can be evaluated in a similar way. The required density (4) is obtained after putting the pieces together.

BIBLIOGRAPHY

Bergé, L., Bouvreyon, C. and Girard, S. (2012) HDclassif: an R package for model-based clustering and discriminant analysis of high-dimensional data. *Journal of Statistical Software* **42**, 1–29.

Bhowmick, D., Davison, A. C., Goldstein, D. R. and Ruffieux, Y. (2006) A Laplace mixture model for identification of differential expression in microarray experiments. *Biostatistics* **7**, 630–641.

Bickel, P. J. and Levina, E. (2004) Some theory for Fisher's discriminant function, naive Bayes, and some alternatives when there are more variables than observations. *Bernoulli* **10**, 989–1010.

Booth, J. G., Casella, G. and Hobert, J. P. (2008) Clustering using objective functions and stochastic search. *Journal of the Royal Statistical Society, series B* **70**, 119–139.

Chang, W.-C. (1983) On using principal components before separating a mixture of two multivariate normal distributions. *Applied Statistics* **32**, 267–275.

Claeskens, G. and Hjort, N. L. (2008) *Model Selection And Model Averaging*. Cambridge: Cambridge University Press.

Dudoit, S., Fridlyand, J. and Speed, T. P. (2002) Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association* **97**, 77–87.

Everitt, B., Landau, S. and Leese, M. (2011) *Cluster Analysis*. Fifth edition. New York: Wiley.

Fraley, C. and Raftery, A. E. (2002) Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association* **97**, 611–631.

Friedman, J. H. (1987) Exploratory projection pursuit. *Journal of the American Statistical Association* **82**, 249–266.

George, E. I. and McCulloch, R. E. (1997) Approaches for Bayesian variable selection. *Statistica Sinica* **7**, 339–373.

Ghahramani, Z. and Beal, M. J. (2000) Variational inference for Bayesian mixtures of factor analyzers. In *Advances in Neural Information Processing Systems*, eds S. A. Solla, T. K. Leen and K. Muller, volume 12, pp. 449–455. MIT Press.

Gohlke, R. S. and McLafferty, F. W. (1993) Early gas chromatography/mass spectrometry. *Journal of the American Society for Mass Spectrometry* **4**, 367–371.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D. and Lander, E. S. (1999) Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science* **286**, 531–537.

Green, P. J. and Mira, A. (2001) Delayed rejection in reversible jump Metropolis–Hastings. *Biometrika* **88**, 1035–1053.

Guo, J., Levina, E., Michailidis, G. and Zhu, J. (2010) Pairwise variable selection for high-dimensional model-based clustering. *Biometrics* **66**, 793–804.

Hall, P., Marron, J. S. and Neeman, A. (2005) Geometric representation of high dimension, low sample size data. *Journal of the Royal Statistical Society, Series B* **67**, 427–444.

Hand, D. J. (2006) Classifier technology and the illusion of progress. *Statistical Science* **21**, 1–14.

Hand, D. J. and Yu, K. (2001) Idiot's Bayes — not so stupid after all? *International Statistical Review* **69**, 385–399.

Hastie, T. J., Tibshirani, R. J. and Friedman, J. H. (2009) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Second edition. New York: Springer.

Heard, N. A., Holmes, C. C. and Stephens, D. A. (2006) A quantitative study of gene regulation involved in the immune response of *Anopheline* mosquitoes: An application of Bayesian hierarchical clustering of curves. *Journal of the American Statistical Association* **101**, 18–29.

Heller, K. A. and Ghahramani, Z. (2005) Bayesian hierarchical clustering. Technical report, Gatsby Computational Neuroscience Unit, University College London.

Kaufman, L. and Rousseeuw, P. J. (1990) *Finding Groups in Data: An Introduction to Cluster Analysis*. New York: Wiley.

Kim, S., Tadesse, M. G. and Vannucci, M. (2006) Variable selection in clustering via Dirichlet process mixture models. *Biometrika* **93**, 877–893.

Lance, G. N. and Williams, W. T. (1967) A general theory of classificatory sorting strategies. 1: Hierarchical systems. *Computer Journal* **9**, 373–380.

Lau, J. W. and Green, P. J. (2007) Bayesian model-based clustering procedures. *Computational Statistics and Data Analysis* **16**, 526–558.

McCulloch, C. E. and Searle, S. R. (2001) *Generalized, Linear, and Mixed Models*. New York: Wiley.

McLachlan, G. J., Bean, R. W. and Peel, D. (2002) A mixture model-based aproach to the clustering of microarray expression data. *Bioinformatics* **18**, 413–422.

McLachlan, G. J. and Peel, D. (2000) *Finite Mixture Models*. New York: Wiley.

McNicholas, P. D. and Murphy, T. B. (2010) Model-based clustering of microarray expression data via latent Gaussian mixture models. *Bioinformatics* **26**, 2705–2712.

Messerli, G., Partovi Nia, V., Trevisan, M., Kolbe, A., Schauer, N., Geigenberger, P., Chen, J., Davison, A. C., Fernie, A. R. and Zeeman, S. C. (2007) Rapid classification of phenotypic mutants of Arabidopsis via metabolite fingerprinting. *Plant Physiology* **143**, 1481–1492.

Mitchell, T. J. and Beauchamp, J. J. (1988) Bayesian variable selection in linear regression (with discussion). *Journal of the American Statistical Association* **83**, 1023–1036.

Pan, W. and Shen, X. (2007) Penalized model-based clustering with application to variable selection. *Journal of Machine Learning Research* **8**, 1145–1164.

Partovi Nia, V. and Davison, A. C. (2012) High-dimensional Bayesian clustering with variable selection: The R package bclust. *Journal of Statistical Software* **47**, 1–22.

Raftery, A. E. and Dean, N. (2006) Variable selection for model-based clustering. *Journal of the American Statistical Association* **101**, 168–178.

Rand, W. M. (1971) Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association* **66**, 846–850.

Scholz, M., Gatzek, S., Sterling, A., Fiehn, O. and Selbig, J. (2004) Metabolite fingerprinting: detecting biological features by independent component analysis. *Bioinformatics* **20**, 2447–2454.

Tadesse, M. G., Sha, N. and Vannucci, M. (2005) Bayesian variable selection in clustering high-dimensional data. *Journal of the American Statistical Association* **100**, 602–617.

Wang, S. and Zhu, J. (2008) Variable selection for model-based high-dimensional clustering and its application to microarray data. *Biometrics* **64**, 440–448.

Witten, D. M. and Tibshirani, R. J. (2010) A framework for feature selection in clustering. *Journal of the American Statistical Association* **105**, 713–726.

Yeung, K. and Ruzzo, W. (2001) Principal component analysis for clustering gene expression data. *Bioinformatics* **17**, 763–774.

Received 14 May 2014
Accepted 13 December 2014