Prior specifications for Dirichlet Process Mixture Models with Gaussian kernels

ISBA - 2022

Michail Papathomas (with Wei Jing and Silvia Liverani)

School of Mathematics and Statistics, University of St Andrews
Queen Mary University, London

01/07/2022

High-dimensional data

In modern epidemiology and genetics, cohort studies create high dimensional databases that include measurements of **tens or hundreds of**

- anthropometric
- lifestyle
- dietary
- environmental exposure

risk factors, as well as **thousands or millions** of genetic-epigenetic-omics variables.

Analysing highly structured data

- It is of great importance to public health and public policy to **understand the effect of risk factors to phenotypes such as cancer** and cardiovascular diseases.
- It is now understood that,
  - risk factors may combine to affect the probability of disease
  - the effect of a particular covariate may only be important in the presence of other covariates.
- Typically, we aim to make inferences and measure our uncertainty on how the risk factors combine to affect the probability of disease.

Problems with Linear modelling for detecting interactions

Linear modelling is the most common approach for detecting interactions between covariates, but...

- In a classical setting, fitting linear models with many parameters sometimes **requires an impractically large vector of observations** for valid inferences (Burton et al., IJE, 2009). Also, **identifiability and collinearity** problems are often present.
- In Bayesian model comparison, the space of models becomes vast, and model search algorithms like the Reversible Jump approach (Green, Bka 1995) require **an impractically large number of iterations** before they converge (Dobra and Massam, St Meth 2010).

Profile regression

Rather than introducing main effects + 2-way + 3-way interactions etc., profile regression is a **Bayesian approach** that **reduces dimensionality** by using as its main unit of inference the **subject's profile**,
e.g. (lives on main road, high exposure of PM10 and NO_2, low physical activity, overweight, genetic marker)

- The subjects are clustered using their **covariate profiles** and **disease status** (with Dirichlet process mixture modelling).

This modelling (usually) reveals a population structure of **high and low risk sub-populations (groups)**.

Notation

For individual i

y_i	outcome of interest
$\mathbf{x}_i = (x_{i1}, \ldots, x_{iP})$	covariate profile
$\mathbf{w}_i$	fixed effects
$z_i = c$	the allocation variable indicates the cluster to which individual i belongs

- Mixture model for the covariates

$$f(\mathbf{x}_i|\theta, \psi) = \sum_c \psi_c f(\mathbf{x}_i|z_i = c, \theta_c)$$

Statistical Framework

- Joint covariate and response model

$$f(\mathbf{x}_i, y_i|\phi, \theta, \psi, \beta) = \sum_c \psi_c f(\mathbf{x}_i|z_i = c, \phi_c) f(y_i|z_i = c, \theta_c, \beta, \mathbf{w}_i)$$

- For example, for Bernoulli outcome

$$\text{logit}\{p(y_i = 1|\theta_c, \beta, \mathbf{w}_i)\} = \theta_c + \beta^T \mathbf{w}_i$$

- The association of the profiles with the response are characterised by the risk effect parameters θ_c
- The above, adopted in Molitor et al. (Biostatistics, 2010), is similar to Bigelow and Dunson (JASA, 2009).

Implementation

We have implemented profile regression in C++, wrapped in the R package PReMiuM (Liverani et al. 2015) for

- binary, binomial, categorical, Normal and Poisson outcome
- Normal and discrete covariates

It can do

- dependent or independent slice sampling (Kalli et al., 2011)
- or truncated Dirichlet process model (Ishwaran and James, 2001)

as well as

- variable selection
- handles missing data
  - we check which cluster the subject is allocated to and then sample
- includes some label switching moves

Clustering with the Dirichlet Process (DP)

- The DP allows for flexible Bayesian clustering and the evaluation of uncertainty for the clustering (i.e. the number of clusters, their composition, and the cluster specific parameters.)
- Extensively used; epidemiology, molecular biology, statistical ecology, social sciences, etc.

The DP clustering model

- The DP (infinite) mixture model is written as,

$$x_i|\theta_i \sim F(\theta_i).$$

$$\theta_i|G \sim G,$$

$$G|G_0, \alpha \sim DP(\alpha, G_0),$$

- Stick breaking representation (Ishwaran+James, 2001),

$$x_i|\boldsymbol{\theta}^*, \boldsymbol{z} \sim F(\theta^*_{z_i}), \qquad i = 1, ..., n$$

$$\theta^*_c|G_0 \sim G_0, \qquad c = 1, 2, ...$$

$$z_i|\psi \sim \sum_{c=1}^{\infty} \psi_c \delta_c(.)$$

$$\psi_c = v_c \prod_{j=1}^{c-1}(1 - v_j),\ v_c|\alpha \sim Beta(1, \alpha).$$

Problems of mixture models for multivariate Gaussian kernels

- My work mostly involves clustering with categorical observations.
- Supervised a student that used Profile regression to model the relation between continuous metabolite observations (from 375 metabolite features identified in that population) and Breast cancer (EPIC study into cancer.)
- Results were not interpretable or according to expectations.
- In the end, the observations were categorised into tertiles, and more interpretable results were obtained.
- Why did the mixture model work in a less satisfactory manner with the continuous observations?

The 'basic' approach

- For datasets with continuous variables, the standard choice for the mixture components is the multivariate normal,

$$f(X_i|\boldsymbol{\theta}_{z_i}) = (2\pi)^{-\frac{J}{2}}|\Sigma_{z_i}|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(X_i-\mu_{z_i})^T\Sigma_{z_i}^{-1}(X_i-\mu_{z_i})\right\}, \quad (1)$$

  where μ_{z_i} and Σ_{z_i} are the mean vector and the covariance matrix for cluster z_i respectively. In this case, $\boldsymbol{\theta}_{z_i} = (\mu_{z_i}, \Sigma_{z_i})$.
- One commonly used setting for the base distribution G_0 is,

$$G_0 = N_J(\mu_c; \mu_0, \Sigma_0) \times InvWishart_J(\Sigma_c; R_0, \kappa_0). \quad (2)$$

- However, when number of dimensions goes beyond, say, 3 or 4, the clustering is likely to fail.

Shortcomings of the Inverse Wishart distribution

- Allocates little mass for variances near zero; Gelman (2006).
- Only one degree of freedom (κ_0) to model the variability and dependence between all matrix elements. Translates to large dependency so that the larger the variances, the larger the correlations in absolute value; O'Hagan (1994).
- Sensitive to the choice of the hyperparameter values Hennig (2015).
- For mixture modelling, the IW can be detrimental to obtaining sensible inferences.

The Hierarchical IW

- The main idea is to put hyperpriors on R_0 and κ_0 so that those hyperparameters can be estimated from the data.
- The prior of Σ_c is (referred to as HIW1),

$$\begin{aligned} \Sigma_c | R_0, \kappa_0 &\sim \mathit{InvWishart}_J(\Sigma_c; R_0, \kappa_0) \\ R_0 | R_1^{-1}, \kappa_1 &\sim \mathit{Wishart}_J(R_0; R_1^{-1}, \kappa_1) \\ \kappa_0 - J | \alpha_{\kappa_0}, \beta_{\kappa_0} &\sim \mathit{InvGamma}(\kappa_0 - J; \alpha_{\kappa_0}, \beta_{\kappa_0}). \end{aligned} \tag{3}$$

- A slightly different specification (referred to as HIW2) is,

$$\begin{aligned} \Sigma_c &\sim \mathit{InvWishart}_J\left(\Sigma_c; \epsilon_0 + J - 1; 2\epsilon_0 \mathit{diag}\left(\frac{1}{\delta_1}, ..., \frac{1}{\delta_J}\right)\right) \\ \delta_j &\sim \mathit{InvGamma}(\delta_j; \alpha_\delta, g_j) \\ (\epsilon_0 - 1) &\sim \mathit{InvGamma}(\epsilon_0 - 1; \alpha_{\epsilon_0}, \beta_{\epsilon_0}). \end{aligned} \tag{4}$$

- Implemented in 'Dppackage' in R (Jara et al. 2011).
- Added flexibility only rectifies shortcomings to some extent.

The separation prior

- Relies on the decomposition so that,

$$\Sigma = SRS, \tag{5}$$

where $S = \{s_{j,j}\}$ is diagonal with st. deviations as diagonal elements and $R = \{r_{i,j}\}$ is a correlation matrix.

- Priors can be specified separately for S and R.
- Use a similar specification to O'Malley and Zaslavsky (2008).

$$\begin{aligned} R_c &\sim \mathit{InvWishart}_J(\kappa_R, R_R^{-1}) \\ (\kappa_R - J) &\sim \mathit{InvGamma}(\kappa_R - J; \alpha_{\kappa_R}, \beta_{\kappa_R}) \\ s_{c,j} &\sim \Gamma^{-1}(s_{c,j}; \alpha_s, \beta_{sj}). \\ \beta_{sj} &\sim \Gamma(\beta_{sj}; \alpha_0, \beta_{0j}). \\ \Sigma_c &= S_c R_c S_c, \end{aligned} \tag{6}$$

where $S_c = \{s_{c,j}\}$ is a diagonal matrix for cluster c and $R_c = \{r_{c,i,j}\}$ is a dense matrix for cluster c.

The log prior

- Relies on the log transformation $A = \log(\Sigma)$, and the spectral decomposition on Σ, $\Sigma = EDE^T$, where D is a diagonal matrix with eigenvalues of Σ as its diagonal elements and E is a orthonormal matrix. The columns of E are normalized eigenvectors, and correspond to the eigenvalues of D.
- A can be decomposed as $A = E \log(D) E^T$, where $\log(D)$ is a diagonal matrix with the logarithm of the eigenvalues of Σ as its diagonal elements.
- Therefore, as long as A is sampled as a symmetric matrix, $\exp(A) = \Sigma$ must be positive definite.
- Approximation for the full conditional of the elements of A, $\boldsymbol{a}_c$ requires that μ_c is known (Leonard and Hsu, 92). Then, the full conditional is approximately Normal; great for a sampler, but approximation may not be good for a mixture model.
- Large number of computations required.

The Sparse prior (1)

- Consider placing the Laplace prior on the off-diagonal elements and the exponential prior on the diagonal elements of the precision matrix Wang (2012), Khondker (2013),

$$\begin{aligned} \boldsymbol{X}_i | T &\sim N_J(\mathbf{0}, T^{-1}); \\ T_{ij} &\sim Laplace(0, 1/m_{0,ij}) \; i \neq j, \\ T_{ii} &\sim Exp(m_{0,ii}), \end{aligned} \tag{7}$$

  where T denotes the precision matrix and the hyperparameter $m_{0,ij}$ can be interpreted as the shrinkage parameter that control the amount of shrinkage of T_{ij} towards zero.

- Note that the $m_{0,ij}$ can be arranged into a matrix denoted by $M_0 = \{m_{0,ij}\}$.

The Sparse prior (2)

- As the values in M_0 largely influence the distributions of T and Σ, we would like to place hyperpriors on them.
- The hyperprior we propose for M_0 is similar to Wang (2012b), however, we also place hyperpriors on the diagonal elements of M_0.

$$p(M_0) \propto C \prod_{i<j} \left\{ \Gamma(m_{0,ij} | \alpha_{m_0}, \beta_{m_0}) \right\} \prod_{i} \left\{ \Gamma(m_{0,ii} | \alpha_{m_i}, \beta_{m_i}) \right\}, \tag{8}$$

where C is a constant of proportionality. (See Jing et al. (2022, arXiv) for more details.)

Two simulations - Dense variance matrices

Table: Simulated data III specifications. $J = 20$. $rep(a, b)$ represents a vector of length b with repeated elements a.

Cluster	μ_c^{true}	Var_c^{true}	ρ_c^{true}	n_c
1	(rep(3,5),rep(12,5),rep(18,5),rep(12,5))	3	0.2	200
2	(rep(12,5),rep(18,5),rep(3,5),rep(18,5))	3	0.5	200
3	(rep(18,5),rep(18,5),rep(12,5),rep(8,5))	3	0.3	200
4	(rep(18,5),rep(3,5),rep(8,5),rep(3,5))	3	0.1	200
5	(rep(8,5),rep(8,5),rep(12,5),rep(3,5))	3	0.7	200

Table: Simulated data IV specifications. $J = 20$.

Cluster	μ_c^{true}	Var_c^{true}	ρ_c^{true}	n_c
1	(rep(3,5),rep(12,5),rep(18,5),rep(12,5))	9	0.2	200
2	(rep(12,5),rep(18,5),rep(3,5),rep(18,5))	9	0.5	200
3	(rep(18,5),rep(18,5),rep(12,5)rep(8,5))	9	0.3	200
4	(rep(18,5),rep(3,5),rep(8,5),rep(3,5))	9	0.1	200
5	(rep(8,5),rep(8,5),rep(12,5),rep(3,5))	9	0.7	200

Simulation results - Run times

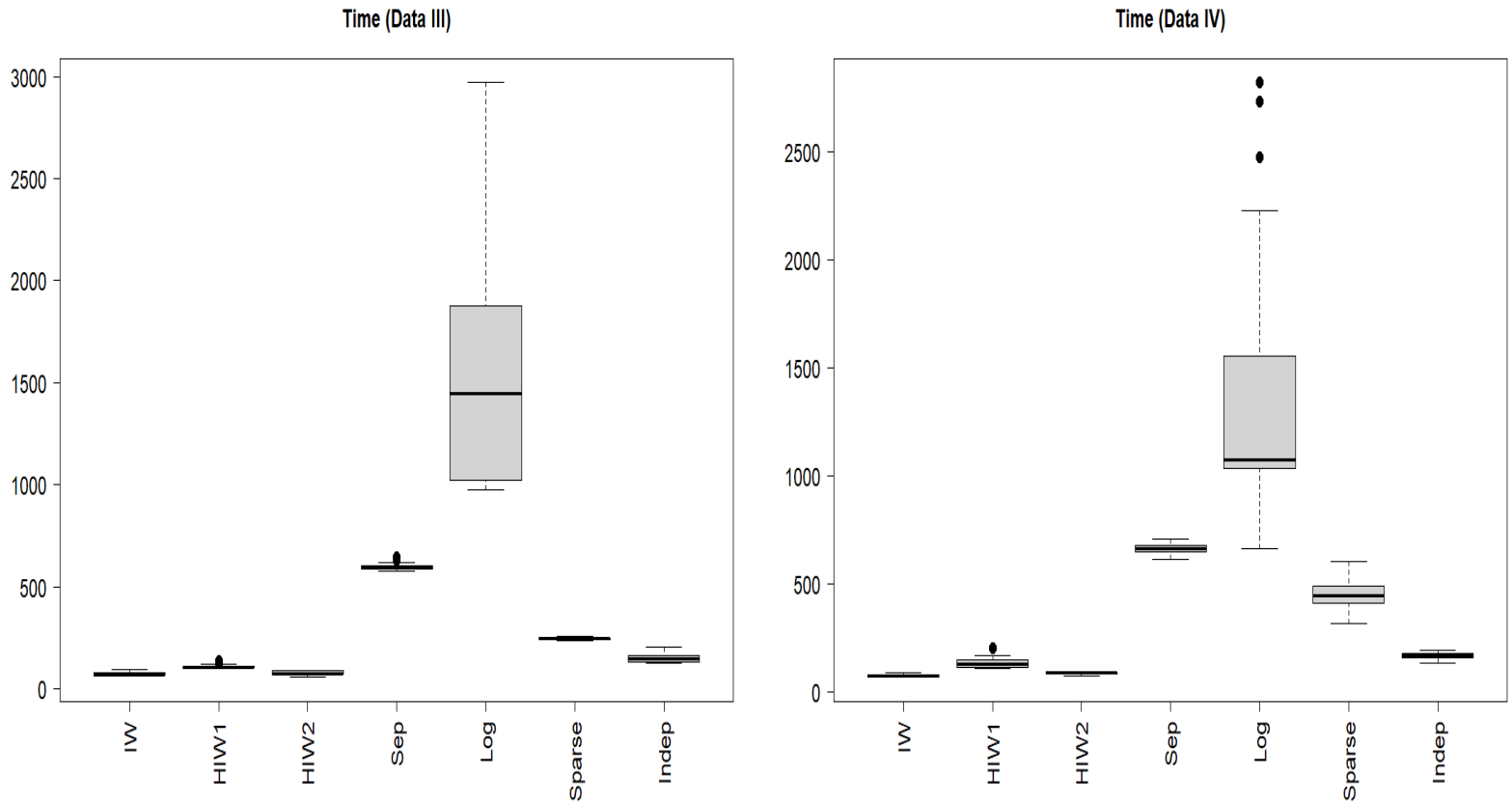

Figure: Computational times for 20 datasets according to Data III specifications (left) and Data IV (right) in seconds.

Simulation results - Rand index

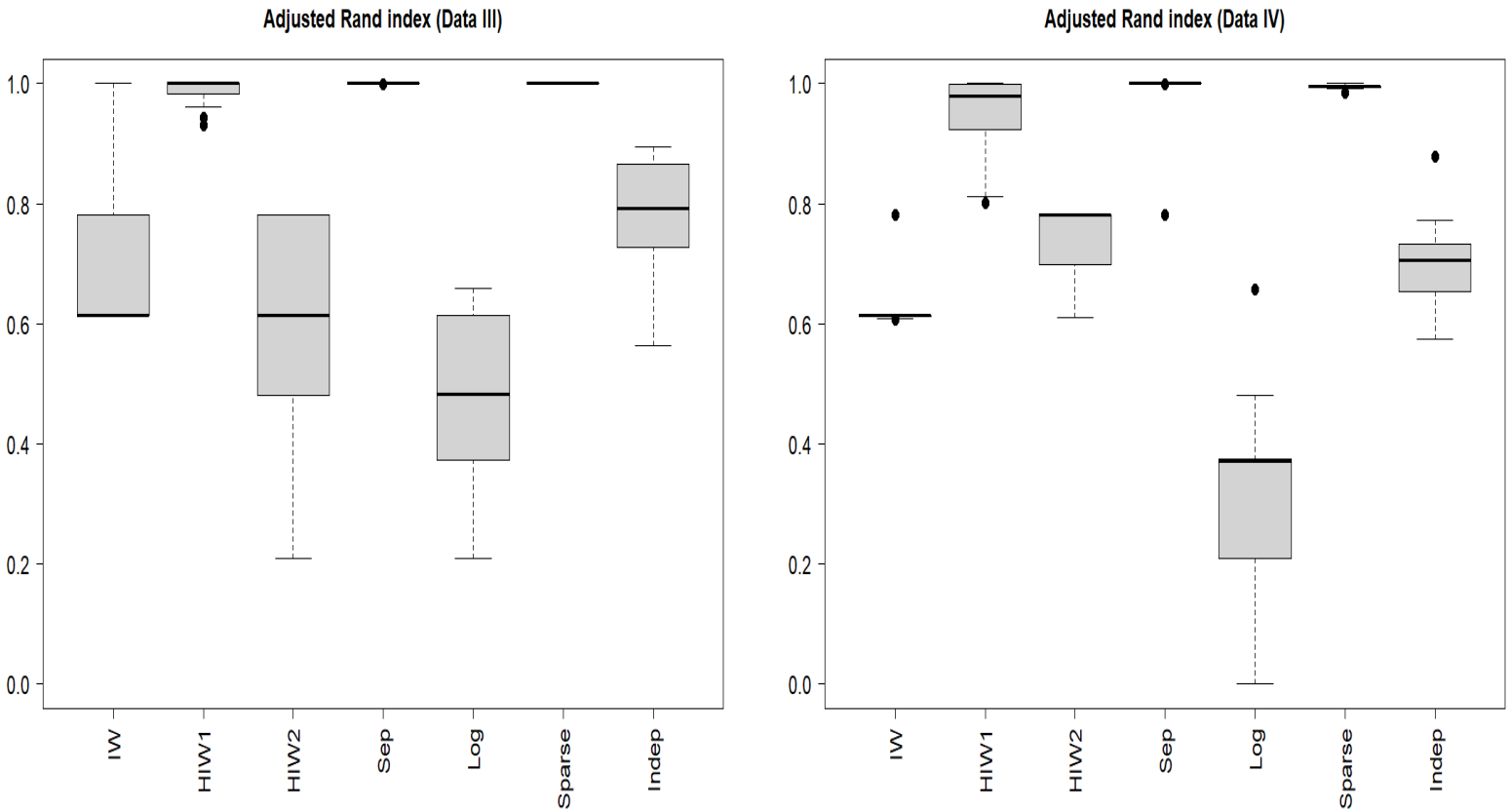

Figure: Boxplots of adjusted rand indices. Datasets simulated according to Data III specifications (left) and Data IV (right).

Simulation results - Simulation IV

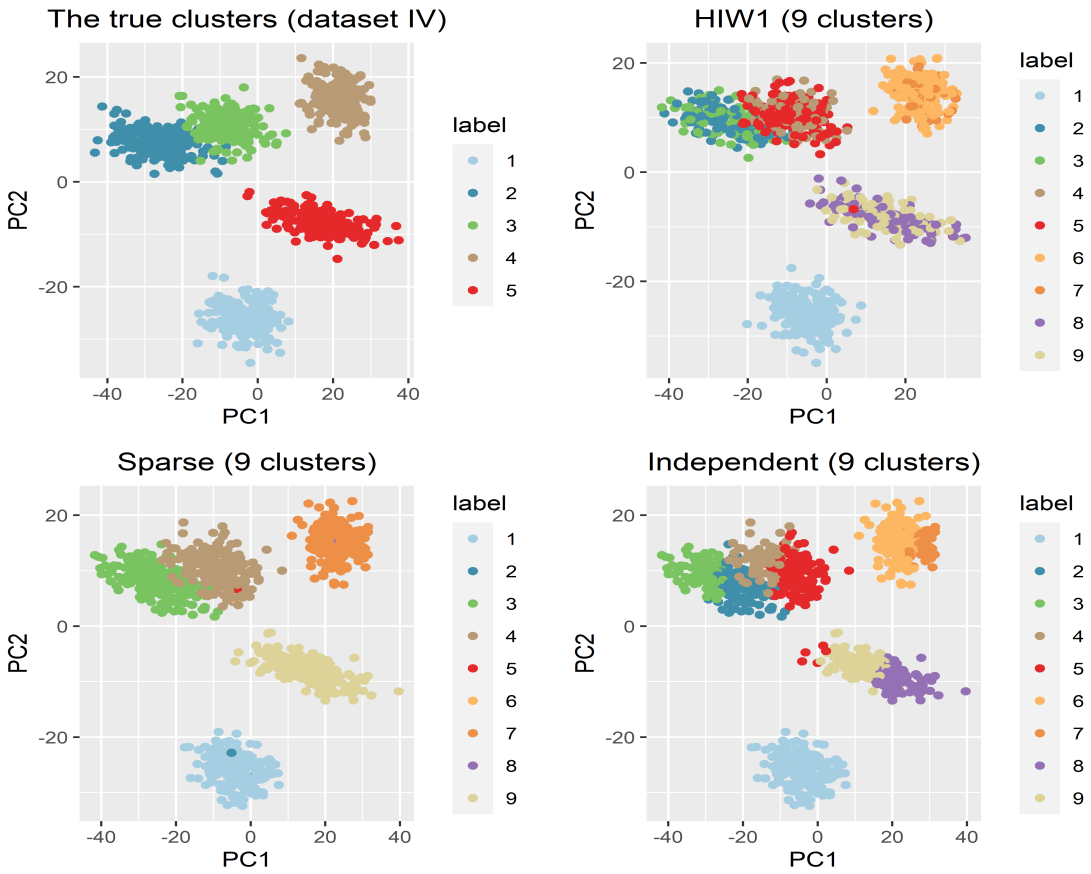

Figure: Reduced space plots of the true partition and one representative partition for the HIW1, sparse and independent priors for simulated data IV.

Two simulations - Sparse variance matrices

Table: Simulated data V specifications. $J = 20$. $rep(a, b)$ represents a vector of length b with repeated elements a.

Cluster	μ_c^{true}	Var_c^{true}	ρ_c^{true}	n_c
1	(rep(3,5),rep(12,5),rep(18,5),rep(12,5))	3	0.7	200
2	(rep(12,5),rep(18,5),rep(3,5),rep(18,5))	3	0.7	200
3	(rep(18,5),rep(18,5),rep(12,5),rep(8,5))	3	0.7	200
4	(rep(18,5),rep(3,5),rep(8,5),rep(3,5))	3	0.7	200
5	(rep(8,5),rep(8,5),rep(12,5),rep(3,5))	3	0.7	200

Table: Simulated data VI specifications. $J = 20$.

Cluster	μ_c^{true}	Var_c^{true}	ρ_c^{true}	n_c
1	(rep(3,5),rep(12,5),rep(18,5),rep(12,5))	9	0.7	200
2	(rep(12,5),rep(18,5),rep(3,5),rep(18,5))	9	0.7	200
3	(rep(18,5),rep(18,5),rep(12,5)rep(8,5))	9	0.7	200
4	(rep(18,5),rep(3,5),rep(8,5),rep(3,5))	9	0.7	200
5	(rep(8,5),rep(8,5),rep(12,5),rep(3,5))	9	0.7	200

Simulation results - Rand index

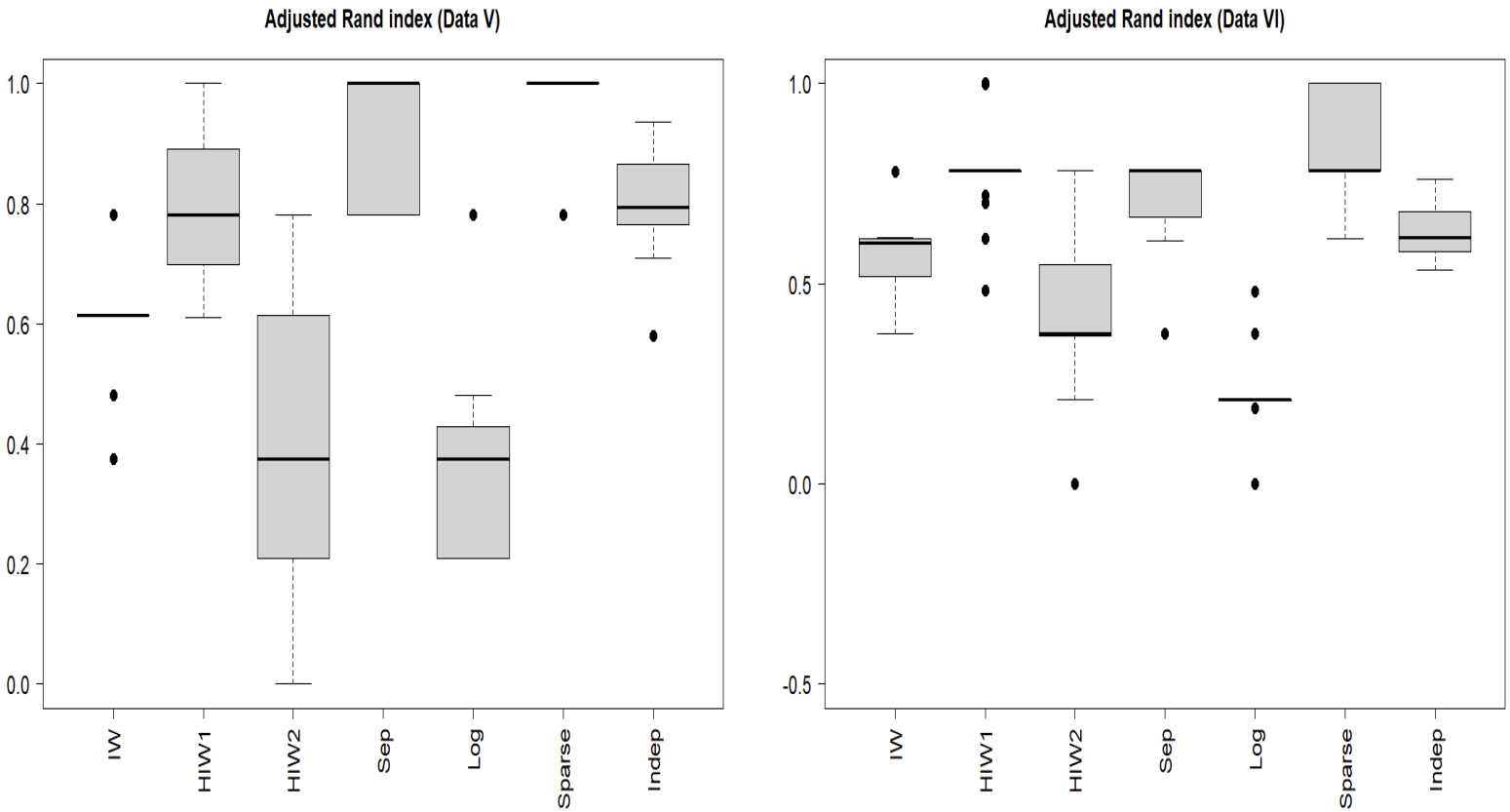

Figure: Boxplots of the adjusted rand indices. Datasets simulated according to data V (left), and according to data VI (right).

RNA-Seq PAN-cancer dataset

- To assess the performance of the different priors, we analyse a real data set where true labels are known.
- The dataset is part of the RNA-Seq PAN-cancer dataset.
- The covariates are gene expression measurements of patients with five distinct types of tumors, BRCA, KIRC, COAD, LUAD and PRAD, as labels.
- The dataset contains 20531 covariates and 801 subjects.
- The number of covariates is much larger than the number of subjects, so we pre-select part of the genes before fitting the DPMM.
- Considering the relatively small number of subjects, we chose $p = 20$ using the ordered p-values after performing ANOVA testing.

Clustering the subjects

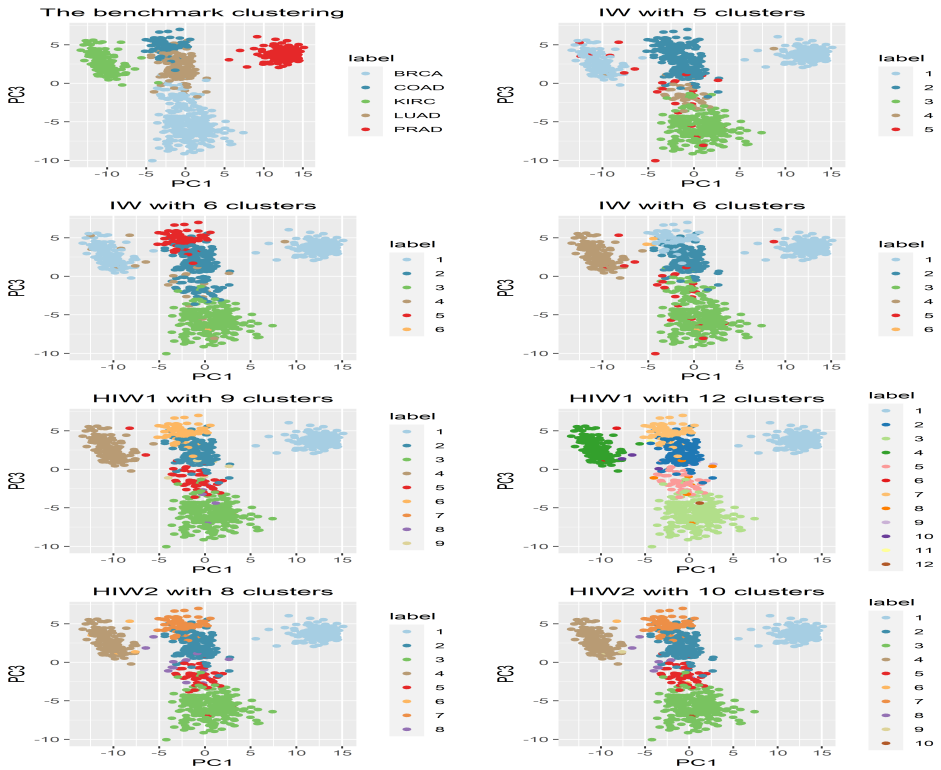

Figure: One representative clustering for a certain number of clusters for priors IW, HIW1 and HIW2.

Clustering the subjects

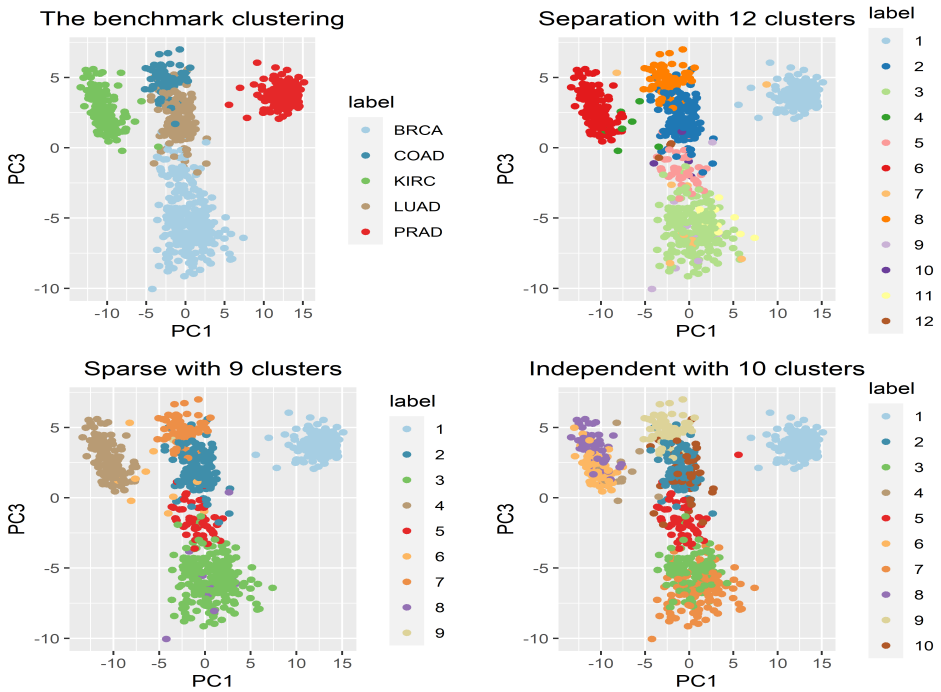

Figure: One representative clustering for a certain number of clusters for the separation, sparse and independent priors.

The Rand index

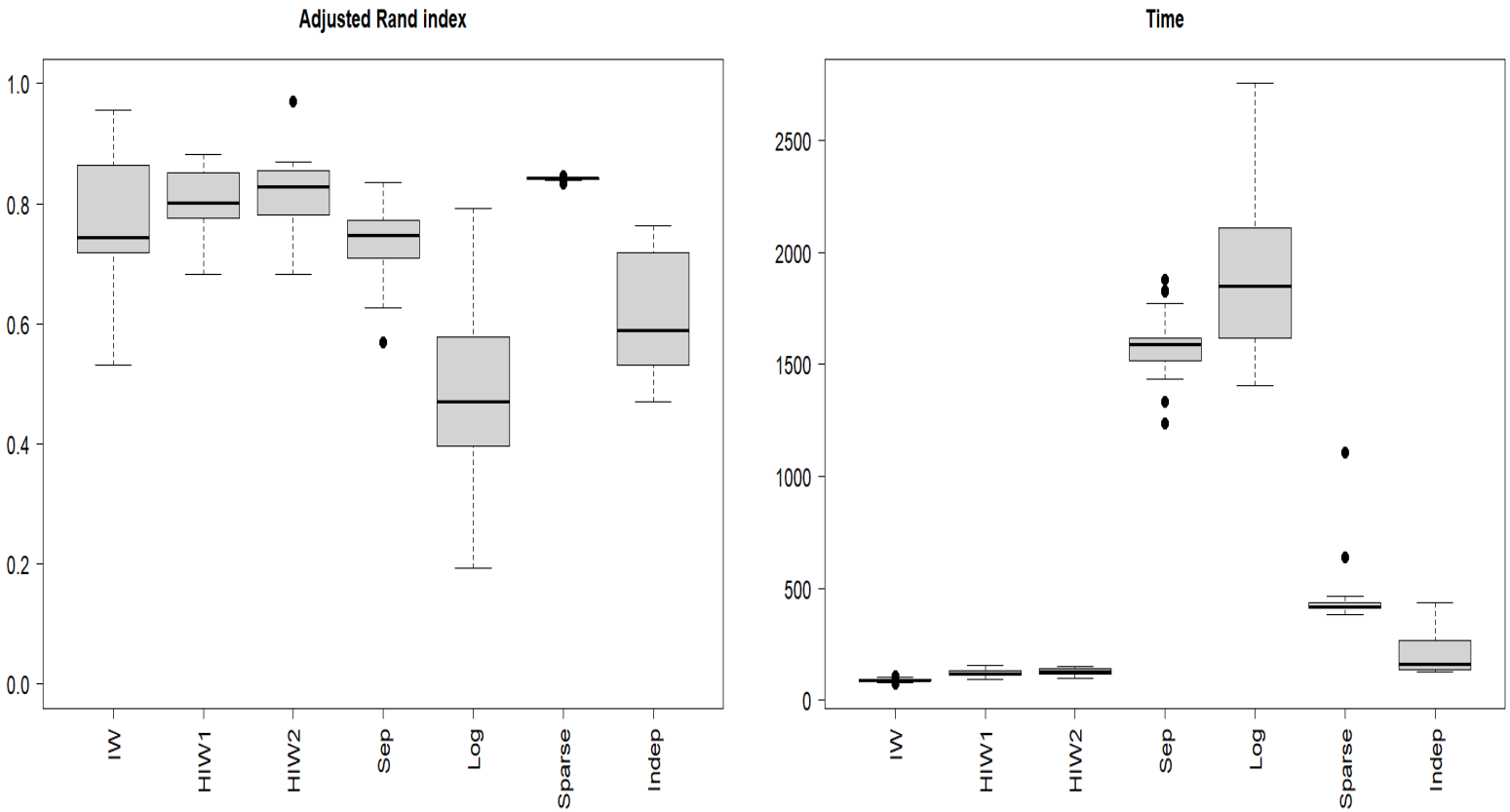

Figure: Boxplots of adjusted Rand indices for different priors (left) and computational times after running 30 seeds (right) in seconds.

Conclusions

- The sparse prior is one that performs best when the dimensionality of the problem increases beyond a handful of variables, among the priors we examined.
- Importantly the sparse prior leads to faster convergence and more reliable MCMC output.
- Not straightforward to fully disentangle the effect of the prior on improving the model (e.g. flexibility) from its effect on the sampler and convergence.
- We are currently working on examining the effect of prior specification on Mixtures of Finite Mixtures and the Telescopic sampler.
- arXiv manuscript discusses the importance of the prior on the Normal mean, as well as the initialisation of the Markov chain.

Some References

- Liverani, S., Hastie, D. I., Azizi, L., Papathomas, M. and Richardson, S. (2015) PReMiuM: An R package for Profile Regression Mixture Models using Dirichlet Processes. Journal of Statistical Software. , 64, Issue 7, pp 1-30.
- J. T. Molitor, M. Papathomas, M. Jerrett and S. Richardson (2010) Bayesian Profile Regression with an Application to the National Survey of Children's Health, Biostatistics, 11, 484-498.
- M. Papathomas, J. Molitor, S. Richardson, E. Riboli and P. Vineis (2011) Examining the joint effect of multiple risk factors using exposure risk profiles: lung cancer in non smokers. Environmental Health Perspectives, 119, 84-91.
- Papathomas, M , Molitor, J, Hoggart, C, Hastie, D and Richardson, S (2012) Exploring data from genetic association studies using Bayesian variable selection and the Dirichlet process: application to searching for gene-gene patterns. Genetic Epidemiology. 36:663-674
- Jing, W., Papathomas, M. and Liverani, S. (2022). Prior specifications for Dirichlet Process Mixture Models with Gaussian kernels. Submitted.
- Bigelow and Dunson (2009). Bayesian Semiparametric Joint Models for Functional Predictors. JASA, 104, 26-36.
- Chung Y, Dunson DB. 2009. Nonparametric Bayes conditional distribution modelling with variable selection. J Am Statist Assoc 104:1646-60
- Yang Y, Dunson DB. 2013. Bayesian conditional tensor factorizations for high-dimensional classification. Preprint.