Variable Selection in Covariate Dependent Random Partition Models: an Application to Urinary Tract Infection

William Barcella

joint with Maria De Iorio Gianluca Baio James Malone-Lee

Dept. of Statistical Science, University College London

email: *william.barcella.13@ucl.ac.uk*

BAYES2015
May 19-22, 2015 (Basel)

Motivation: Lower Urinary Tract Symptoms (LUTS)

Study:

Relation between Urinary Tract Infection (UTI) and LUTS

Dataset:

- patients diagnosed with UTI (*i.e.* WBC≥ 1, white blood cells count)
- all women over 18 y.o. at first attendance visit
- $y_i = \log(\text{WBC}_i)$ for $i = 1, \ldots, 1341$
- $\boldsymbol{x}_i$ contains 34 binary indicators (LUTS profile)

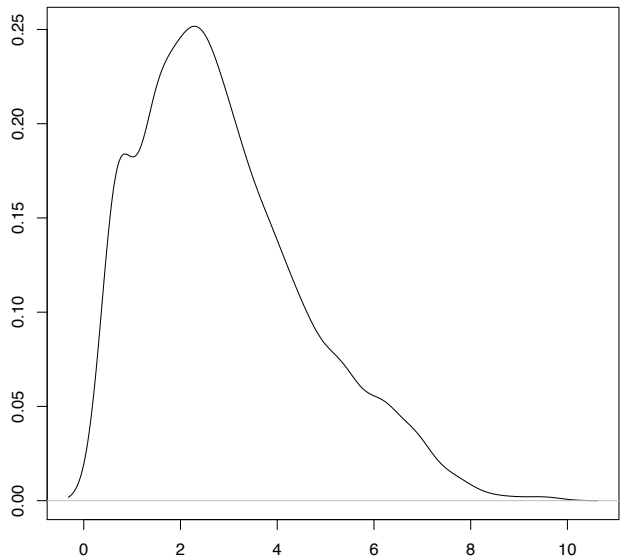


Figure: Kernel density estimation of log(WBC)

Clustering and Variable (Model) Selection

Typical Medical Problem

- $\boldsymbol{y} = (y_1, \ldots, y_n)$: individual level outcomes (response)
- $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ individual level profiles (covariates)

$\Rightarrow$ **investigate the relation** between $\boldsymbol{y}$ and $\boldsymbol{X}$ ($e.g.\ y_i \sim N(\boldsymbol{x}_i\boldsymbol{\beta}, \lambda)$)

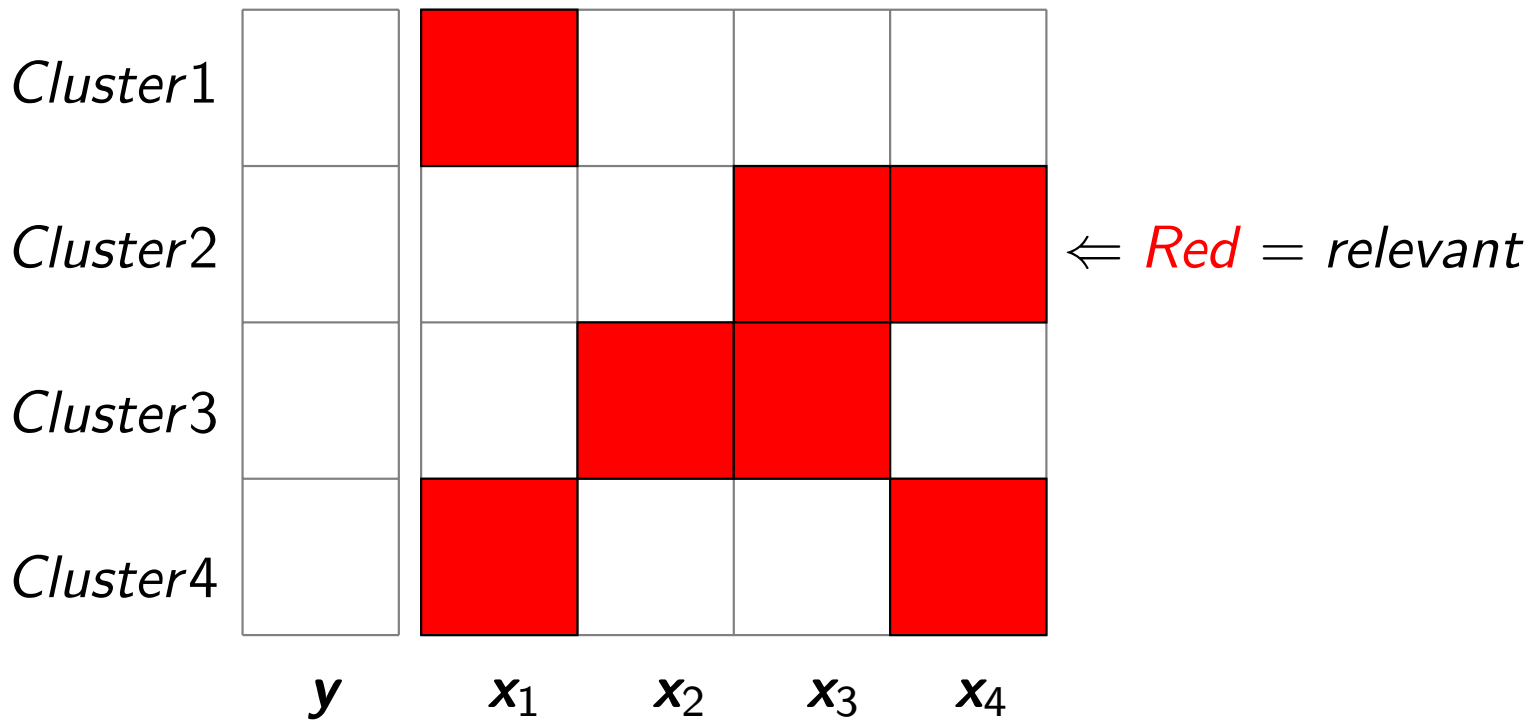
Clustering: do we expect the same relation for all $i = 1, \ldots, n$?

Variable Selection: if $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{iD})$, do we expect all D covariates to affect/explain y_i?

Clustering and Variable Selection

Further Complication

One (set of) covariate(s) may be relevant in explaining the outcome variable for a subset of patients



Objectives

Model that performs **regression analysis** of $\boldsymbol{y}$ on $\boldsymbol{X}$ within

1. **clusters** of individuals based on both **response** and **profiles**
   - patients with similar profiles should be a priori more likely to co-cluster
   - predict similar responses for similar profiles
2. **selecting important covariates** explaining $\boldsymbol{y}$ (within each cluster)

Model-Based Clustering

Regression model

$$y_i \mid \boldsymbol{\beta}_i, \lambda \sim \text{Normal}(y_i \mid \boldsymbol{x}_i \boldsymbol{\beta}_i, \lambda)$$

Infinite Mixture Model

$s_i \in \{1, 2, \ldots\}$: cluster assignment

$$y_1, \ldots, y_n \mid \{\psi_k, \boldsymbol{\beta}_k\}_{k=1}^{\infty}, \lambda \sim \sum_{k=1}^{\infty} \psi_k N(y_i \mid \boldsymbol{x}_i \boldsymbol{\beta}_{s_i=k}, \lambda)$$

- **nonparametric model:** unbounded number of degrees of freedom
- how to achieve this: **Dirichlet Process**

Dirichlet Process

Dirichlet Process (DP): distribution over distributions (Ferguson [1973])

Stick-breaking construction (Sethuraman [1991])

If $G \sim \mathrm{DP}(\alpha, G_0)$ then

$$
\begin{aligned}
G &= \sum_{k=1}^{\infty} \psi_k \delta_{\boldsymbol{\beta}_k} \\
\psi_k &= \xi_k \prod_{h=1}^{k-1} (1 - \psi_h) \\
\xi_1, \xi_2, \ldots &\overset{iid}{\sim} \mathrm{Beta}(1, \alpha) \\
\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \ldots &\overset{iid}{\sim} G_0
\end{aligned}
$$

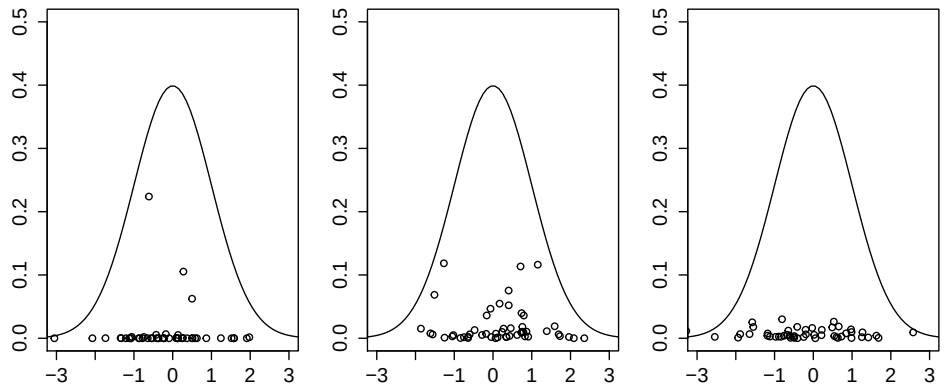


Figure: Three samples from $\mathrm{DP}(\alpha, G_0)$. The continuous line is G_0 (standard Gaussian). α is 1, 10 and 100 from the left panel respectively. G is represented by the points.

Dirichlet Process Mixture Models (DPMM)

Infinite mixture of regressions via **DPMM** (Antoniak [1974]):

$$
\begin{aligned}
y_i \mid \boldsymbol{\beta}_i &\sim \text{Normal}(y_i \mid \boldsymbol{x}_i \boldsymbol{\beta}_i, \lambda) \\
\boldsymbol{\beta}_i \mid G &\sim G \qquad \left(= \sum_{k=1}^{\infty} \psi_k \delta_{\boldsymbol{\beta}_k} \Rightarrow \text{Discrete distribution} \right) \\
G &\sim \text{DP}(\alpha, G_0)
\end{aligned}
$$

Under G : $p(\boldsymbol{\beta}_i = \boldsymbol{\beta}_{i'}) > 0$

- Observations share the **same** $\boldsymbol{\beta} \Rightarrow$ belong to the **same cluster**

$\Rightarrow$ DPMM induces a **random partition** of $\{1, \ldots, n\}$

Task 1: clustering with covariates information

Express the sequence $\psi_1, \psi_2, \ldots$ as function of $\boldsymbol{x}$ using an **auxiliary model** for the covariates (*e.g.* Müller et al. [1996], Müller et al. [2011], etc)

- $\boldsymbol{x}_i$'s become random with Bernoulli model

$$y_i, \boldsymbol{x_i} \mid \{\boldsymbol{\beta}_k, \boldsymbol{\zeta}_k, \psi_k\}_{k=1}^{\infty} \sim \sum_{k=1}^{\infty} \psi_k p(\boldsymbol{x}_i \mid \boldsymbol{\zeta}_k) N(y_i \mid \boldsymbol{x}_i \boldsymbol{\beta}_k, \lambda)$$

Difference:

$$\begin{aligned} (\boldsymbol{\beta}_i, \boldsymbol{\zeta}_i) \mid G &\sim G = \sum_{k=1}^{\infty} \psi_k (\delta_{\boldsymbol{\beta}_k} \times \delta_{\boldsymbol{\zeta}_k}) \\ G &\sim \mathrm{DP}(\alpha, G_0) \end{aligned}$$

with $G_0 = G_{0\beta} \times G_{0\zeta}$ (convenient choices are Normal and Beta)

Task 2: variable selection

BNP models are ideal for simultaneous clustering and variable selection

Objective: select subsets of covariates most associated with the response, allowing different subsets of covariates for different clusters

Modify $G_0 = G_{0\beta} \times G_{0\zeta}$:

Spike and Slab distribution

$$G_{0\beta}(\boldsymbol{\beta}) = \prod_{d=1}^{D} [\omega_d \delta_0(\beta_{\cdot d}) + (1 - \omega_d) \text{Normal}(\beta_{\cdot d} \mid m_d, \tau_d)]$$

$\Rightarrow$ allow β to be **exactly zero in some cluster**

NOTE Spike and Slab priors already used for variable selection (see Kim et al. [2009]), but not in clustering with covariates information

Random Partition Model with Covariate Selection (RPMS)

Summary of the resulting RPMS, constructed via DPMM:

- Sampling (and auxiliary) model

$$y_i, \boldsymbol{x}_i \mid \boldsymbol{\beta}_i, \boldsymbol{\zeta}_i \sim \text{Normal}(y_i \mid \boldsymbol{x}_i \boldsymbol{\beta}_i, \lambda) \prod_{d=1}^{D} \text{Bernoulli}(x_{id} \mid \zeta_{id})$$

- Prior distribution for $\boldsymbol{\beta}_i, \boldsymbol{\zeta}_i$

$$\begin{aligned} (\boldsymbol{\beta}_i, \boldsymbol{\zeta}_i) \mid G &\sim G \\ G &\sim \text{DP}(\alpha, G_0) \end{aligned}$$

- Base measure for G_0 (*within-cluster-prior* for $\boldsymbol{\beta}$ and $\boldsymbol{\zeta}$)

$$G_0 = \prod_{d=1}^{D} \{[\omega_d \delta_0(\beta_{\cdot d}) + (1 - \omega_d) \text{ Normal}(\beta_{\cdot d} \mid m_d, \tau_d)] \text{Beta}(\zeta_{\cdot d} \mid a_\zeta, b_\zeta)\}$$

- Hyperprior distributions chosen for **conjugacy** or **computational advantages**.

Bayesian Inference

Posterior Inference

$$p(\theta \mid \text{Data}) \propto p(\text{Data} \mid \theta)p(\theta)$$

Efficient **MCMC** algorithm based on **Gibbs samplers** for sampling from posteriors (auxiliary variable algorithm by Neal [2000])

Predictive Inference

$$p(\tilde{y} \mid \boldsymbol{y}, \boldsymbol{X}, \tilde{\boldsymbol{x}}) = \int p(\tilde{y} \mid \tilde{\boldsymbol{x}}, \boldsymbol{\beta}, \boldsymbol{\zeta}) dp(\boldsymbol{\beta}, \boldsymbol{\zeta} \mid \boldsymbol{y}, \boldsymbol{X}, \tilde{\boldsymbol{x}})$$

Obtainable in **Gibbs fashion** sampling from the posterior of $\boldsymbol{\beta}$ and $\boldsymbol{\zeta}$ integrating over the cluster allocation given the new profile $\tilde{\boldsymbol{x}}$

Application on WBC and LUTS

We apply RPMS to evaluate the relation between LUTS profile and level of UTI (in terms of log(WBC):

- clustering output
- variable selection

Competitor: a model with spike and slab distribution but without a model on the covariates (we call it SSP)

Clustering Output – Posterior distribution of k

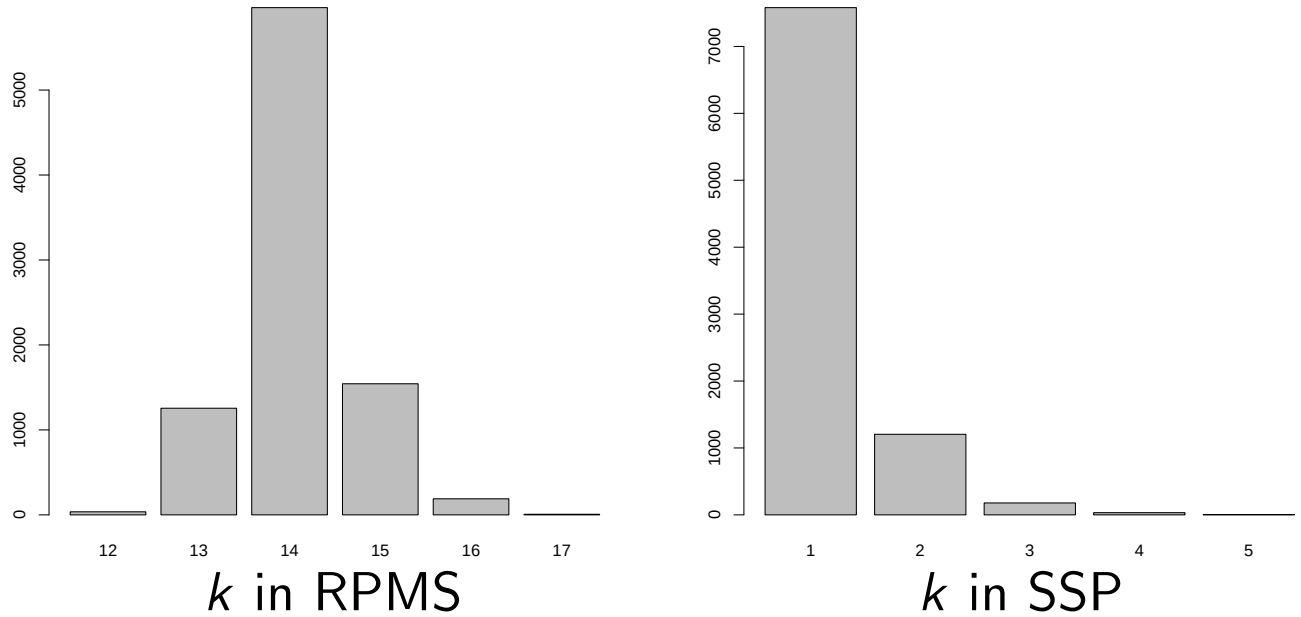
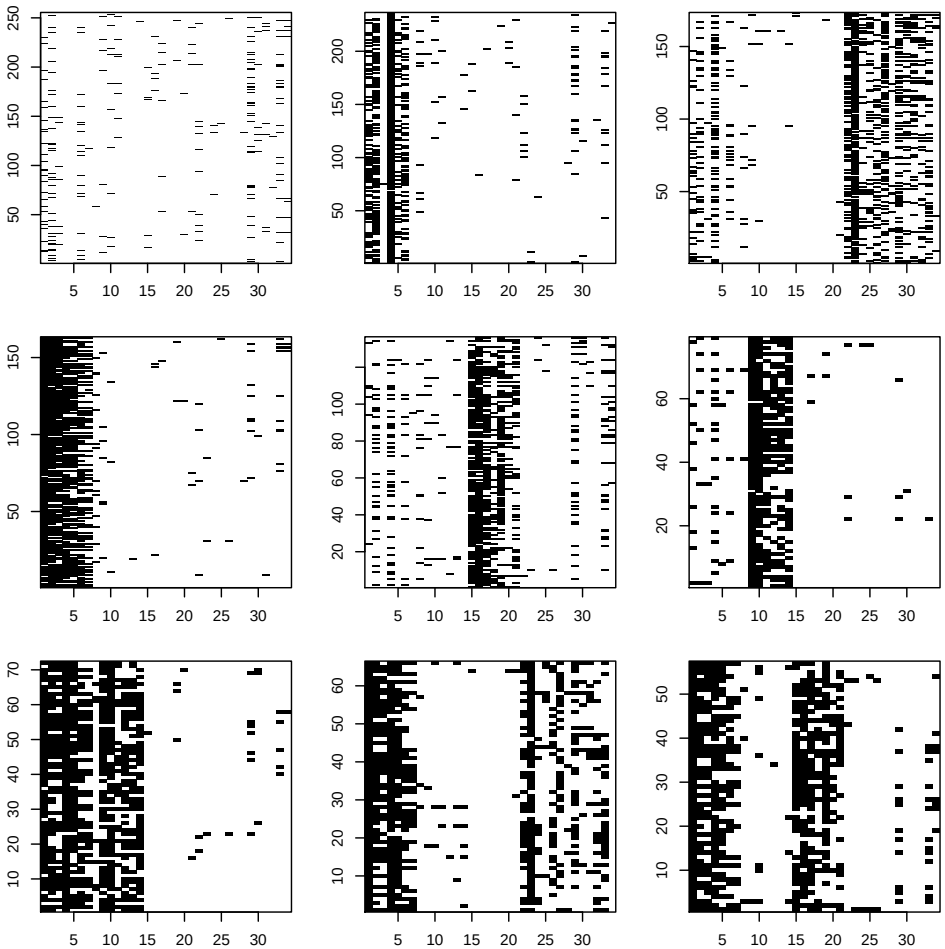


Figure: Posterior distribution of the number of clusters k for RPMS and for SSP models.

NOTE: RPMS accounts also for the variability within the covariates

Clustering Output – Binder estimate

Minimise: $L(\hat{\boldsymbol{s}}, \boldsymbol{s}) = \sum_{i<i'} \left(l_1 \cdot I_{\{\hat{s}_i \neq \hat{s}_{i'}\}} I_{\{s_i = s_{i'}\}} + l_2 \cdot I_{\{\hat{s}_i = \hat{s}_{i'}\}} I_{\{s_i \neq s_{i'}\}} \right)$



- urgency symptoms $\rightarrow$ from 1 to 8
- stress incontinence symptoms $\rightarrow$ from 9 to 14
- voiding symptoms $\rightarrow$ from 15 to 21
- pain symptoms $\rightarrow$ from 22 to 34

Categories of symptoms (or combinations of categories) are associated to different values β

Variable Selection – Fixing the partition

Alternative 1 $\Rightarrow 1 - p(\beta^*_{jd} = 0 \mid \hat{\boldsymbol{s}}, \ldots)$ where $\hat{\boldsymbol{s}}$ is the Binder estimate

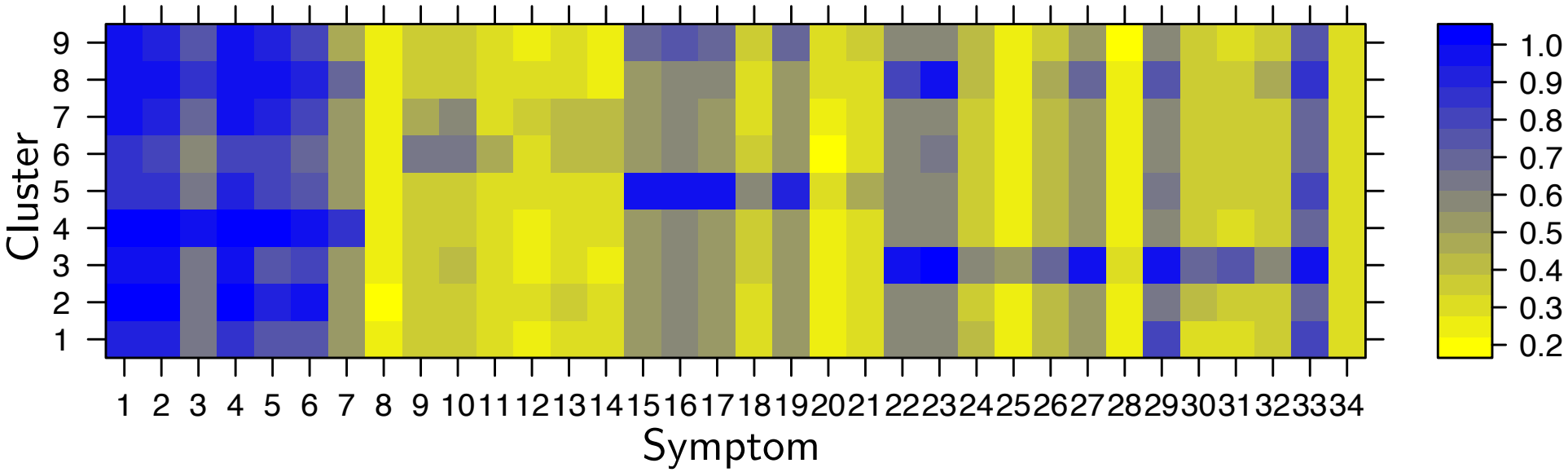
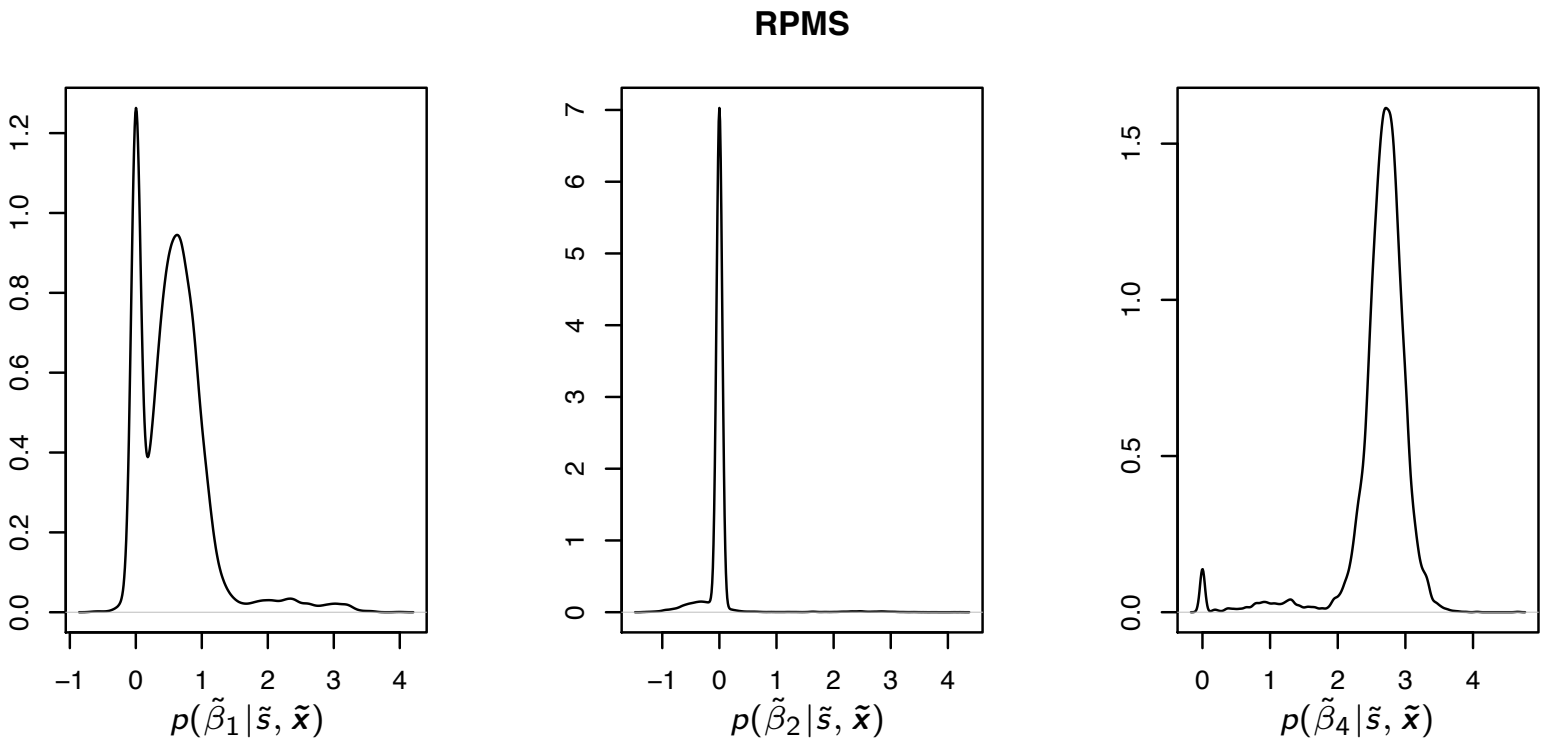


Figure: Probability of inclusion, *i.e.* $\beta \neq 0$, for each symptoms in the 9 biggest clusters of the partition estimated by minimizing the Binder loss function.

Variable Selection – Fixing the profile

Alternative 2 $\Rightarrow$ fixing $\tilde{\boldsymbol{x}}$ and check the posterior of β
e.g. $\tilde{\boldsymbol{x}}$: x_1, x_2 and x_4 activated



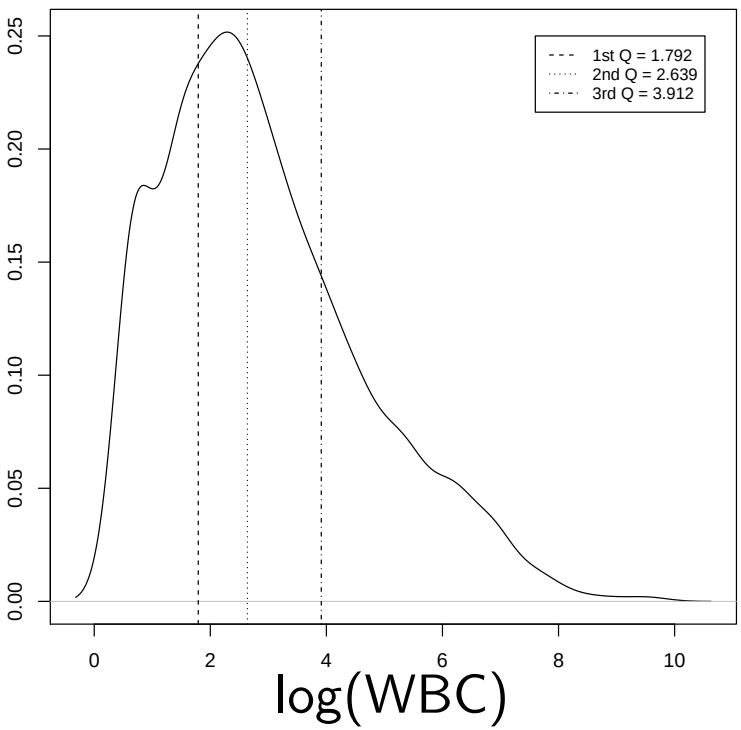
NOTE SSP has posterior distributions independent on $\tilde{\boldsymbol{x}}$

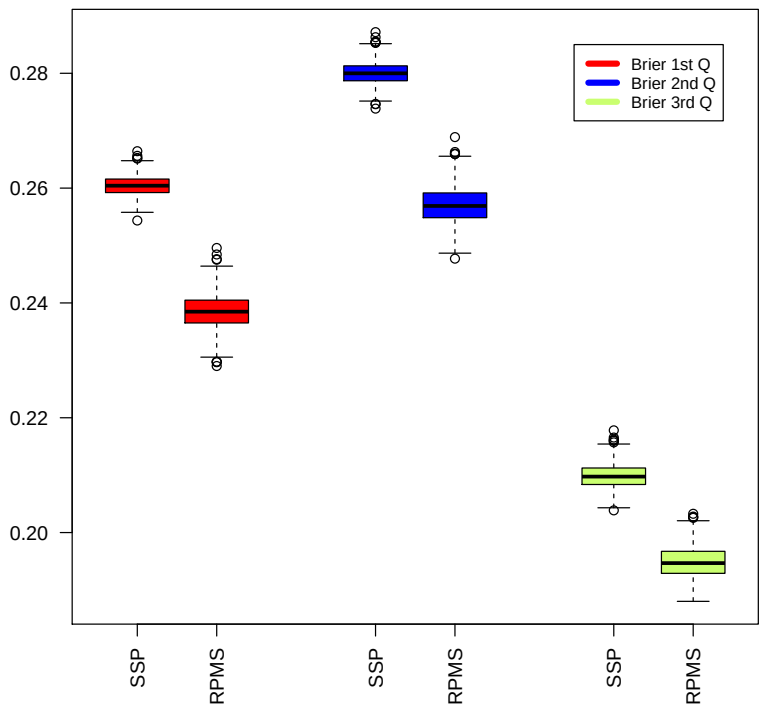
Variable Selection – Effects on predictive inference

Compare posterior distributions of **Brier score**

$\text{Brier}_{(q)} = \frac{1}{n}\sum_{i=1}^{n}(f_i^{(q)} - y_i^{(q)})^2$

- $y_i^{(q)}$: 1 if $y_i > q$-th quartile, 0 otherwise
- $f_i^{(q)}$: $p(\tilde{y} > q\text{-th quartile} \mid \boldsymbol{y})$





Conclusions

RPMS (Barcella et al. [2015]) is a model based on a DPM of regressions that performs simultaneously

- clustering with covariates (by modelling the covariates)
- within clusters variable selection in terms of explanatory power on y (by spike and slab prior distribution)

The resulting method has been applied to investigate the relation between LUTS and WBC finding:

1. LUTS categories significant in explaining levels of WBC
2. importance of the urgency symptoms
3. improved predictions of WBC levels using LUTS profile as predictor

References

Antoniak, C. E. (1974). Mixtures of dirichlet processes with applications to bayesian nonparametric problems. *The annals of statistics*, pages 1152–1174.

Barcella, W., De Iorio, M., Baio, G., and Malone-Lee, J. (2015). Variable selection in covariate dependent random partition models: an application to urinary tract infection. *arXiv preprint arXiv:1501.03537*.

Blackwell, D. and MacQueen, J. B. (1973). Ferguson distributions via pólya urn schemes. *The annals of statistics*, pages 353–355.

Ferguson, T. S. (1973). A bayesian analysis of some nonparametric problems. *The annals of statistics*, pages 209–230.

Kim, S., Dahl, D. B., and Vannucci, M. (2009). Spiked dirichlet process prior for bayesian multiple hypothesis testing in random effects models. *Bayesian analysis (Online)*, 4(4):707.

Müller, P., Erkanli, A., and West, M. (1996). Bayesian curve fitting using multivariate normal mixtures. *Biometrika*, 83(1):67–79.

Müller, P., Quintana, F., and Rosner, G. L. (2011). A product partition model with regression on covariates. *Journal of Computational and Graphical Statistics*, 20(1).

Neal, R. M. (2000). Markov chain sampling methods for dirichlet process mixture models. *Journal of computational and graphical statistics*, 9(2):249–265.

Sethuraman, J. (1991). A constructive definition of dirichlet priors. Technical report, DTIC Document.

The End