Oliver M. Crook[1,2,3] / Laurent Gatto[4] / Paul D. W. Kirk[3,5]

Fast approximate inference for variable selection in Dirichlet process mixtures, with an application to pan-cancer proteomics

[1] Department of Applied Mathematics and Theoretical Physics, University of Cambridge, Cambridge, UK, E-mail: omc25@cam.ac.uk
[2] Department of Biochemistry, Cambridge Centre for Proteomics, University of Cambridge, Cambridge, UK, E-mail: omc25@cam.ac.uk
[3] MRC Biostatistics Unit, School of Clinical Medicine, University of Cambridge, Cambridge, UK, E-mail: omc25@cam.ac.uk, paul.kirk@mrc-bsu.cam.ac.uk
[4] UCLouvain, de Duve Institute, Brussels, Belgium. https://orcid.org/0000-0002-1520-2268.
[5] University of Cambridge, Cambridge Institute of Therapeutic Immunology & Infectious Disease (CITIID), Cambridge Biomedical Campus Cambridge, United Kingdom of Great Britain and Northern Ireland, E-mail: paul.kirk@mrc-bsu.cam.ac.uk

Abstract:
The Dirichlet Process (DP) mixture model has become a popular choice for model-based clustering, largely because it allows the number of clusters to be inferred. The sequential updating and greedy search (SUGS) algorithm (Wang & Dunson, 2011) was proposed as a fast method for performing approximate Bayesian inference in DP mixture models, by posing clustering as a Bayesian model selection (BMS) problem and avoiding the use of computationally costly Markov chain Monte Carlo methods. Here we consider how this approach may be extended to permit variable selection for clustering, and also demonstrate the benefits of Bayesian model averaging (BMA) in place of BMS. Through an array of simulation examples and well-studied examples from cancer transcriptomics, we show that our method performs competitively with the current state-of-the-art, while also offering computational benefits. We apply our approach to reverse-phase protein array (RPPA) data from The Cancer Genome Atlas (TCGA) in order to perform a pan-cancer proteomic characterisation of 5157 tumour samples. We have implemented our approach, together with the original SUGS algorithm, in an open-source R package named sugsvarsel, which accelerates analysis by performing intensive computations in C++ and provides automated parallel processing. The R package is freely available from: https://github.com/ococrook/sugsvarsel

Keywords: Bayesian clustering, cancer proteomics, variable selection

DOI: 10.1515/sagmb-2018-0065

1 Introduction

Bayesian nonparametric methods have become commonplace in the statistics and machine learning literature due to their flexibility and wide applicability. For model-based clustering, Dirichlet process (Ferguson 1973; 1974) mixture models have become particularly popular (Antoniak, 1974; Lo, 1984; Escobar, 1994; Escobar & West, 1995; Blei & Jordan, 2006), partly because they allow the number of clusters supported by the data to be inferred. By introducing latent selection indicators, these models can be extended to perform variable selection for clustering (Kim, Tadesse & Vannucci, 2006), which is particularly relevant in high-dimensional settings (Law, Figueiredo & Jain, 2004; Constantinopoulos, Titsias & Likas, 2006). There are now several approaches for model-based clustering and variable selection (see Fop & Murphy, 2018, for a recent review), but current Markov chain Monte Carlo (MCMC) algorithms for Bayesian inference in Dirichlet process (DP) mixture models (e.g. Neal, 2000; Jain & Neal, 2004) are computationally costly, and often infeasible for large datasets.

A number of algorithms have been proposed for fast approximate inference in DP and related mixture models, which make possible the analysis of datasets with large numbers of observations. In the present paper, we focus on the sequential updating and greedy search (SUGS) algorithm (Wang & Dunson, 2011; Zhang et al., 2014), which we describe in more detail in Section 2.2. However, there are many other approximate inference

Oliver M. Crook, Paul D. W. Kirk are the corresponding authors.

procedures, a (non-exhaustive, but representative) selection of which we now briefly describe. Variational Bayes (VB) approaches for approximate inference in mixture models have a long history (Attias 1999; 2000), and were extended to DP mixture models by Blei and Jordan (2006). Despite well-known limitations in terms of generally underestimating the variance of the posterior, variational techniques have enabled (approximate) Bayesian inference to be applied to a large class of models and "big data" settings, and are now a mainstay of modern computational Bayesian statistics (Blei, Kucukelbir & McAuliffe, 2016). We note that SUGS was previously shown by Wang and Dunson (2011) to be 10 times faster than VB (largely due to the authors finding that VB required a computationally costly initialisation step in order to provide good results), while performing comparably to VB in terms of model fit. Daumé III (2007) provided an alternative approximate inference strategy that uses fast search algorithms to seek the maximum *a posteriori* (MAP) allocation of observations to clusters, and demonstrated that these techniques permit clustering of very large datasets. The results obtained depend upon the order in which observations are considered, and hence Daumé III (2007) considered a number of ordering strategies. *Bayesian hierarchical clustering* (Heller & Ghahramani, 2005; Savage et al., 2009; Cooke et al., 2011; Darkins et al., 2013) is another method for performing approximate inference for a DP mixture model that also identifies a single optimal clustering structure, but does so using an agglomerative hierarchical clustering approach that determines which clusters to merge at each step on the basis of computed marginal likelihoods. In contrast, by revisiting the widely used k-means algorithm from a Bayesian nonparametric viewpoint, Kulis and Jordan (2012) proposed a novel hard clustering algorithm called *DP-means*, which was subsequently generalised beyond the Gaussian mixtures case (Jiang, Kulis & Jordan, 2012) and was also adapted to cluster large sequencing datasets (Jiang et al., 2016). The MAP-DP approach of Raykov, Boukouvalas, and Little (2016a) is an approximate maximum *a posteriori* inference algorithm for DP mixtures, which has also been proposed as a principled alternative to k-means (Raykov et al., 2016b), but which – in contrast to DP-means – inherits the "rich get richer" property of the DP mixture model, and allows standard model selection and model fit diagnostics to be used (Raykov, Boukouvalas & Little, 2016a). Despite the advances provided by the above methods in terms of reduced computational cost and scalability to large datasets, we note that without variable selection all of these approaches may be ill-suited in high-dimensional settings.

In the spirit of the original SUGS algorithm, here we pose clustering and variable selection as a Bayesian model selection (BMS) problem. We consider variable selection for clustering in terms of partitioning variables into those which are relevant and those which are irrelevant for defining the clustering structure, and thereby pose the problem as one of using BMS to select both a partition of the variables and a partition of the observations. We moreover consider the benefits of performing Bayesian model averaging (BMA) (Madigan & Raftery, 1994; Hoeting et al., 1999) for summarising the SUGS output. For ease of exposition, we focus on the case of DP Gaussian mixtures, but note that all of our methods extend straightforwardly to other distributions for which conjugate priors may be chosen.

We consider a range of simulation settings and well-studied examples from cancer transcriptomics to show that our methods perform competitively with the current state-of-the-art. Having established the utility of our approach, we consider an application to reverse-phase protein arrays (RPPA) datasets in order to characterise the pan-cancer functional proteome. Such datasets have the potential to provide a deeper understanding of the biomolecular processes at work in cancer cells, and have previously been shown to offer additional insights beyond what may be captured by genomics or transcriptomics datasets (Akbani et al., 2014). Here we consider RPPA data for 5157 tumour samples obtained from The Cancer Genome Atlas (TCGA).

Section 2 recaps DP mixture models and the SUGS algorithm, then describes our extensions to SUGS including variable selection and BMA. Section 3 evaluates our method on simulated datasets and compares it with other approaches to clustering and variable selection. We then apply our method to a large proteomics dataset, highlighting its applicability. In the final section, we make some concluding remarks and discuss limitations and extensions. Our methods are implemented in an R package: https://github.com/ococrook/sugsvarsel.

2 Methods

2.1 Dirichlet process mixtures

We provide a very brief recap of DP mixture models, mainly to introduce notation, and refer to the overview provided in Section 3 of Teh et al. (2006) for further details. Let $G \sim DP(\beta P_0)$ where $\beta > 0$ is the DP concentration parameter, P_0 is the base probability measure, and G is a random probability measure. We consider a Pólya urn scheme in which we have independent and identically distributed (i.i.d.) random variables $\theta_1, \theta_2, \ldots$ distributed according to G. Computing the sequential conditional distributions of θ_i given $\theta_1, \ldots, \theta_{i-1}$, upon marginalising out the random G, we obtain (Blackwell & MacQueen, 1973):

$$\theta_i|\theta_1,\dots,\theta_{i-1} \sim \frac{\beta}{\beta+i-1}P_0 + \frac{1}{\beta+i-1}\sum_{l=1}^{i-1}\delta_{\theta_l},\ \ i=1,\dots,n, \tag{1}$$

where δ_θ is a probability measure with mass concentrated at θ. It is clear from this equation that for any $r = 1,2,\dots,i-1$, the probability that θ_i is equal to θ_r is given by $\sum_{l=1}^{i-1}\mathbb{I}(\theta_l=\theta_r)/(\beta+i-1)$, where $\mathbb{I}(X)=1$ if X is true and $\mathbb{I}(X)=0$ otherwise. Thus θ_i has non-zero probability to be equal to one of the previous draws, and it is this clustering property that makes the DP a suitable prior for mixture models.

The DP mixture model is obtained by introducing an additional parametric probability distribution, F. More precisely, let observations x_i be modelled according to the following hierarchical model:

$$\begin{aligned} G &\sim DP(\beta P_0),\\ \theta_i|G &\sim G,\\ x_i|\theta_i &\sim F(\theta_i), \end{aligned} \tag{2}$$

where F denotes the conditional distribution of the observation x_i given θ_i. For example, when F is chosen to be a Gaussian random variable we arrive at the DP Gaussian mixture model (also referred to as the infinite Gaussian mixture model; Rasmussen, 2000).

When performing inference for such models, it is common to introduce a set of latent variables (cluster labels) $z_1, \dots, z_n$ associated with the observations, such that z_i is the cluster label for observation x_i. From the above specification of the DP mixture model, it follows that the conditional prior distribution of z_i given $z_{-i} = (z_1,\dots,z_{i-1})$ is categorical with:

$$\pi_{ik} := P(z_i=k|z_{-i},\beta) = \begin{cases} \dfrac{n_k}{\beta+i-1}, & \text{for } k=1,..,K-1\\ \dfrac{\beta}{\beta+i-1}, & \text{for } k=K, \end{cases} \tag{3}$$

where $\beta > 0$ is the DP concentration parameter, $n_k := \sum_{l=1}^{i-1}\mathbb{I}(z_l=k)$ is the number of previous observations allocated to cluster k, and $K = \max\{z_{-i}\}+1$. Larger values of β encourage observations to be allocated to new clusters, hence β plays a role in controlling the number of clusters.

Inference for DP mixture models can performed using computationally intensive MCMC methods (Neal, 2000; Jain & Neal, 2004). However, as we discuss below, here we are interested in the SUGS algorithm for approximate inference, proposed by Wang and Dunson (2011).

2.2 Sequential updating and greedy search (SUGS)

SUGS is a sequential approach for allocating observations to clusters, which (greedily) allocates the i-th observation to a cluster, given the allocations of the previous $i-1$ observations. Suppose that observations $x_{-i} = (x_1,\dots,x_{i-1})$ have previously been allocated to clusters. As described in Wang and Dunson (2011), the posterior probability of allocating observation i to cluster k according to the DP mixture model formulation above is given by:

$$P(z_i=k|x_i,x_{-i},z_{-i},\beta) = \frac{\pi_{ik}L_{ik}(x_i)}{\sum_{l=1}^{K}\pi_{ik}L_{il}(x_i)}, \tag{4}$$

where π_{ik} is defined as in Equation (3), and

$$L_{ik} = \int f(x_i|\theta_k)p(\theta_k|x_{-i},z_{-i})\,d\theta_k \tag{5}$$

is the conditional marginal likelihood associated with x_i given allocation to cluster k and the cluster allocations for observations $1,\dots,i-1$, with $f(x_i|\theta_k)$ denoting the likelihood associated with x_i as a function of θ_k. If k is a cluster to which previous observations have already been allocated, then $p(\theta_k|x_{-i},z_{-i})$ is the posterior distribution of θ_k given the observations previously allocated to cluster k; i.e. $p(\theta_k|x_{-i},z_{-i}) \propto p_0(\theta_k)\prod_{j:z_j=k,1\le j\le i-1} f(x_j|\theta_k)$, where $p_0(\theta_k)$ is the prior on the cluster-specific parameters, θ_k. For a new cluster, i.e. for $k = K$, we have $p(\theta_k|x_{-i}, z_{-i}) = p_0(\theta_k)$. If p_0 is taken to be conjugate for the likelihood f, then the posterior and conditional marginal likelihood are available analytically.

Assuming that the concentration parameter β is given and that conjugate priors are taken, the above suggests a computationally efficient deterministic clustering algorithm (the SUGS algorithm). That is, z_1 is initialised as $z_1 = 1$, and then subsequent observations are sequentially allocated to clusters by setting $z_i = \arg\max_{k\in\{1,\ldots,K\}} P(z_i = k|x_i, x_{-i}, z_{-i}, \beta)$, where we recall that $K = \max\{z_{-i}\} + 1$ may change after each sequential allocation.

2.2.1 Dealing with unknown β

The DP concentration parameter β directly influences the number of clusters, thus we treat this as a random variable to be inferred, in the same way as in Wang and Dunson (2011). In particular, let $\hat{\boldsymbol{\beta}} = (\hat\beta_1, \ldots, \hat\beta_L)$ be a discrete grid of permissible values for β with a large range, and then define the prior for β to be discrete with the following form:

$$p_0(\beta|\kappa_1, \ldots, \kappa_L) = \sum_{l=1}^{L} \kappa_l \mathbb{1}(\beta = \hat\beta_l), \tag{6}$$

where $\kappa_l = p(\beta = \hat\beta_l)$. Further defining $\phi_l^{(i-1)} = p(\beta = \hat\beta_l|x_{-i}, z_{-i})$ and $\pi_{ikl} = p(z_i = k|\beta = \hat\beta_l, z_{-i})$, the β parameter may be marginalised in Equation (4) to obtain:

$$p(z_i = k|x_{-i}, x_i, z_{-i}) = \frac{\sum_{l=1}^{L} \phi_l^{(i-1)} \pi_{ikl} L_{ik}(x_i)}{\sum_{l=1}^{L} \phi_l^{(i-1)} \sum_{k=1}^{K} \pi_{ikl} L_{ik}(x_i)}, \tag{7}$$

where $\pi_{ikl} := p(z_i = k|\beta = \hat\beta_l, z_{-i})$ is given by Equation (3); $\phi_l^{(0)} = \kappa_l$; and:

$$\phi_l^{(i)} = p(\beta = \hat\beta_l|x_{-i}, x_i, z_{-i}, z_i) = \frac{\phi_l^{(i-1)} \pi_{iz_il}}{\sum_{s=1}^{L} \phi_s^{(i-1)} \pi_{iz_is}} \tag{8}$$

may be calculated sequentially for $i = 1, \ldots, n$. The SUGS algorithm for allocating observations to clusters when β is unknown is then as presented in Algorithm 1.

Algorithm 1: The SUGS algorithm, when the DP precision parameter β is allowed to be unknown.

```
Input  : Data $X = \{x_i\}_{i=1}^n$, Prior $p_0(\theta)$,
         Hyperparameters $\{\kappa_l\}_{l=1}^L$
Output : Cluster allocations $Z = \{z_i\}_{i=1}^n$
Initialise $z_1 = 1$, $K = 2$, and $\{\phi_l^{(0)} = \kappa_l\}_{l=1}^L$;
Evaluate $p(\theta_1|z_1, x_1) \propto p_0(\theta_1) f(x_1|\theta_1)$;
Calculate $\{\phi_l^{(1)}\}_{l=1}^L$, according to Eq. (8);
for $i = 2$ to $N$ do
    for $k = 1$ to $K$ do
        Calculate $L_{ik}$ according to Eq. (5);
        Evaluate $p(z_i = k|x_1, \ldots, x_i, z_1, \ldots, z_{i-1})$ according to Eq. (7);
    end
    Set $z_i = \arg\max_{k=1,\ldots,K}(p(z_i = k|x_1, \ldots, x_i, z_1, \ldots, z_{i-1}))$;
    Set $K = \max\{z_1, \ldots, z_i\} + 1$;
    for $l = 1$ to $L$ do
        Calculate $\phi_l^{(i)}$, according to Eq. (8);
    end
    Evaluate $p(\theta_{z_i}|x_1, \ldots, x_i, z_1, \ldots z_i) \propto p_0(\theta_{z_i}) \prod_{j:z_j=z_i, 1\leq j\leq i} f(x_j|\theta_{z_i})$;
end
```

2.2.2 Formulation of Bayesian model selection problem

A notable limitation of the (deterministic) SUGS algorithm as presented so far is that the clustering structure obtained is dependent upon the initial ordering of the observations. To remove this dependence, Wang and Dunson (2011) consider multiple permutations of this ordering, and pose SUGS as a Bayesian model selection (BMS) problem. More concretely, the algorithm is repeated for many random orderings of the data and a final partition of the observations is then chosen by optimising an appropriate objective function for BMS, such as the marginal likelihood (ML):

$$L(X|Z) = \prod_{k=1}^{K} \int_{\theta_k} \left[\prod_{i:z_i=k} f(x_i|\theta_k) \right] p_0(\theta_k) d\theta_k. \tag{9}$$

In practice, Wang and Dunson (2011) advocate optimising the *pseudo*-marginal likelihood (PML), since they found that the marginal likelihood to often produce many small clusters. The PML is given by:

$$\begin{aligned} \mathrm{PML}_z(X) &= \prod_{i=1}^{N} p(x_i|X_{n\backslash -i}, z_{n\backslash -i}) \\ &= \prod_{i=1}^{N} \int_{\theta} p(x_i|\theta) p(\theta|X_{n\backslash -i}, z_{n\backslash -i}) d\theta \\ &= \prod_{i=1}^{N} \sum_{k=1}^{K} P(z_i = k|X_{n\backslash -i}, z_{n\backslash -i}) \int_{\theta_k} f(x_i|\theta_k) p(\theta_k|X_{n\backslash -i}, z_{n\backslash -i}) d\theta_k, \end{aligned} \tag{10}$$

where, defining $X = \{x_1, \ldots, x_n\}$ and $Z = \{z_1, \ldots, z_n\}$, we have $X_{n\backslash -i} = X\backslash\{x_i\}$ is the set of all observations except the *ith*, and similarly $z_{n\backslash -i} = Z\backslash\{z_i\}$. In addition, Wang and Dunson (2011) remark that that $p(x_i|X, Z)$ can be used to approximate $p(x_i|X_{n\backslash -i}, z_{n\backslash -i})$ to speed up computations and that this approximation is accurate for large sample sizes.

2.3 SUGS for variable selection

Irrelevant variables in high-dimensions can present a considerable challenge for clustering models and algorithms, because the number of variables with no clustering structure can overwhelm those where a clustering structure exists (Witten & Tibshirani, 2010). There have been many approaches to model-based clustering and variable selection (e.g. Raftery & Dean, 2006; Maugis, Celeux & Martin-Magniette, 2009), and we direct readers to Fop and Murphy (2018) for a recent review. However, many of these scale poorly with increasing dataset dimension, and/or require the number of clusters to be determined as a separate analysis step. To address these challenges, here we extend the SUGS algorithm to simultaneously perform clustering and variable selection, and refer to the resulting procedure as *SUGSVarSel*.

Since we are in the high-dimensional setting, we assume for simplicity that variables are independent given the cluster allocations (which, in the Gaussian case, is equivalent to assuming a diagonal structure for the covariance matrix). Let $x_{i,d}$ be the *dth* element of the *ith* observation vector, with $d = 1, \ldots, D$, and D the number of variables. Introducing indicator variables γ_d, which is 1 if the *dth* variable is relevant for the clustering structure and 0 if not, we follow a common approach from the literature (Law, Figueiredo & Jain, 2004; Tadesse, Sha & Vannucci, 2005; Kim, Tadesse & Vannucci, 2006) and assume that the cluster conditional likelihood can be factorised as follows:

$$f(x_i|\theta, \theta_0, z_i = k) = \prod_{d=1}^{D} f(x_{i,d}|\theta_{k,d})^{\mathbb{1}(\gamma_d=1)} f(x_{i,d}|\theta_{0,d})^{\mathbb{1}(\gamma_d=0)}, \tag{11}$$

where θ_0 are "global" (i.e. not cluster-specific) parameters. In other words, the variables for which $\gamma_d = 1$ are modelled by a mixture distribution with cluster-specific parameters $\theta_{k,\,d}$, while the variables for which $\gamma_d = 0$ are modelled by a single component with (global, not cluster-specific) parameters $\theta_{0,d}$. Having introduced the D indicator variables γ_d, we now extend the SUGS algorithm in order to estimate them.

2.3.1 The SUGSVarSel algorithm

Given a realisation of the indicator variables, $\Gamma = \{\gamma_1, \dots, \gamma_D\}$, we may plug the cluster conditional likelihood given in Equation (11) into Equation (5) and proceed as before in order to identify a clustering, Z.

Conversely, suppose we have a realisation, Z, of the set of component allocation variables, but that the indicator variables Γ are unknown. In this case, the posterior probabilities associated with the variable indicators are given by:

$$P(\gamma_d = 1|X, Z) = \frac{p_0(\gamma_d = 1)}{B} \prod_{k \in Z} \int_{\theta_{k,d}} \left(\prod_{i:z_i = k} f(x_{i,d}|\theta_{k,d}) \right) p_0(\theta_{k,d}) d\theta_{k,d} \tag{12}$$

$$P(\gamma_d = 0|X, Z) = \frac{p_0(\gamma_d = 0)}{B} \int_{\theta_{0,d}} \left(\prod_{i:z_i = k} f(X_d|\theta_{0,d}) \right) p_0(\theta_{0,d}) d\theta_{0,d}, \tag{13}$$

where $p_0(\gamma_d = q)$ indicates the prior probability that $\gamma_d = q$, and B is a normalising constant that ensures that $p(\gamma_d = 0|X, Z)$ and $p(\gamma_d = 1|X, Z)$ sum to 1. Thus, given a realisation, Z, of the set of component allocation variables, a greedy approach to finding γ_d is to set $\gamma_d = \arg\max_{q \in \{0,1\}} P(\gamma_d = q|X, Z)$.

Given an initial realisation of the indicator variables, $\Gamma = \Gamma^{(0)}$, the above suggests an iterative strategy in which at each iteration we use the SUGS algorithm to find a partition $Z^{(t)}$ given $\Gamma^{(t-1)}$, and then greedily update the indicator variables according to Equations (12) and (13) above in order to obtain $\Gamma^{(t)}$ given $Z^{(t)}$. This algorithm, which we refer to as SUGSVarSel, is presented in Algorithm 2.

```
Algorithm 2: The SugsVarSel algorithm
Input  : Data $X = \{x_i\}_{i=1}^n$, Priors $p_0(\theta)$ and $p_0(y)$,
         Hyperparameters $\{\kappa_l\}_{l=1}^L$, Initial Indicator Switches $\Gamma^{(0)}$,
         Maximum Iterations $T$.
Output : Cluster allocation $Z = \{z_i\}_{i=1}^n$
         Variable switches $\Gamma = \{y_d\}_{d=1}^D$
Initialise $z_1 = 1$, $K = 2$, and $\{\phi_l^{(0)} = \kappa_l\}_{l=1}^L$;
Evaluate $p(\theta_1|z_1, x_1) \propto p_0(\theta_1) f(x_1|\theta_1)$;
Calculate $\{\phi_l^{(1)}\}_{l=1}^L$, according to Eq. (8);
while $t \leq T$ do
    for $i = 2$ to $N$ do
        for $k = 1$ to $K$ do
            Calculate $L_{ik}$ given $\Gamma^{(t-1)}$, according to Eqs. (5) and (11);
            Evaluate $p(z_i = k|x_1, \dots, x_i, z_1, \dots, z_{i-1})$ according to Eq.
             (7);
        end
        Set $z_i = \arg\max_{k=1,\dots,K}(p(z_i = k|x_1, \dots, x_i, z_1, \dots, z_{i-1}))$;
        Set $K = \max\{z_1, \dots, z_i\} + 1$;
        for $l = 1$ to $L$ do
            Calculate $\phi_l^{(i)}$, according to Eq. (8);
        end
        Evaluate, using the cluster conditional likelihood in Eq. (11),
         $p(\theta_{z_i}|x_1, \dots, x_i, z_1, \dots z_i) \propto p_0(\theta_{z_i}) \prod_{j:z_j = z_i, 1 \leq j \leq i} f(x_j|\theta_{z_i})$;
    end
    for $d = 1$ to $D$ do
        Calculate $p(y_d = r|X, Z)$, according to Eqs. (12) and (13);
        Set $y_d = \arg\max_{r \in \{0,1\}}(p(y_d = r|X, Z))$;
    end
    $t \leftarrow t + 1$
end
```

2.3.2 Initialisation strategies for SUGSVarSel

Like the SUGS algorithm, the output of SUGSVarSel depends upon the initial ordering of the observations. It moreover depends upon the initialisation of the variable selection switches, $\Gamma^{(0)}$. To address this latter issue, we propose a random sub-sampling initialisation strategy. This is as follows: first randomly select p_1 variables (with $1 < p_1 \leq D$) and apply SUGSVarSel on this new dataset $\tilde{X}$ of size $n \times p_1$ with a small number of random orderings of the observations (we find 10 works in practice). The initial indicator for the variables of $\tilde{X}$, which we write as $\tilde{\Gamma}^{(0)}$, are set as all-on ($\gamma_d = 1$ for these p_1 variables). $\tilde{\Gamma}^{(0)}$ is held the same for each of the random orderings. For each of the random orderings, this approach outputs $\tilde{Z}$ for all observations but $\tilde{\Gamma}$ for only a subset of size p_1 of the variables. To obtain Γ for all D variables, we use the cluster allocations $\tilde{Z}$ and the full data X to compute probabilities for the remaining variables using 12 and 13. We then greedily assign the indicator variables. A single best model generated by these random orderings is selected using the ML. This procedure returns a $\Gamma_1 \in \{0,1\}^D$; that is, variable selection switches with some variables switched on and other variables switched off. We repeat this process for a total of M random sub-samples of the variables to produce a set of clusterings $Z_1, \dots, Z_M$ and a set of variables $\Gamma_1, .., \Gamma_M$. These variable sets are then used as initial inputs $\Gamma^{(0)} = \Gamma_i$ for $i = 1, \dots,$ M for the SUGSVarSel algorithm (which is now run using all variables $p = D$) with Q new random orderings (again we find 10 is sufficient in practice). This SUGSVarSel with sub-sampling initialisation strategy returns Q models for each random sub-sample of the variables. Thus, we have QM models from which to choose. For each model obtained in this way, we calculate the marginal likelihood (see Section 2.2.2). We can then perform BMS to obtain a single "best" model, or we can use Bayesian model averaging (BMA; see next section).

2.4 Bayesian model-averaged co-clustering matrices

2.4.1 Bayesian model averaging

The output of our algorithm is a set of clusterings, associated variables and a marginal likelihood. One can select a single "best" model amongst these possible clustering, however we can also average over these models to capture the model uncertainty. The idea is called Bayesian model averaging (BMA) and we apply the method to clustering and variable selection (Madigan & Raftery, 1994; Hoeting et al., 1999; Russell, Murphy & Raftery, 2015).

For each model we form a co-clustering matrix S. S is defined in the following way:

$$S_{ij} = \begin{cases} 0, & \text{if } z_i \neq z_j \\ 1, & \text{if } z_i = z_j. \end{cases} \tag{14}$$

That is the *ijth* entry of S is 1 if observation x_i and x_j are in the same cluster and 0 otherwise. We note that the S is invariant to relabelling and the number of clusters. Now, suppose we have M models $\mathcal{M}_1, \dots, \mathcal{M}_M$, letting X be our observations and θ_m be the parameters associated with model $\mathcal{M}_m$. The posterior probability for $\mathcal{M}_m$ is given by

$$p(\mathscr{M}_m|X) = \frac{p(X|\mathscr{M}_m)p_0(\mathscr{M})}{\sum_{l=1}^{M} p(X|\mathscr{M}_l)p_0(\mathscr{M}_l)}, \tag{15}$$

where

$$P(X|\mathscr{M}_m) = \int P(X|\theta_m, \mathscr{M}_m)P(\theta_m|\mathscr{M}_m)\, d\theta_m. \tag{16}$$

The marginal likelihood (16) is the key quantity for model comparison and can be interpreted as the weight given to each proposed model. Further note the two sources of averaging: the averaging over the parameters in the ML and the averaging over the models in equation (15). We suppose that *a priori* all models are equally likely, choosing the prior on each model to be $p_0(\mathscr{M}_m) = 1/M$. One computational challenge that (15) gives us is computing the summation, since it can involve evaluating possibly thousands of models. To overcome this, one can discount models that are poor at describing our observations comparatively to our best model. More precisely, let us form Occam's window (Hoeting et al., 1999):

$$\mathscr{W} = \left\{ M_k : \frac{\max_l(p(M_l|X))}{p(M_k|D)} \leq K \right\}, \tag{17}$$

where K is a tuning parameter. Occam's window is the set of all possible models within a reasonable Bayes factor from the best model under consideration. The summation in (15) is then replaced with a summation over the set $\mathscr{W}$.

2.4.2 Averaging the co-clustering matrices

We can form the Bayesian model-averaged co-clustering matrix (BMAC) by taking the set of co-clustering matrices $S_{\mathscr{W}}$ and averaging, weighting by their ML:

$$S_{BMAC} = \frac{p(X|\mathscr{M}_m)S_m}{\sum_{l\in\mathscr{W}} p(X|\mathscr{M}_l)}. \tag{18}$$

The BMA of the variable set can be found in the same way by averaging over the weighted variable sets for each model:

$$\mathscr{F}_{BMA} = \frac{p(X|\mathscr{M}_m)\mathscr{F}_m}{\sum_{l\in\mathscr{W}} p(X|\mathscr{M}_l)}, \tag{19}$$

where we denote by $\mathscr{F}_m$ the variable set associated with model $\mathcal{M}_m$.

3 Comparisons with the state-of-the-art

We compare sugsVarSel to a number of alternative algorithms, and demonstrate the performance of our method in two situations. The first is the $p > n$ paradigm, where the number of variables exceeds the number of observations. The second situation considers $n > p$ for $n = 1000$, while simultaneously considering different proportions of variables being relevant. In both cases, we consider a variety of scenarios, for which different proportions of the variables are relevant.

3.1 Alternative methods for clustering and variable selection

We compare our method relative to the current state-of-the-art, including methods that do and do not peform variable selection. These include: mclust, a finite mixture model based clustering method (Fraley & Raftery, 2002; Fraley et al., 2012; Scrucca et al., 2016); DP-means, a non-parametric interpretation of K-means (Kulis & Jordan, 2012); clustvarsel, a finite mixture model method with variable selection (Raftery & Dean, 2006; Maugis, Celeux & Martin-Magniette, 2009; Scrucca & Raftery, 2014); the original sequential updating and greedy search algorithm (Wang & Dunson, 2011) as implemented in our sugsvarsel R package; and VarSelLCM, a model-based clustering and variable selection approach using the integrated complete-data likelihood (Marbac & Sedki, 2017).

3.2 High-dimensional example

In the first example, we simulate a mixture of 3 Gaussians with mixture proportions 0.5, 0.3, 0.2 centred at $(0, 0, .., 0), (2, 2, \dots, 2), (-2, -2, \dots, -2)$ respectively, each with variance-covariance matrix equal to the identity. The irrelevant variables are simulated from a standard Gaussian. First, we simulate 100 observations from this model with 200 variables and explore varying the number of relevant variables.

When running SUGS and SUGSVarSel we use the same prior specification for both methods and 30 random orderings of the data. Throughout this article, we always perform 2 iterations of variable selection in the SUGSVarSel algorithm. To initialise variable selection in SUGSVarSel, we subsample 10% of the variables 20 times to produce an initial variable selection set. For SUGS we choose the partition with maximal PML (as advised in the original SUGS paper by Wang and Dunson 2011), while for SUGSVarSel we select the result with maximal ML. Prior choices for SUGS and SUGSVarSel can be found in the Supplementary Material. For mclust and clustvarsel, we find the appropriate number of clusters using a sequential search up to a maximum of 9 possible clusters. We then use then Bayesian Information Criterion (BIC) to select an appropriate model (Schwarz, 1978). For DP-means we repeat the algorithm over a range of penalty parameters $\lambda = \{0.01, 0.1, 1, 10, 100, 200, 400, 600, 800, 1000\}$ and select the partition which minimises the DP-means objective function. For VarSelLCM we run the algorithm up to a maximum of 9 possible clusters and select an appropriate model using the Maximum Integrated Complete-data Likelihood (MICL) (Marbac & Sedki, 2017, 2018). All methods are run in serial for fair comparison.

Results are presented in Table 1–Table 4. In all tables, we provide runtimes for each of the methods, indicate the proportion of relevant and irrelevant variables that each method correctly identified (for methods without variable selection this is reported as 1 for relevant and 0 for irrelevant variables), and report the adjusted Rand index (Rand, 1971; Hubert & Arabie, 1985) between the clustering produced and the truth. We repeat all methods for 10 different random realisation of the datasets to produce a distribution of scores. We report the median scores, along with the upper and lower quartiles.

Table 1: High-dimensional simulation example where 100 observations are simulated from a Gaussian mixture distribution with 3 components and 200 variables, in which 50% of variables are relevant.

Method	Time, secs	Correct relevant variables	Correct irrelevant variables	ARI
mclust	<1	1	0	1 [1, 1]
DP-means	<1	1	0	0.60 [0.37, 0.66]
clustvarsel	14280.8 [10431.6, 20310.4]	0.47 [0.45, 0.48]	1 [1, 1]	1 [1, 1]
SUGS	0.92 [0.90, 0.97]	1	0	0.955 [0.90, 0.97]
SUGSVarSel	24.6 [23.8, 24.9]	1 [1, 1]	1 [1, 1]	1 [1, 1]
VarSelLCM	620.0 [574.9, 650.8]	1 [1, 1]	1 [1, 1]	1 [1, 1]

Table 2: High-dimensional simulation example where 100 observations are simulated from a Gaussian mixture distribution with 3 components and 200 variables, in which 25% of variables are relevant.

Method	Time, secs	Correct relevant variables	Correct irrelevant variables	ARI
mclust	<1	1	0	1 [1, 1]
DP-means	<1	1	0	0.74 [0.70, 0.79]
clustvarsel	1852.3 [1185.2, 5880.8]	0.02 [0.02, 0.02]	0.847 [0.812, 0.945]	0.01 [0.00, 0.04]
SUGS	2.07 [1.89, 2.16]	1	0	0.78 [0.72, 0.84]
SUGSVarSel	21.9 [21.9, 22.1]	1 [1, 1]	1 [1, 1]	1 [1, 1]
VarSelLCM	487.7 [481.3, 494.1]	1 [1, 1]	1 [1, 1]	1 [1, 1]

Table 3: High-dimensional simulation example where 100 observations are simulated from a Gaussian mixture distribution with 3 components and 200 variables, in which 10% of variables are relevant.

Method	Time, secs	Correct relevant variables	Correct irrelevant variables	ARI
mclust	<1	1	0	0 [0, 0]
DP-means	<1	1	0	0 [0, 0]
clustvarsel	3095.8 [2377.3, 3302.7]	0.05 [0.05, 0.10]	0.803 [0.778, 0.854]	0 [0, 0]
SUGS	5.02 [4.76, 5.23]	1	0	0.18 [0.13, 0.21]
SUGSVarSel	19.7 [19.5, 19.9]	1 [1, 1]	1 [1, 1]	1 [1, 1]
VarSelLCM	523.5 [521.6, 532.0]	1 [1, 1]	1 [1, 1]	1 [1, 1]

Table 4: High-dimensional simulation example where 100 observations are simulated from a Gaussian mixture distribution with 3 components and 200 variables, in which 5% of variables are relevant.

Method	Time, secs	Correct relevant variables	Correct irrelevant variables	ARI
mclust	<1	1	0	0 [0, 0]
DP-means	<1	1	0	0 [0, 0]
clustvarsel	2183.1 [802.5, 2992.4]	0.1 [0.1, 0.1]	0.879 [0.814, 0.959]	0 [0, 0]

SUGS	6.30 [6.07, 10.11]	1	0	0.04 [0.02, 0.05]
SUGSVarSel	19.9 [19.7, 20.5]	1 [1, 1]	1 [1, 1]	1 [1, 1]
VarSelLCM	583.5 [521.6, 532.0]	1 [1, 1]	1 [1, 1]	1 [1, 1]

It is evident that methods that do not perform variable selection such as mclust and SUGS perform poorly when there are many irrelevant variables. The performance of clustvarsel here seems volatile and performs poorly at correctly selecting relevant features. VarSelLCM and SUGSVarSel are competitive in terms variable selection and clustering. However, VarSelLCM requires an exhaustive search over the number of clusters, which makes this method computationally costly to apply when the number of clusters is not known. SUGSVarSel outperforms all variable selection and clustering methods in terms of speed, while also automatically inferring the number of clusters in the data. We proceed to evaluate the performance of SUGSVarSel on large simulated datasets.

3.2.1 Increasing the number of observations

We simulate the same distribution as before, but instead sample 1000 observations and only 100 variables and the irrelevant variable are simulated from a standard Gaussian distribution. All priors are the same as in the previous analysis and we sub-sample 10% of the variables 10 times to produce an initial variable selection set. We repeat SUGS and SUGSVarSel for 10 random orderings of the data. We compare the scalable methods mclust, DP-means, SUGS, SUGSVarSel and VarSelLCM, where 25%, 10%, 5% of the variable are relevant. For SUGS we choose the partition with maximal PML, while for SUGSVarSel we select the result with maximal ML. For VarSelLCM we run the algorithm for possible number of clusters 1 through 4 and select an appropriate model using the MICL, as previously. Results are presented in Table 5–Table 7.

Table 5: Simulation example where 1000 observations are simulated from a Gaussian mixture distribution with 3 components and 100 variables, in which 25% of variables are relevant.

Method	Time, secs	Correct relevant variables	Correct irrelevant variables	ARI
mclust	11.2 [10.9, 11.6]	1	0	0 [0, 0]
DP-means	22.3 [21.7, 23.1]	1	0	0.30 [0.25, 0.36]
SUGS	3.4 [3.1, 3.6]	1	0	0.98 [0.97, 0.98]
SUGSVarSel	31.2 [30.7, 31.8]	1 [1, 1]	1 [1, 1]	1 [1, 1]
VarSelLCM	3596.8 [2639.5, 7537.7]	1 [1, 1]	1 [1, 1]	1 [1, 1]

Table 6: Simulation example where 1000 observations are simulated from a Gaussian mixture distribution with 3 components and 100 variables, in which 10% of variables are relevant.

Method	Time, secs	Correct relevant variables	Correct irrelevant variables	ARI
mclust	11.0 [10.7, 11.4]	1	0	0 [0, 0]
DP-means	21.4 [21.0, 21.8]	1	0	0.11 [0.02, 0.22]
SUGS	5.1 [4.9, 5.3]	1	0	0.01 [0.01, 0.04]
SUGSVarSel	33.3 [33.0, 33.8]	1 [1, 1]	1 [1, 1]	0.90 [0.80, 0.97]
VarSelLCM	1938.5 [1852.3, 1973.9]	1 [1, 1]	1 [1, 1]	0.997 [0.994, 0.997]

Table 7: Simulation example where 1000 observations are simulated from a Gaussian mixture distribution with 3 components and 100 variables, in which 5% of variables are relevant.

Method	Time, secs	Correct relevant variables	Correct irrelevant variables	ARI
mclust	11.4 [11.2, 15.7]	1	0	0 [0, 0]

DP-means	22.0 [21.1, 22.7]	1	0	0 [0, 0]
SUGS	6.3 [5.6, 11.1]	1	0	0 [0, 0]
SUGSVarSel	60.8 [59.8, 64.2]	1 [1, 1]	1 [0.99, 1]	0.78 [0.54, 0.92]
VarSelLCM	2688.8 [2588.9, 2878.6]	1 [1, 1]	1 [1, 1]	0.943 [0.931, 0.945]

Mclust, SUGS and DP-means produce poor quality clusterings, because irrelevant variables present in the data render finding the true underlying clustering structure challenging. SUGSVarSel and VarSelLCM produce high quality answers in all situations but SUGSVarSel is 2 orders of magnitude faster. However, to alleviate the computational burden we searched up to a maximum of 4 clusters in VarSelLCM, providing it with an easier opportunity to produce high quality clusterings. In applications to real data this would have to be much larger, adding considerably to computational time, whereas the inference of the number of clusters is automatic in SUGSVarSel.

3.3 Advantages of Bayesian model averaging

As an example, we simulate a dataset with 30 observations from a mixture of 3 Gaussians, where two of the Gaussians are isotropic and centred (2, 2) and (−3, −3), respectively, each with mixing weights 0.4. The third component has mixture weight 0.2 and is centered at (−3, 4) but the covariance matrix is 2 on the diagonals and 1 on the off diagonals, violating our independence assumption. We additionally include 2 components of irrelevant variables generated from standard Gaussians. Our prior specifications are set as in the previous section. Simply using the ML to pick a partition results in an ARI of 0.635 between the clustering produced and the truth. However, we can also perform BMA and then summarise our co-clustering. We applied hierarchical clustering with average linkage to compute a clustering, which has previously be applied to posterior similarity matrices (Medvedovic, Yeung & Bumgarner, 2004; Fritsch & Ickstadt, 2009; Liverani et al., 2015) (see Supplementary Material for complete details). This clustering then produces an ARI of 0.875. The heatmap of the co-clustering matrix is plotted in Figure 1, allowing us to visualise the uncertainty in the clustering.

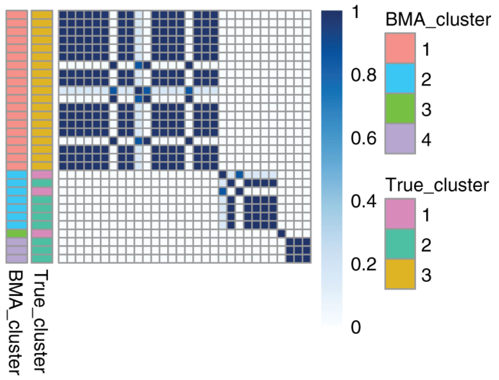

Figure 1: A heatmap of the BMA co-clustering matrix, where dark blue indicates the probability of being in the same cluster is 1 and white indicates a probability of 0 of belonging to the same cluster. The component annotation bar indicates the true component labels and the cluster annotation bar indicates the clustering obtained from summarising the BMA co-clustering matrix.

4 Applications to cancer subtyping

4.1 Application to leukaemia dataset

In this section, we apply SUGSVarSel to real biological datasets. The first is a well-studied genomic clustering problem: the separation of acute myeloid leukaemia (AML) and the B/T-cell subtypes of acute lymphoblastic leukemia (ALL) samples on the basis of microarray transcriptomic data. We use the dataset described by Golub et al. (1999), which comprises 38 samples, 27 of which are ALL (8 T-cell and 19 B-cell related), and 11 of which are AML cases. Initial preprocessing is performed as in Dudoit, Fridlyand, and Speed (2002), which reduces the dimension of the dataset from 6817 to 3051 genes. In Dudoit, Fridlyand, and Speed (2002), a further dimension reduction step is performed that makes use of the AML and ALL class labels, so that only those genes that have a high ratio of their between-class to within-class sums of squares are retained. Here we instead wish to adopt a completely unsupervised approach, so that we may use the known ALL-AML class label in order to validate our results.

We select the 200 most variable genes and then normalise, so the expression values for each gene are mean-centred at 0 with variance 1. 200 genes were chosen because this led to good predictive performance in previous analysis of these data (Golub et al., 1999; Dudoit, Fridlyand & Speed, 2002). We then apply SUGSVarSel to the resultant dataset. We sub-sample 10% of the variables 20 times to produce an initial variable selection set, and run the algorithm for 100 random orderings. We adopt our default priors and summarise the output using BMA. A final summary clustering is obtained by performing hierarchical clustering with average linkage (Fritsch & Ickstadt, 2009). We use the ARI to compare our results to the truth (of 3 classes) and repeat the process 10 times and report the average results.

Results are illustrated in Figure 2. The final clustering result provides an ARI of 0.831, which is in line with previous analyses preformed on this dataset (Golub et al., 1999; Dudoit, Fridlyand & Speed, 2002). The algorithm selects a total of 92 genes, including TCL1, TCRB, IL8, EPB72, IL7R, TCRG, NFIL6, which are all known to be associated with leukaemia (Natsuka et al., 1992; Pekarsky, Hallas & Croce, 2001; Van der Velden et al., 2004; Kuett et al., 2015; Chen, Tsau & Lin, 2010; Shochat et al., 2011). A full list of the selected genes (including their descriptions) can be found in the Supplementary Material. The advantage of our analysis over other methods is that we did not need to specify the number clusters – the algorithm automatically inferred 3 clusters in the data, which have excellent correspondence to the known classes of AML and ALL, as well as the 2 ALL subgroups.

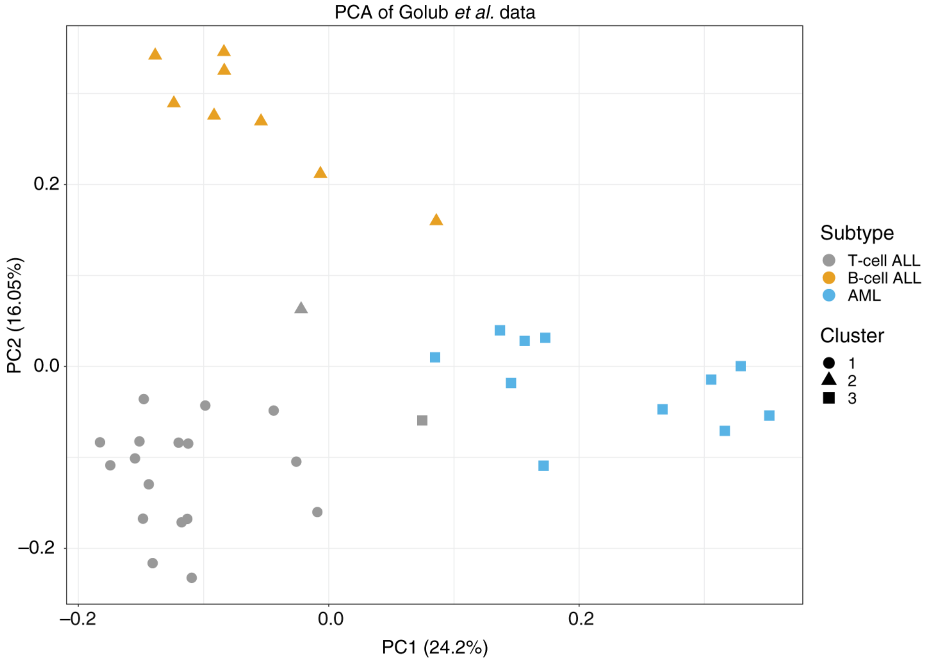

Figure 2: A PCA plot of the microarray expression data of 38 patients from the Golub et al. (1999) dataset, using the 200 most variable genes. The different symbols indicate the clustering produced by the SUGSVarSel algorithm after summarising the BMA co-clustering matrix using hierarchical clustering with average linkage. The colours indicate the annotated sub-types.

To assess the importance of variable selection, we also apply mclust and the original SUGS algorithm to the data. We run the mclust algorithm performing a systematic search to select the number of clusters, up to a maximum of 9, and select the number of cluster which maximises the BIC. This criterion selects 3 clusters and clustering produced gives an adjusted Rand index of 0.627 – the inclusion of irrelevant variables has led to reduced cluster quality. We run SUGS using our default prior choices and using the PML criterion to select a clustering. The algorithm was run for 100 random ordering and we repeated the process 10 times, reporting an average ARI of 0. The lack of variable selection renders SUGS unable to produce a meaningful clustering. In Figure 3, we visualise the BMA co-clustering matrix for these data when applying the SUGSVarSel algorithm.

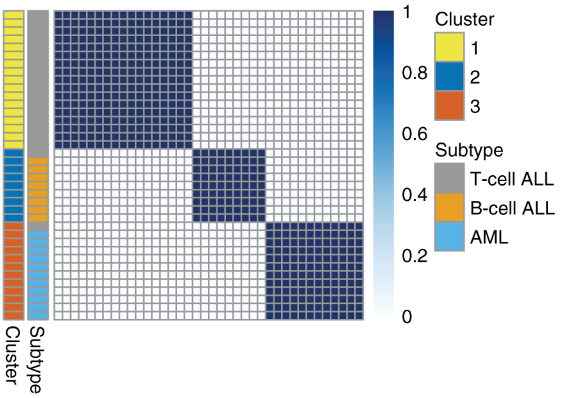

Figure 3: A heatmap of the BMA co-clustering matrix for the 38 patients, when applying SUGSVarSel, demonstrating the added benefit of visualising uncertainty. The annotation bars of the left indicate the correspondence between the clusters and the subtypes.

4.2 Application to TCGA breast cancer dataset

We demonstrate SUGSVarSel on a further genomics dataset. We analyse an expression dataset for breast cancer tumour data from The Cancer Genome Atlas (TCGA) (Network, 2012), which we pre-process in the same way as in Lock and Dunson (2013). The processed expression dataset comprises 348 tumours with 645 genes, of which 14 belong to the PAM50 (Prediction Analysis of Microarray) group of genes (Parker et al., 2009).

Analysis was performed in the following way. We first standardise our data so that each column is mean-centred with variance 1. We then subsample 10% of the variables 64 times to produce an initial variable set. We then apply the SUGSVarSel algorithm with default settings. We summarise our output by performing BMA and then hierarchical clustering with average linkage.

SUGSVarSel reveals two clusters in the dataset, the second of which is significantly associated with Basal-like tumours (Fisher test, $p < 0.0001$). The algorithm selects 245 variables to discriminate between the groups. We perform PCA before and after variable selection to demonstrate that the reduced variable set produces more separable and therefore more interpretable clusters. Furthermore, the algorithm selected 13 out of a total of 14 of the PAM50 genes, which is significantly better than random (Fisher Test, $p < 0.0001$).

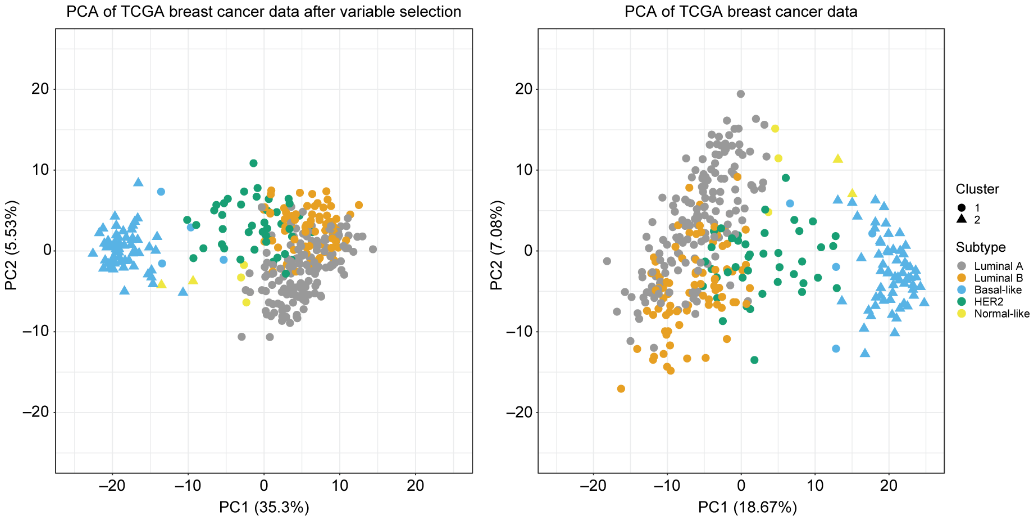

Figure 4: PCA plot on the TCGA breast cancer data, where clusters produced by SUGSVarSel are indicated by shape and subtypes by colour. The left PCA plot demonstrates smaller and tighter clusters using only the variables that remained after variable selection. In the right hand plot all variable were used to produce the plot.

There is perhaps concern that variable selection could remove relevant genes for clustering, in the situation where we have a highly informative set of variables. We consider the following task to cluster the breast cancer genes using the PAM50 genes from the total unprocessed dataset (that is without the filtering of Lock and Dunson (2013)), of which there are 48. We apply the SUGSVarSel in identical fashion to before, sub-sampling 10% of the variables 4 time to produce an initial variable set. We obtain 5 clusters which correspond well to the different breast cancer subgroups.

Cluster 1 is associated with Luminal A cancers, cluster B is associated with Luminal cancers, cluster 3 with basal-like tumours, cluster 4 contains mostly HER2 type breast cancers (chi-squared $p < 0.0001$). Thus, hardly surprisingly, the cluster produce on the PAM50 data coincide well with the PAM50 subgroups. Furthermore, 87.5% of the genes were selected which is more than we expect given our prior, telling us this was a highly informative set of genes.

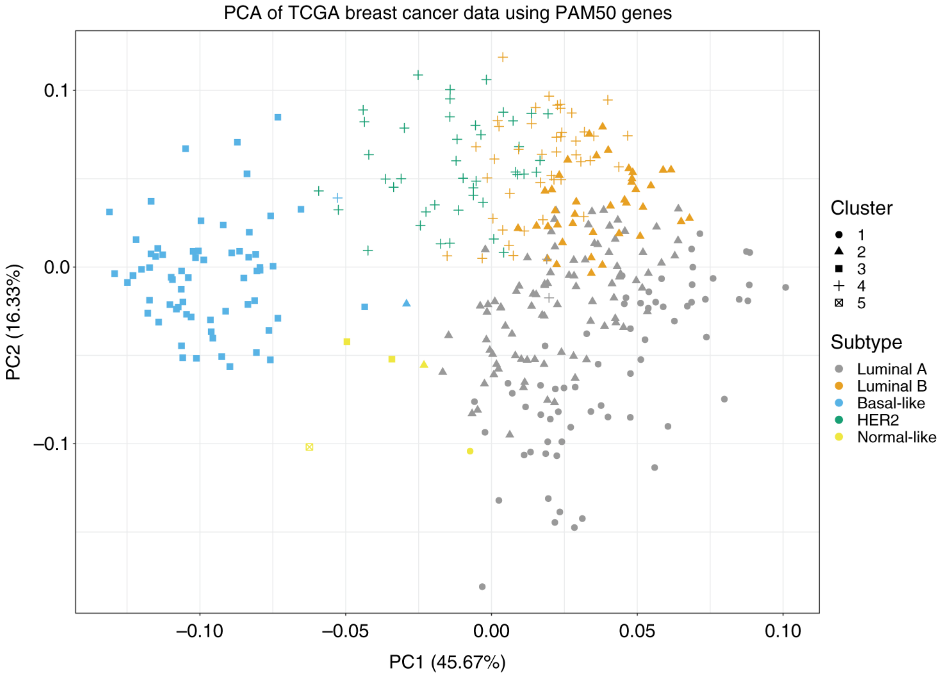

Figure 5: PCA plot on the TCGA breast cancer data using 48 of the PAM50 genes, where clusters produced by SUGSVarSel are indicated by shape and subtypes by colour.

The clusterings shown in Figure 4 and Figure 5 demonstrate that the variables we use for clustering are critically important. The two different pre-filtering choices led to results of varying quality and biological meaning. This is strong evidence in support of model-based variable selection rather than ad-hoc preprocessing.

5 Pan-cancer proteomic characterisation

In this section we apply our method to The Cancer Proteome Atlas (TCPA) datasets (Li et al., 2013; Akbani et al., 2014; Städler et al., 2017). The dataset contains a large number of tumours and cell line samples with protein expression levels generated using reverse-phase protein arrays (RPPAs). Our method allows us to perform a number of tasks on this data; in particular, for each cancer we can detect possible subgroups and the relevant proteins which discriminate these subgroups. We can also perform a pan-cancer analysis to explore the differences and similarities between cancers. Pan-cancer studies can unravel inter-cancer relationships which are important for developing new clinical targets (Weinstein et al., 2013; Uhlen et al., 2017; Berger et al., 2018; Hoadley et al., 2018). Recent pan-cancer analyses have suggested that cancers should be classified based on their molecular signatures rather than tissue of origin (Berger et al., 2018; Hoadley et al., 2018) and this motivates our analysis.

As is usual with this data there are irrelevant variables so methods that do not perform variable selection such as mclust and SUGS are ill-suited. Furthermore, there is little *a priori* knowledge about the number of clusters and so methods such as VarSelLCM and clustvarsel which require an exhaustive search of the number of clusters are inappropriate. To perform the analysis on all cancer sets would be prohibitively slow for the slowest of analysis methods.

The TCPA datasets contain data on 19 cancer types and the description of these cancers can be found in Supplementary Material. The total dataset consists of over 5000 tumour samples with only a few samples for some cancers and hundreds of samples for others and several hundreds of proteins. The merged PAN-Can 19 level 4 dataset is used in the following analysis, since it is appropriate for multiple disease analysis. More information about the data can be found here http://tcpaportal.org/tcpa/, where the data itself can also be downloaded. In addition, we standardise the expression levels for each protein so that they are zero-centred with unit variance.

The following table demonstrates the number of cases for each cancer type (Table 8).

Table 8: A table indicating the different cancer types and the number of observations from each of those cancers.

ACC	BLCA	BRCA	COAD	GBM	HNSC	KIRC	KIRP	LGG	LUAD
46	127	820	327	205	203	445	208	257	234
LUSC	**OV**	**PAAD**	**PRAD**	**READ**	**SKCM**	**STAD**	**THCA**	**UCEC**	
192	411	105	164	129	207	299	374	404	

We only keep proteins which have been measured on all cancers, which total 217 and so our dataset has a total of 5157 tumour samples with 217 variables. We apply SUGSVarSel to this data by first sub-sampling 10% of the variables 43 (a fifth of the total number of variables) times. Using the same priors as in previous analysis we analyses this data using the SUGSVarSel algorithm, running the algorithm for 50 random orderings, thus exploring a total of 2150 models. We summarise the BMA clustering using hierarchical clustering with average linkage. The summarised clustering contains 60 clusters, however many of these clusters contain only a few observations. Reassuringly there are 18 clusters with more than 20 observations and we focus on these for our analysis. A table summarising the clusters, along with results from hierarchical clustering, can be found in the Supplementary material, in Figure 6 is a heatmap of the clusterings:

In addition, we plot a heatmap of the data with the clustering produce by SUGSVarSel using only the proteins selected by the algorithm (Figure 7).

It is rare that a cancer associates with a single cluster, however there are evident relationships between cancers and clusters. Cluster A contains predominately womens' cancers (OV, UCEC, BRCA), while cluster B contains a large spread of cancers. Clusters C, E and F contain the cancers of the digestive tract (STAD, COAD and READ). Cluster D contains a subgroups of breast cancers (BRCA), while cluster G contains solely kidney cancer (KIRC). Clusters H and I contain cancers of the brain (LGG, GBM). Cluster J and P contain aero-digestive cancers (HNSC, LUAD LUSC). Thyroid cancer (THCA) is spread across clusters K, L and B, whilst KIRP is predominately found in cluster M. Pancreatic cancer (PAAD) is split across clusters N and B. Cluster O contains the majority of breast cancer patients. Prostate cancer (PRAD) is dominantly found in Q, while R forms a small cluster of stomach cancers. This is in line with other analyses performed on these data (Akbani et al., 2014; Hoadley et al., 2014; Şenbabaoğlu et al., 2016). A total of 147 proteins were selected as relevant for clustering.

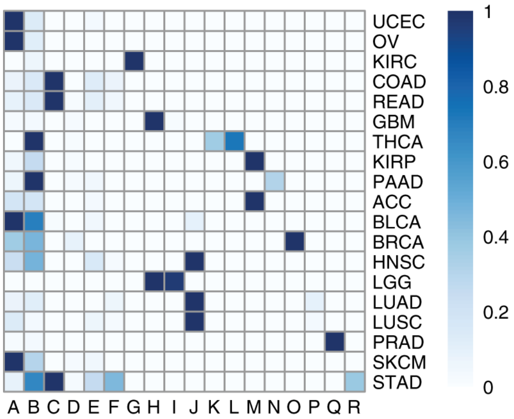

Figure 6: A heatmap indicating the correspondence between clusters produced by the SUGSVarSel algorithm and the different cancer types.

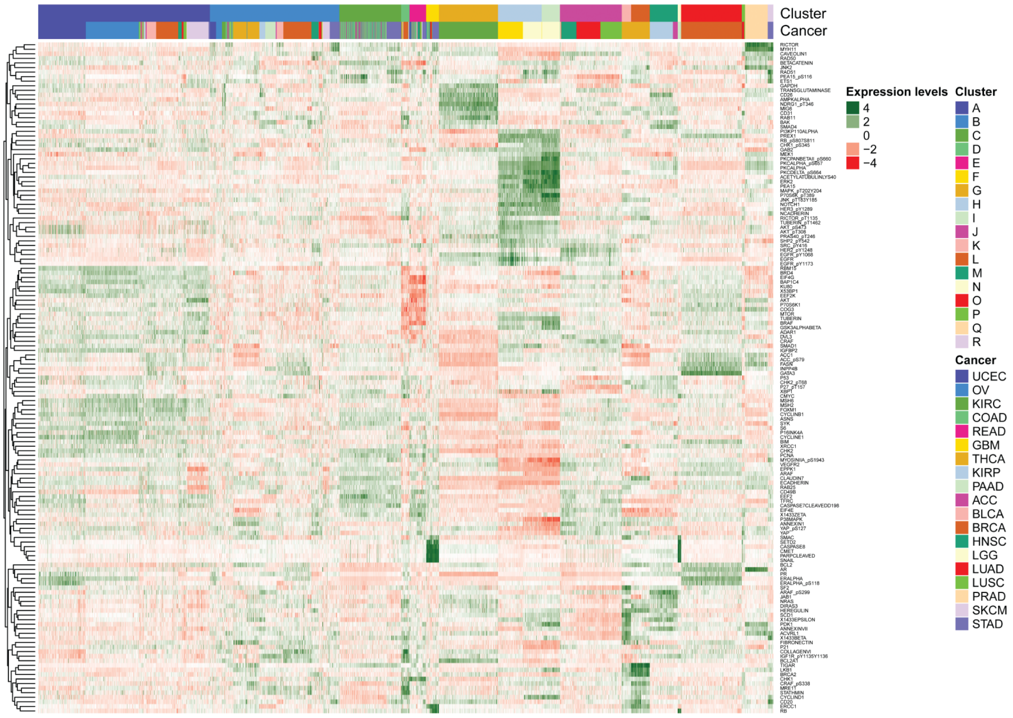

Figure 7: A heatmap of the expression data using the clustering produced by the SUGSVarSel algorithm applied to the pan-cancer TCPA dataset. The annotation bars on the top of plot indicate the different cancers and clusters.

We now consider an illustrative example. Figure 6 shows us that clusters K and L contain only thyroid cancers. It is of biological interest to see what drives the differences between these clusters as they could define clinically relevant thyroid subgroups. Considering only the 147 selected proteins, we plot the expression profile for the 20 proteins (Figure 8), with smallest *p*-value, which are significantly different between clusters K and L (T-test (Welch, 1947), $p < 0.00001$, using Benjamini-Hochberg correction (Benjamini & Hochberg, 1995)).

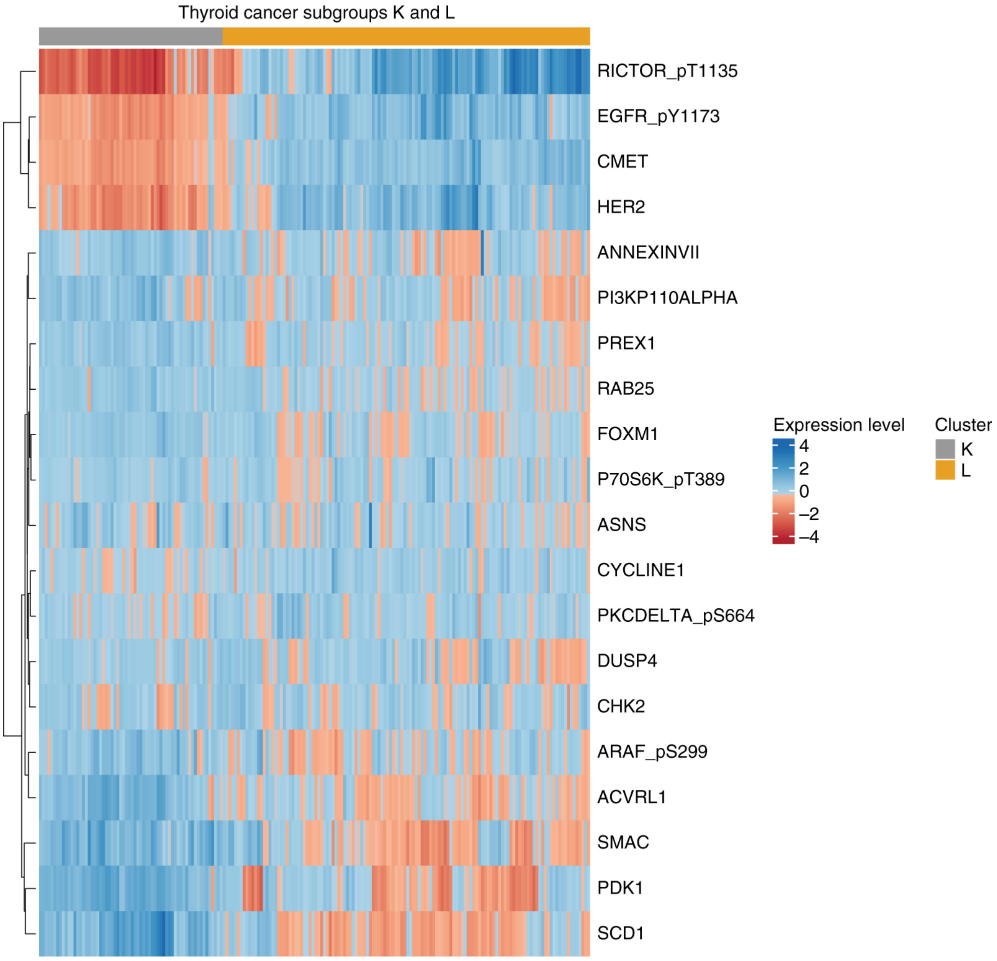

Figure 8: A heatmap of the expression TCPA data for the thyroid subgroups. We have plotted the expression for only the top 20 proteins which are significantly different between clusters K and L.

We do not observe an over representation of any of the thyroid cancers subtypes within each of these clusters (see Table 9).

Table 9: A table showing the distribution of 3 different THCA subtypes across the clusters K and L produce from the SUGSVarSel algorithm.

	K	**L**
Thyroid papillary carcinoma – classical/usual	31	72
Thyroid papillary carcinoma – follicular (>= 99% follicular patterned)	17	25
Thyroid papillary carcinoma – tall cell (>= 50% tall cell features)	2	6

Note that this information was not available for all patients.

6 Conclusion

In this article we presented SUGSVarSel, an extension to the SUGS algorithm of Wang and Dunson (2011) to allow variable selection. We demonstrated that when irrelevant variables are present the quality of the clustering can be degraded and clusters become more challenging to interpret. SUGSVarSel allows the flexibility of a Bayesian nonparametric approach but inference is considerably faster than using MCMC. Indeed, the SUGSVarSel algorithm infers the number of clusters automatically and performs inference for the Dirichlet process hyperparameter. This is in contrast to most clustering with variable selection methods which require a systematic search over the number of clusters.

Whilst our method is approximate it performs competitively with other commonly used approaches. Furthermore, we take advantage of exploring many models by performing Bayesian model averaging, which is important for exploring uncertainty in our clustering. We remark that model uncertainty and the application of BMA is rarely explored in clustering tasks. We have provided an R package to facilitate dissemination of our method utilising C++ to accelerate intensive computations and parallel processing features to make further computational gains

Application to two cancer transcriptomic datasets show the clear benefit of simultaneously performing variable selection and clustering. We demonstrate that variable selection improves interpretation of these datasets, providing the genes that drive the clustering structure of the data, as well as identifying those that are irrelevant for clustering. We further applied our method to a pan-cancer proteomic dataset for which none of the current model-based clustering and variable selection methods are suitable. SUGSVarSel is able to provide a characterisation of 5157 tumour samples, demonstrating clustering relationships across cancer types based on their molecular signature rather the tissue of origin.

There are a number of ways in which our proposed method could be extended. Firstly, our assumption that variables are conditionally independent given the cluster allocations might be unrealistic for some datasets. In such cases, more elaborate variable selection methods might be desirable, although this is likely to come at increased computational cost. Furthermore, we have assumed conjugacy throughout, so that the marginal likelihood in Equation (5) may be evaluated analytically. As noted in the original SUGS paper of Wang and Dunson (2011), one possible way to extend to non-conjugate cases would be to approximate this marginal likelihood, e.g. using a Laplace approximation.

Funding

Medical Research Council, Funder Id: http://dx.doi.org/10.13039/501100000265, Wellcome Trust Mathematical Genomics and Medicine student supported financially by the School of Clinical Medicine, University of Cambridge. Grant Number: MC_UU_00002/10, MC_UU_00002/13.

References

Akbani, R., P. K. S. Ng, H. M. J. Werner, M. Shahmoradgoli, F. Zhang, Z. Ju, W. Liu, J.-Y. Yang, K. Yoshihara, J. Li, S. Ling, E. G. Seviour, P. T. Ram, J. D. Minna, L. Diao, P. Tong, J. V. Heymach, S. M. Hill, F. Dondelinger, N. Städler, L. A. Byers, F. Meric-Bernstam, J. N. Weinstein, B. M. Broom, R. G. W. Verhaak, H. Liang, S. Mukherjee, Y. Lu and G. B. Mills (2014): "A pan-cancer proteomic perspective on The Cancer Genome Atlas." Nat. Commun., 5, 3887.

Antoniak, C. E. (1974): "Mixtures of dirichlet processes with applications to Bayesian nonparametric problems." Ann. Statist., 2, 1152–1174.

Attias, H. (1999): "Inferring parameters and structure of latent variable models by variational bayes." In: Proc. 15th Conf. on Uncertainty in Artificial Intelligence. San Francisco, CA, USA, Morgan Kaufmann Publishers Inc., pp. 21–30.

Attias, H. (2000): "A variational Bayesian framework for graphical models." In: Solla, S. A., Leen, T. K. Müller, K. editors, Advances in Neural Information Processing Systems 12. Denver, USA, MIT Press, pp. 209–215.

Benjamini, Y. and Y. Hochberg (1995): "Controlling the false discovery rate: a practical and powerful approach to multiple testing." J. Roy. Stat. Soc. B Met., 57, 289–300.

Berger, A. C., A. Korkut, R. S. Kanchi, A. M. Hegde, W. Lenoir, W. Liu, Y. Liu, H. Fan, H. Shen, V. Ravikumar, A. Rao, A. Schultz, X. Li, P. Sumazin, C. Williams, P. Mestdagh, P. H. Gunaratne, C. Yau, R. Bowlby, A. G. Robertson, D. G. Tiezzi, C. Wang, A. D. Cherniack, A. K. Godwin, N. M. Kuderer, J. S. Rader, R. E. Zuna, A. K. Sood, A. J. Lazar, A. I. Ojesina, C. Adebamowo, S. N. Adebamowo, K. A. Baggerly, T.-W. Chen, H.-S. Chiu, S. Lefever, L. Liu, K. MacKenzie, S. Orsulic, J. Roszik, C. S. Shelley, Q. Song, C. P. Vellano, N. Wentzensen, Cancer Genome Atlas Research Network, J. N. Weinstein, G. B. Mills, D. A. Levine and R. Akbani (2018): "A comprehensive pan-cancer molecular study of gynecologic and breast cancers." Cancer Cell, 33, 690–705.e9.

Blackwell, D. and J. B. MacQueen (1973): "Ferguson distributions via polya urn schemes." Ann. Statist., 1, 353–355.

Blei, D. M. and M. I. Jordan (2006): "Variational inference for Dirichlet process mixtures." Bayesian Anal., 1, 121–143.

Blei, D. M., A. Kucukelbir and J. D. McAuliffe (2016): "Variational inference: a review for statisticians." J. Am. Stat. Assoc., 112, 859–877.

Chen, A. H., Y.-W. Tsau and C.-H. Lin (2010): "Novel methods to identify biologically relevant genes for leukemia and prostate cancer from gene expression profiles." BMC Genomics, 11, 274.

Constantinopoulos, C., M. K. Titsias and A. Likas (2006): "Bayesian feature and model selection for Gaussian mixture models." IEEE Trans. Pattern Anal. Mach. Intell., 28, 1013–1018.

Cooke, E. J., R. S. Savage, P. D. W. Kirk, R. Darkins and D. L. Wild (2011): "Bayesian hierarchical clustering for microarray time series data with replicates and outlier measurements." BMC Bioinformatics, 12, 399.

Darkins, R., E. J. Cooke, Z. Ghahramani, P. D. W. Kirk, D. L. Wild and R. S. Savage (2013): "Accelerating Bayesian hierarchical clustering of time series data with a randomised algorithm." PLoS One, 8, e59795.

Daumé III, H. (2007): Fast search for Dirichlet process mixture models. In: Meila M., Shen, X. editors, AISTATS. San Juan, Puerto Rico, pp. 83–90.

Dudoit, S., J. Fridlyand and T. P. Speed (2002): "Comparison of discrimination methods for the classification of tumors using gene expression data." J. Am. Stat. Assoc., 97, 77–87.

Escobar, M. D. (1994): "Estimating normal means with a dirichlet process prior." J. Am. Stat. Assoc., 89, 268–277.

Escobar, M. D. and M. West (1995): "Bayesian density estimation and inference using mixtures." J. Am. Stat. Assoc., 90, 577–588.

Ferguson, T. S. (1973): "A Bayesian analysis of some nonparametric problems." Ann. Statist., 1, 209–230.

Ferguson, T. S. (1974): "Prior distributions on spaces of probability measures." Ann. Statist., 2, 615–629.

Fop, M. and T. B. Murphy (2018): "Variable selection methods for model-based clustering." Stat. Surv., 12, 1–48.

Fraley, C. and A. E. Raftery (2002): "Model-based clustering, discriminant analysis and density estimation." J. Am. Stat. Assoc., 97, 611–631.

Fraley, C., A. E. Raftery, T. B. Murphy and L. Scrucca (2012). mclust Version 4 for R: normal mixture modeling for model-based clustering, classification, and density estimation.

Fritsch, A. and K. Ickstadt (2009): "Improved criteria for clustering based on the posterior similarity matrix." Bayesian Anal., 4, 367–391.

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

Heller, K. and Z. Ghahramani (2005): "Bayesian hierarchical clustering." In: Proceedings of the 22nd International Conference on Machine Learning. Bonn, Germany.

Hoadley, K. A., C. Yau, D. M. Wolf, A. D. Cherniack, D. Tamborero, S. Ng, M. D. Leiserson, B. Niu, M. D. McLellan, V. Uzunangelov, J. Zhang, C. Kandoth, R. Akbani, H. Shen, L. Omberg, A. Chu, A. A. Margolin, L. J. Van't Veer, N. Lopez-Bigas, P. W. Laird, B. J. Raphael, L. Ding, A. G. Robertson, L. A. Byers, G. B. Mills, J. N. Weinstein, C. Van Waes, Z. Chen, E. A. Collisson, Cancer Genome Atlas Research Network, C. C. Benz, C. M. Perou, J. M. Stuart (2014): "Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin." Cell, 158, 929–944.

Hoadley, K. A., C. Yau, T. Hinoue, D. M. Wolf, A. J. Lazar, E. Drill, R. Shen, A. M. Taylor, A. D. Cherniack, V. Thorsson, R. Akbani, R. Bowlby, C. K. Wong, M. Wiznerowicz, F. Sanchez-Vega, A. G. Robertson, B. G. Schneider, M. S. Lawrence, H. Noushmehr, T. M. Malta, Cancer Genome Atlas Network, J. M. Stuart, C. C. Benz and P. W. Laird (2018): "Cell-of-origin patterns dominate the molecular classification of 10,000 tumors from 33 types of cancer." Cell, 173, 291–304.

Hoeting, J. A., D. Madigan, A. E. Raftery and C. T. Volinsky (1999): "Bayesian model averaging: a tutorial." Statist. Sci., 14, 382–417.

Hubert, L. and P. Arabie (1985): "Comparing partitions." Journal of Classification, 2, 193–218.

Jain, S. and R. M. Neal (2004): "A split-merge markov chain monte carlo procedure for the dirichlet process mixture model." J. Comput. Graph. Stat., 13, 158–182.

Jiang, K., B. Kulis and M. I. Jordan (2012): "Small-variance asymptotics for exponential family dirichlet process mixture models." In: Advances in Neural Information Processing Systems 25. Lake Tahoe, Nevada.

Jiang, L., Y. Dong, N. Chen and T. Chen (2016): "DACE: a scalable DP-means algorithm for clustering extremely large sequence data." Bioinformatics, 33, 834–842.

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

Kuett, A., C. Rieger, D. Perathoner, T. Herold, M. Wagner, S. Sironi, K. Sotlar, H.-P. Horny, C. Deniffel, H. Drolle and M. Fiegl (2015): "Il-8 as mediator in the microenvironment-leukaemia network in acute myeloid leukaemia." Sci. Rep., 5, 18411.

Kulis, B. and M. I. Jordan (2012): "Revisiting k-means: new algorithms via Bayesian nonparametrics." In: International Conference on Machine Learning.

Law, M. H. C., M. A. T. Figueiredo and A. K. Jain (2004): "Simultaneous feature selection and clustering using mixture models." IEEE Trans. Pattern Anal. Mach. Intell., 26, 1154–1166.

Li, J., Y. Lu, R. Akbani, Z. Ju, P. L. Roebuck, W. Liu, J.-Y. Yang, B. M. Broom, R. G. Verhaak, D. W. Kane, C. Wakefield, J. N Weinstein, G. B. Mills and H. Liang (2013): "TCPA: a resource for cancer functional proteomics data." Nat. Methods, 10, 1046–1047.

Liverani, S., D. I. Hastie, L. Azizi, M. Papathomas and S. Richardson (2015): "PReMiuM: An R package for profile regression mixture models using Dirichlet processes." J. Stat. Softw., 64, 1.

Lo, A. Y. (1984): "On a class of Bayesian nonparametric estimates: i. density estimates." Ann. Statist., 12, 351–357.

Lock, E. F. and D. B. Dunson (2013): "Bayesian consensus clustering." Bioinformatics, 29, 2610–2616.

Madigan, D. and A. E. Raftery (1994): "Model selection and accounting for model uncertainty in graphical models using Occam's window." J. Am. Stat. Assoc., 89, 1535–1546.

Marbac, M. and M. Sedki (2017): "Variable selection for model-based clustering using the integrated complete-data likelihood." Stat. Comput., 27, 1049–1063.

Marbac, M. and M. Sedki (2018): "VarSelLCM: an R/C++ package for variable selection in model-based clustering of mixed-data with missing values." Bioinformatics, 35, 1255–1257.

Maugis, C., G. Celeux and M.-L. Martin-Magniette (2009): "Variable selection for clustering with gaussian mixture models." Biometrics, 65, 701–709.

Medvedovic, M., K. Y. Yeung and R. E. Bumgarner (2004): "Bayesian mixture model based clustering of replicated microarray data." Bioinformatics, 20, 1222–1232.

Natsuka, S., S. Akira, Y. Nishio, S. Hashimoto, T. Sugita, H. Isshiki and T. Kishimoto (1992): "Macrophage differentiation-specific expression of NF-IL6, a transcription factor for interleukin-6." Blood, 79, 460–466.

Neal, R. M. (2000): "Markov chain sampling methods for dirichlet process mixture models." J. Comput. Graph. Stat., 9, 249–265.

Network, C. G. A. (2012): "Comprehensive molecular portraits of human breast tumours." Nature, 490, 61–70.

Parker, J. S., M. Mullins, M. C. Cheang, S. Leung, D. Voduc, T. Vickery, S. Davies, C. Fauron, X. He, Z. Hu, J. F. Quackenbush, I. J. Stijleman, J. Palazzo, J. S. Marron, A. B. Nobel, E. Mardis, T. O. Nielsen, M. J. Ellis, C. M. Perou and P. S. Bernard (2009): "Supervised risk predictor of breast cancer based on intrinsic subtypes." J. Clin. Oncol., 27, 1160–1167.

Pekarsky, Y., C. Hallas and C. M. Croce (2001): "The role of TCL1 in human T-cell leukemia." Oncogene, 20, 5638.

Raftery, A. E. and N. Dean (2006): "Variable selection for model-based clustering." J. Am. Stat. Assoc., 101, 168–178.

Rand, W. M. (1971): "Objective criteria for the evaluation of clustering methods." J. Am. Stat. Assoc., 66, 846–850.

Rasmussen, C. E. (2000): "The infinite gaussian mixture model." In: Advances in Neural Information Processing Systems 12, Denver, USA, volume 12, pp. 554–560.

Raykov, Y. P., A. Boukouvalas and M. A. Little (2016a): "Simple approximate MAP inference for Dirichlet processes mixtures." Electron. J. Statist., 10, 3548–3578.

Raykov, Y. P., A. Boukouvalas, F. Baig and M. A. Little (2016b): "What to do when k-means clustering fails: a simple yet principled alternative algorithm." PLoS One, 11, e0162259.

Russell, N., T. B. Murphy and A. E. Raftery (2015): "Bayesian model averaging in model-based clustering and density estimation." arXiv preprint arXiv:1506.09035.

Savage, R. S., K. Heller, Y. Xu, Z. Ghahramani, W. M. Truman, M. Grant, K. J. Denby and D. L. Wild (2009): "R/BHC: fast Bayesian hierarchical clustering for microarray data." BMC Bioinformatics, 10, 242.

Schwarz, G. (1978): "Estimating the dimension of a model." Ann. Statist., 6, 461–464.

Scrucca, L. and A. E. Raftery (2014): "clustvarsel: a package implementing variable selection for model-based clustering in R." J. Stat. Softw., 84, 1–28.

Scrucca, L., M. Fop, T. B. Murphy and A. E. Raftery (2016): "mclust 5: clustering, classification and density estimation using Gaussian finite mixture models." R J, 8, 205–233.

Şenbabaoğlu, Y., S. O. Sümer, F. Sánchez-Vega, D. Bemis, G. Ciriello, N. Schultz and C. Sander (2016): "A multi-method approach for proteomic network inference in 11 human cancers." PLoS Comput. Biol., 12, e1004765.

Shochat, C., N. Tal, O. R. Bandapalli, C. Palmi, I. Ganmore, G. Te Kronnie, G. Cario, G. Cazzaniga, A. E. Kulozik, M. Stanulla, M. Schrappe, A. Biondi, G. Basso, D. Bercovich, M. U. Muckenthaler, S. Izraeli (2011): "Gain-of-function mutations in interleukin-7 receptor-α (IL7R) in childhood acute lymphoblastic leukemias." J. Exp. Med., 208, 901–908.

Städler, N., F. Dondelinger, S. M. Hill, R. Akbani, Y. Lu, G. B. Mills and S. Mukherjee (2017): "Molecular heterogeneity at the network level: high-dimensional testing, clustering and a TCGA case study." Bioinformatics, 33, 2890–2896.

Tadesse, M. G., N. Sha and M. Vannucci (2005): "Bayesian variable selection in clustering high-dimensional data." J. Am. Stat. Assoc., 100, 602–617.

Teh, Y. W., M. I. Jordan, M. J. Beal and D. M. Blei (2006): "Hierarchical dirichlet processes." J. Am. Stat. Assoc., 101, 1566–1581.

Uhlen, M., C. Zhang, S. Lee, E. Sjöstedt, L. Fagerberg, G. Bidkhori, R. Benfeitas, M. Arif, Z. Liu, F. Edfors, K. Sanli, K. von Feilitzen, P. Oksvold, E. Lundberg, S. Hober, P. Nilsson, J. Mattsson, J. M. Schwenk, H. Brunnström, B. Glimelius, T. Sjöblom, P. H. Edqvist, D. Djureinovic, P. Micke, C. Lindskog, A. Mardinoglu and F. Ponten (2017): "A pathology atlas of the human cancer transcriptome." Science, 357, eaan2507.

Van der Velden, V., M. Brüggemann, P. Hoogeveen, M. de Bie, P. Hart, T. Raff, H. Pfeifer, S. Lüschen, T. Szczepański, E. Van Wering, M. Kneba and J. J. van Dongen (2004): "TCRB gene rearrangements in childhood and adult precursor-B-ALL: frequency, applicability as MRD-PCR target, and stability between diagnosis and relapse." Leukemia, 18, 1971.

Wang, L. and D. B. Dunson (2011): "Fast Bayesian inference in dirichlet process mixture models." J. Comput. Graph. Stat., 20, 196–216.

Weinstein, J. N., E. A. Collisson, G. B. Mills, K. R. M. Shaw, B. A. Ozenberger, K. Ellrott, I. Shmulevich, C. Sander, J. M. Stuart, Cancer Genome Atlas Research Network (2013): "The cancer genome atlas pan-cancer analysis project." Nat. Genet., 45, 1113–1120.

Welch, B. L. (1947): "The generalization of 'student's' problem when several different population variances are involved." Biometrika, 34, 28–35.

Witten, D. M. and R. Tibshirani (2010): "A framework for feature selection in clustering." J. Am. Stat. Assoc., 105, 713–726.

Zhang, X., D. J. Nott, C. Yau and A. Jasra (2014): "A sequential algorithm for fast fitting of dirichlet process mixture models." J. Comput. Graph. Stat., 23, 1143–1162.

Supplementary Material: The online version of this article offers supplementary material (DOI: https://doi.org/10.1515/sagmb-2018-0065).