Modelling highly-collinear data: Bayesian profile regression in epidemiology

Dr Silvia Liverani

Outline

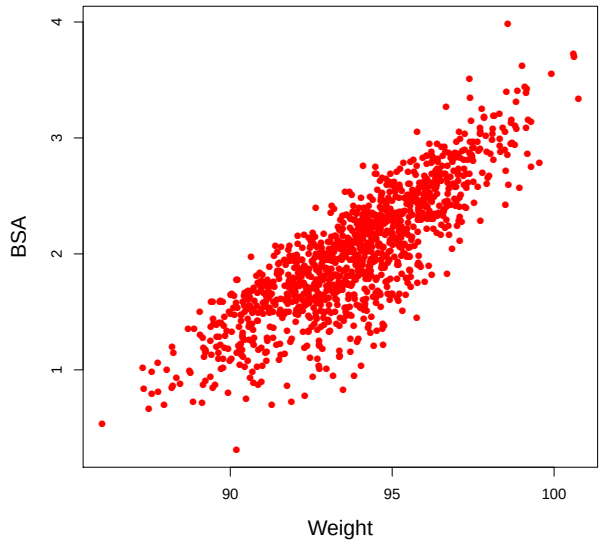
- Motivation
- Profile regression
- Convergence, mixing issues and label switching
- Applications to epidemiology

Model instability due to highly correlated predictors

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_p x_{pi} + \varepsilon_i$$

Example Researchers are interested in determining if a relationship exists between **blood pressure** (y = BP, in mm Hg) and

- **weight** (x_1 = Weight, in kg)
- **body surface area** (x_2 = BSA, in sq m)
- duration of hypertension, basal pulse, stress index, ...

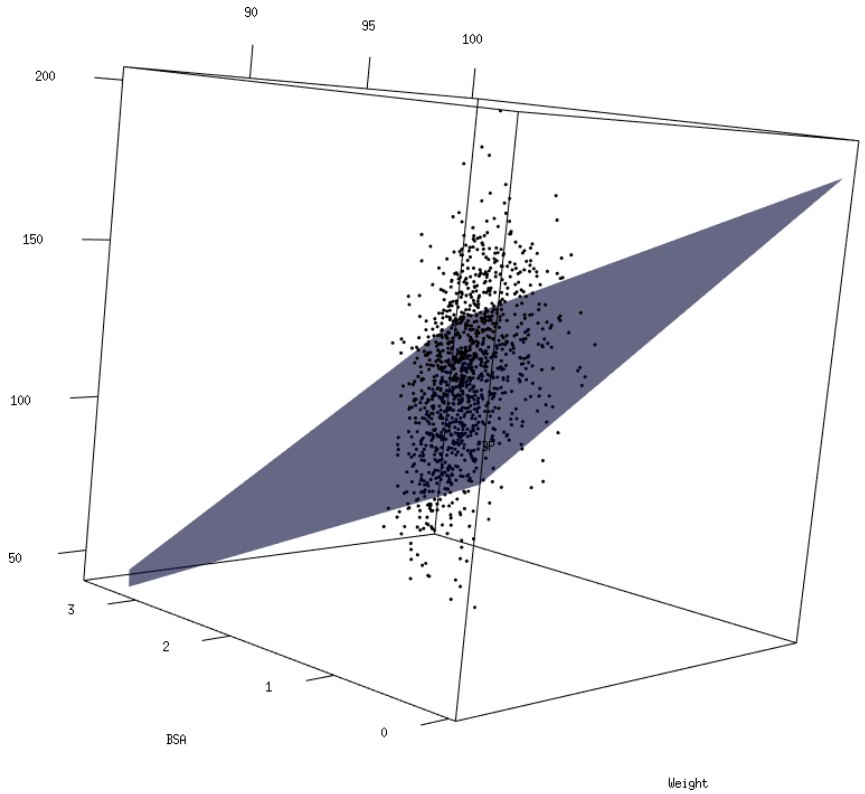


Model instability due to highly correlated predictors

Model	$\hat{\beta}_1$	SE $\hat{\beta}_1$	$\hat{\beta}_2$	SE $\hat{\beta}_2$
$y_i = \beta_1 x_{1i} + \ldots + \varepsilon_i$	2.64	0.30	–	–
$y_i = \beta_2 x_{2i} + \ldots + \varepsilon_i$	–	–	3.34	1.33
$y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \varepsilon_i$	6.58	0.53	-20.44	2.28

Model instability due to highly correlated predictors

Model	$\hat{\beta}_1$	SE $\hat{\beta}_1$	$\hat{\beta}_2$	SE $\hat{\beta}_2$
$y_i = \beta_1 x_{1i} + \ldots + \varepsilon_i$	2.64	0.30	–	–
$y_i = \beta_2 x_{2i} + \ldots + \varepsilon_i$	–	–	3.34	1.33
$y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \varepsilon_i$	6.58	0.53	-20.44	2.28

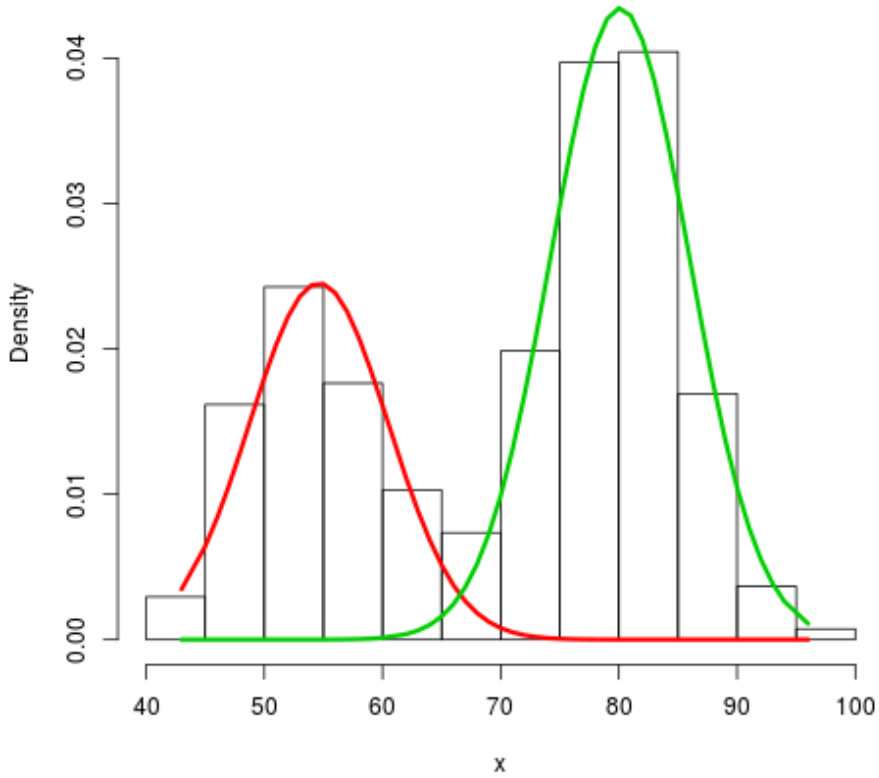


How to deal with collinearity

- Aggregate all the highly correlated variables of interest
- Look at each variable individually
- Clustering
  - Models that include main effects and interactions of increasing order become quickly unwieldy and lose power
  - By studying how the subjects typically cluster into groups and the profiles of the subjects within each group, for example, we characterise how combinations of risk factors may affect the risk of disease

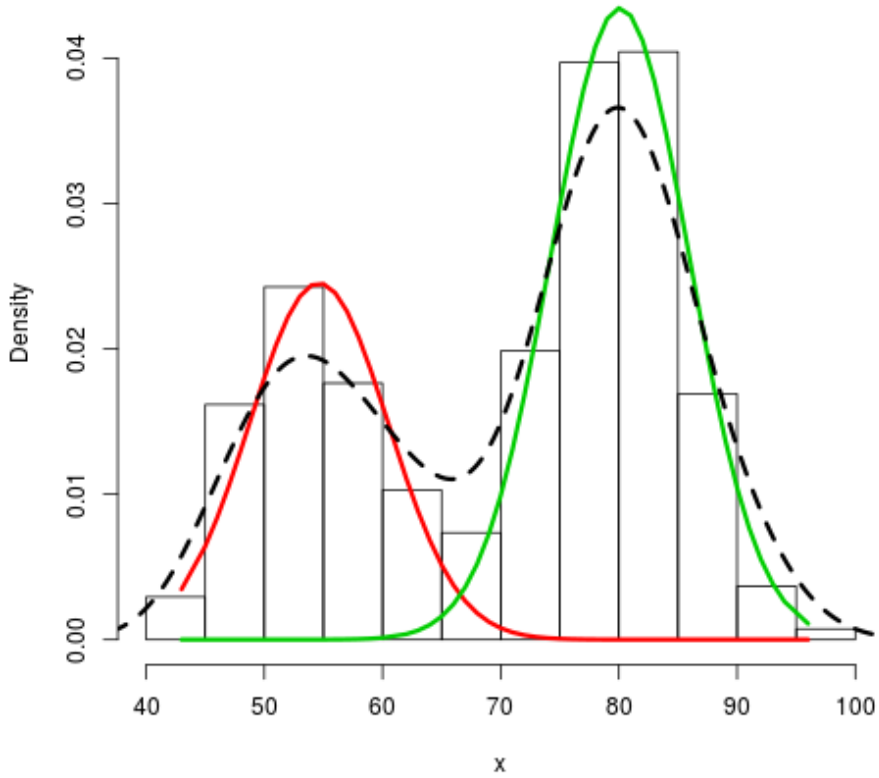
Mixture models for clustering

$$\sum_c \psi_c \text{Normal}(x|\mu_c, \sigma_c)$$



Mixture models for clustering

$$f(x) = \sum_c \psi_c \text{Normal}(x|\mu_c, \sigma_c)$$



Profile regression

- Flexible but tractable non-parametric Bayesian mixture model
- For application in the health sciences, it is important to perform joint modelling of covariate pattern and response

$$f(\mathbf{x}_i, y_i|\phi, \boldsymbol{\theta}, \psi, \boldsymbol{\beta}, \boldsymbol{z}, \boldsymbol{w}_i) = \sum_{c=1}^{\infty} \psi_c f(\mathbf{x}_i|z_i = c, \phi_c) f(y_i|z_i = c, \theta_c, \boldsymbol{\beta}, \mathbf{w}_i)$$

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

Profile regression

$$f(\mathbf{x}_i, y_i|\boldsymbol{\phi}, \boldsymbol{\theta}, \boldsymbol{\psi}, \boldsymbol{\beta}, \boldsymbol{z}, \boldsymbol{w}_i) = \sum_{c=1}^{\infty} \psi_c f(\mathbf{x}_i|z_i = c, \phi_c) f(y_i|z_i = c, \theta_c, \boldsymbol{\beta}, \mathbf{w}_i)$$

Profile regression

$$f(\mathbf{x}_i, y_i|\boldsymbol{\phi}, \boldsymbol{\theta}, \boldsymbol{\psi}, \boldsymbol{\beta}, \boldsymbol{z}, \boldsymbol{w}_i) = \sum_{c=1}^{\infty} \psi_c f(\mathbf{x}_i|z_i = c, \phi_c) f(y_i|z_i = c, \theta_c, \boldsymbol{\beta}, \mathbf{w}_i)$$

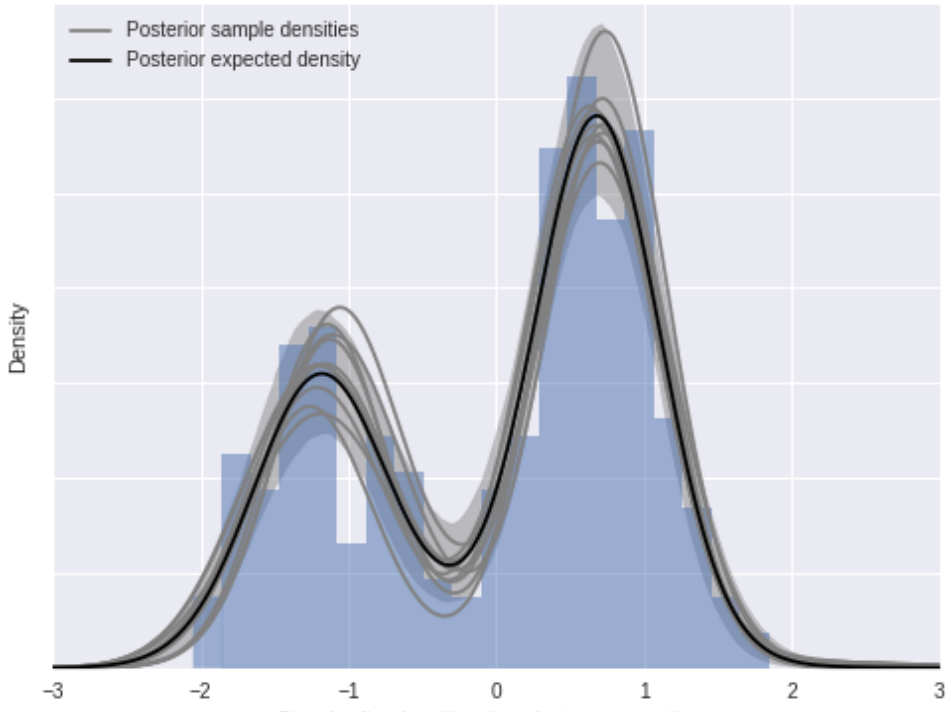
- Joint response and covariate models. For example,

$$f(\mathbf{x}_i|z_i = c, \phi_c) = \prod_{j=1}^{J} \phi_{z_i,j,x_{i,j}}$$

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

Dirichlet process mixture model

$$
\begin{aligned}
P &\sim \mathrm{DP}(\alpha, P_0) \\
\Theta_i &\sim P \\
x_i|\Theta_i &\sim F_{\Theta_i}
\end{aligned}
$$



Profile regression

Stick-breaking priors (constructive definition of the Dirichlet Process)

$$\mathbb{P}(z_i = c|\boldsymbol{\psi}) = \psi_c \qquad \psi_1 = V_1 \qquad V_c \sim \text{Beta}(1, \alpha)$$

$$\psi_c = V_c \prod_{l<c} (1 - V_l)$$

Profile regression

Stick-breaking priors (constructive definition of the Dirichlet Process)

$$\mathbb{P}(z_i = c|\boldsymbol{\psi}) = \psi_c \qquad \psi_1 = V_1 \qquad V_c \sim \text{Beta}(1, \alpha)$$

$$\psi_c = V_c \prod_{l<c} (1 - V_l)$$

Profile regression

Stick-breaking priors (constructive definition of the Dirichlet Process)

$$\mathbb{P}(z_i = c|\boldsymbol{\psi}) = \psi_c \qquad \psi_1 = V_1 \qquad V_c \sim \text{Beta}(1, \alpha)$$

$$\psi_c = V_c \prod_{l<c} (1 - V_l)$$

Profile regression

Stick-breaking priors (constructive definition of the Dirichlet Process)

$$\mathbb{P}(z_i = c|\boldsymbol{\psi}) = \psi_c \qquad \psi_1 = V_1 \qquad V_c \sim \text{Beta}(1, \alpha)$$

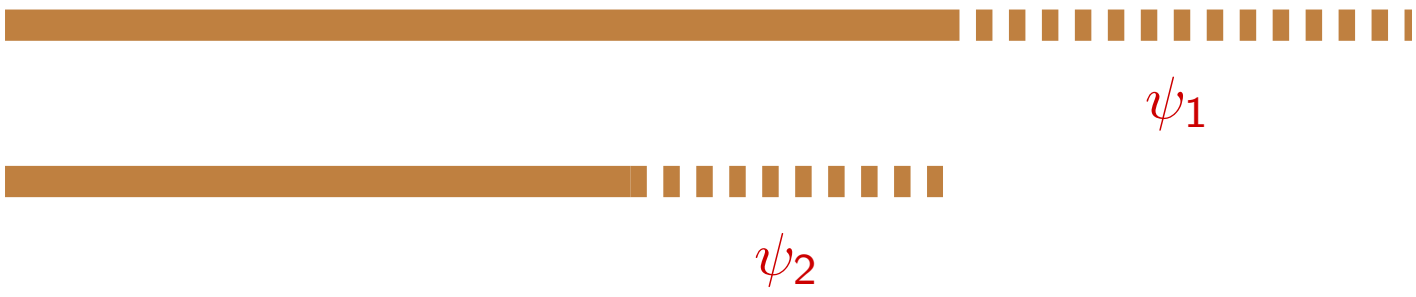
$$\psi_c = V_c \prod_{l<c} (1 - V_l)$$

Profile regression

Stick-breaking priors (constructive definition of the Dirichlet Process)

$$\mathbb{P}(z_i = c|\boldsymbol{\psi}) = \psi_c \qquad \psi_1 = V_1 \qquad V_c \sim \text{Beta}(1, \alpha)$$

$$\psi_c = V_c \prod_{l<c} (1 - V_l)$$

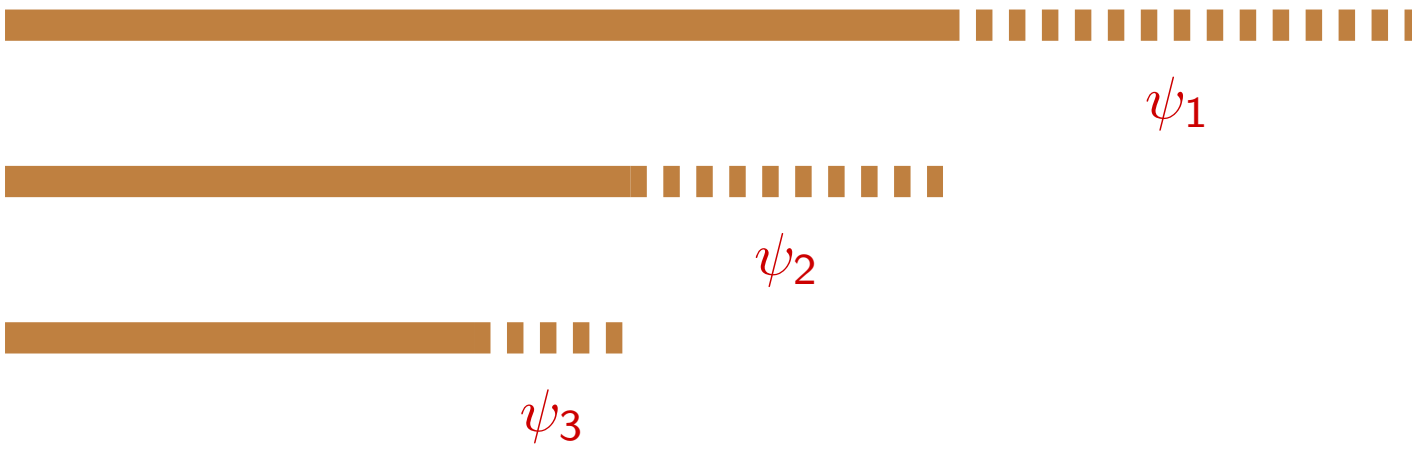


Profile regression

Stick-breaking priors (constructive definition of the Dirichlet Process)

$$\mathbb{P}(z_i = c|\boldsymbol{\psi}) = \psi_c \qquad \psi_1 = V_1 \qquad V_c \sim \text{Beta}(1, \alpha)$$

$$\psi_c = V_c \prod_{l<c} (1 - V_l)$$

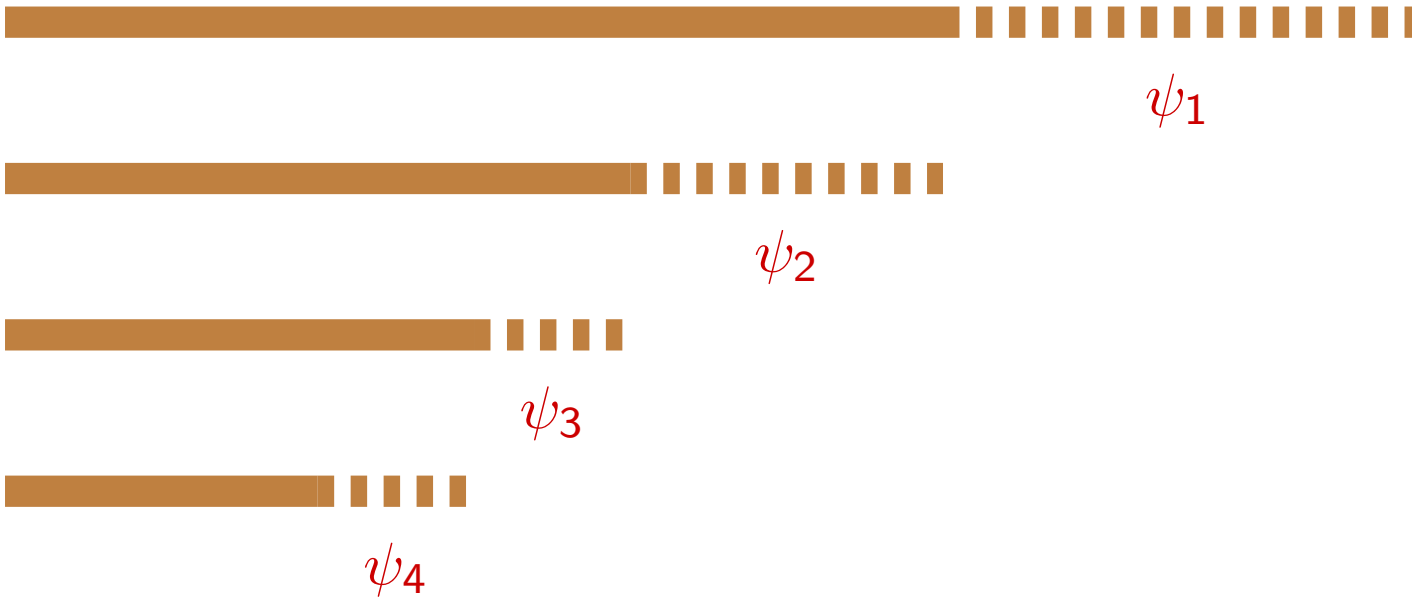


Profile regression

Stick-breaking priors (constructive definition of the Dirichlet Process)

$$\mathbb{P}(z_i = c|\boldsymbol{\psi}) = \psi_c \qquad \psi_1 = V_1 \qquad V_c \sim \text{Beta}(1, \alpha)$$

$$\psi_c = V_c \prod_{l<c} (1 - V_l)$$



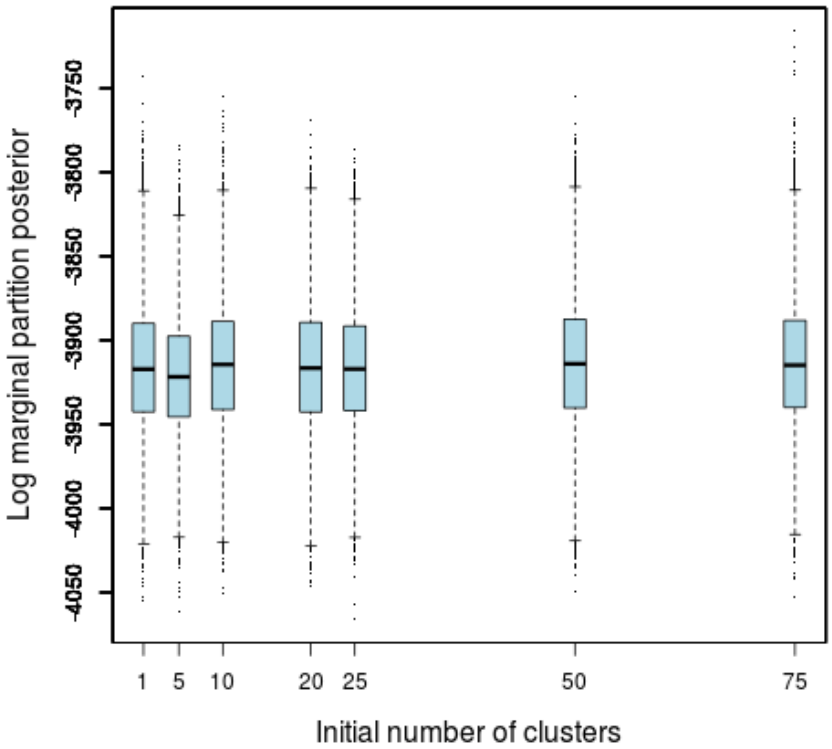
Methods and implementation

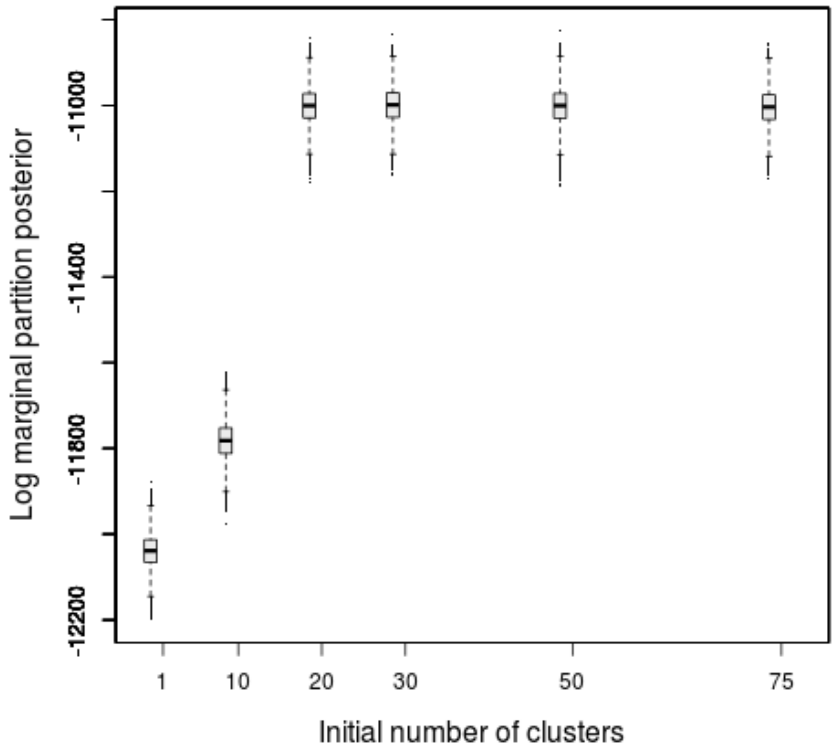
- Implementation of the methods in the R package PReMiuM (Liverani et al, Journal of Statistical Software, 2015)
  - Includes binary, binomial, categorical, Normal, Poisson and survival outcome, spatial CAR term, Normal and/or discrete covariates
  - Dependent or independent slice sampling or truncated Dirichlet process model, fixed or random α, or use the Pitman-Yor process prior
  - Can handles missing data and runs predictive scenarios
- Modelling collinear and spatially correlated data (Liverani et al, 2016)
  - Spatial conditional autoregressive component to allow for spatial correlation
- Quantile Profile Regression (submitted)
  - Asymmetric Laplace distribution to model the tails more accurately
- Mixing issues (Hastie, Liverani et al, Statistics and Computing, 2015)

Marginal partition posterior

We can use the marginal partition posterior to assess convergence.

$$p(z|y,x,w) \propto p(x|z)p(y|z,w)p(z)$$





Spatial Modelling

For example, the likelihood component for Gaussian response becomes

$$f(y_i|z_i = c, \theta_c, \boldsymbol{\beta}, \sigma^2_Y, u_i, W_i) = \frac{1}{\sqrt{2\pi\sigma^2_Y}} \exp(-\frac{1}{2\sigma^2_Y}(Y_i - \lambda_i)^2)$$

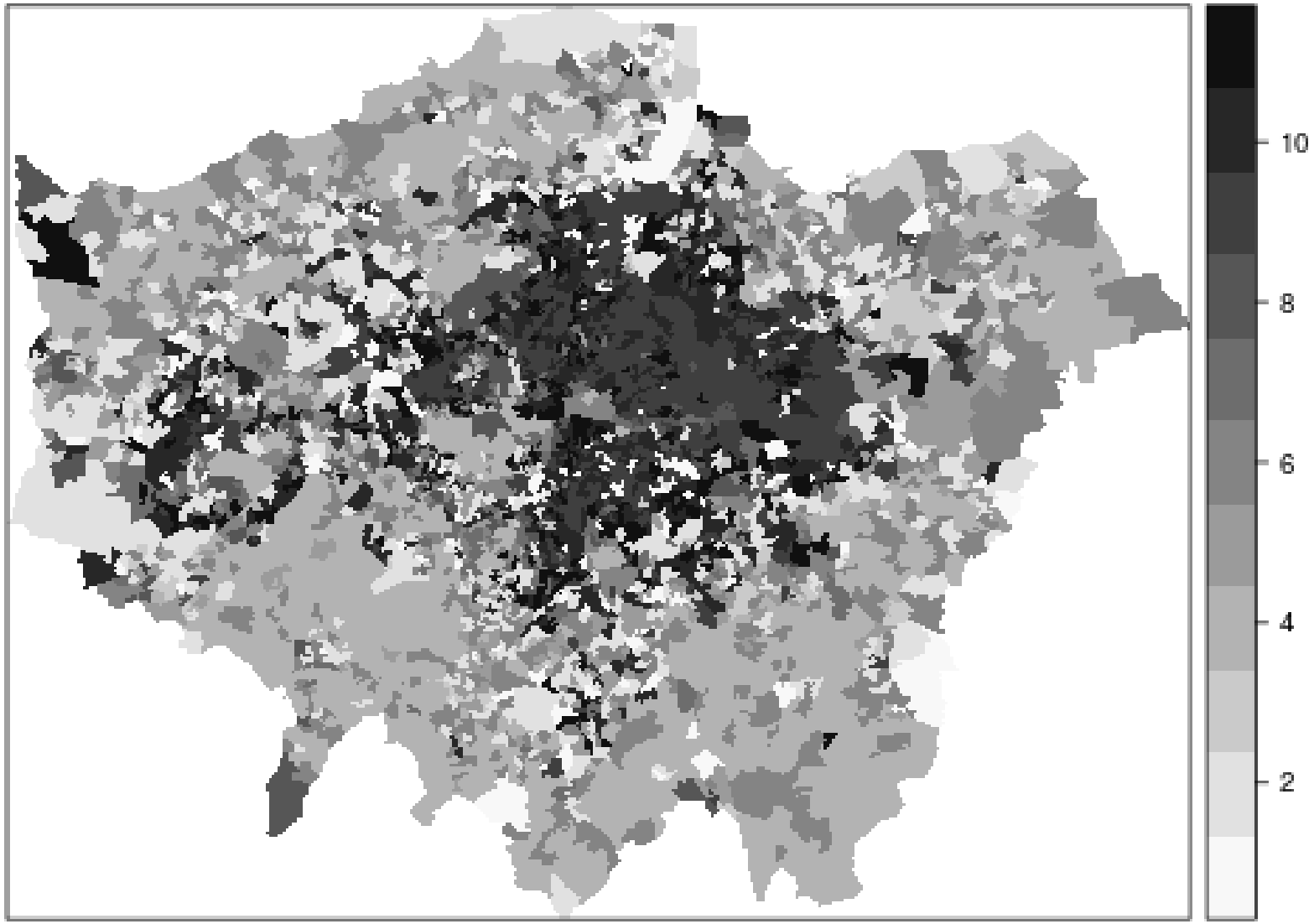
where $\lambda_i = \theta_{Z_i} + W_i\boldsymbol{\beta} + u_i$ and $u = (u_1, \cdots, u_n) \sim N(0, \tau\mathbf{P})$ with $\mathbf{P} = \{P_{ij}\}$ a precision matrix such that

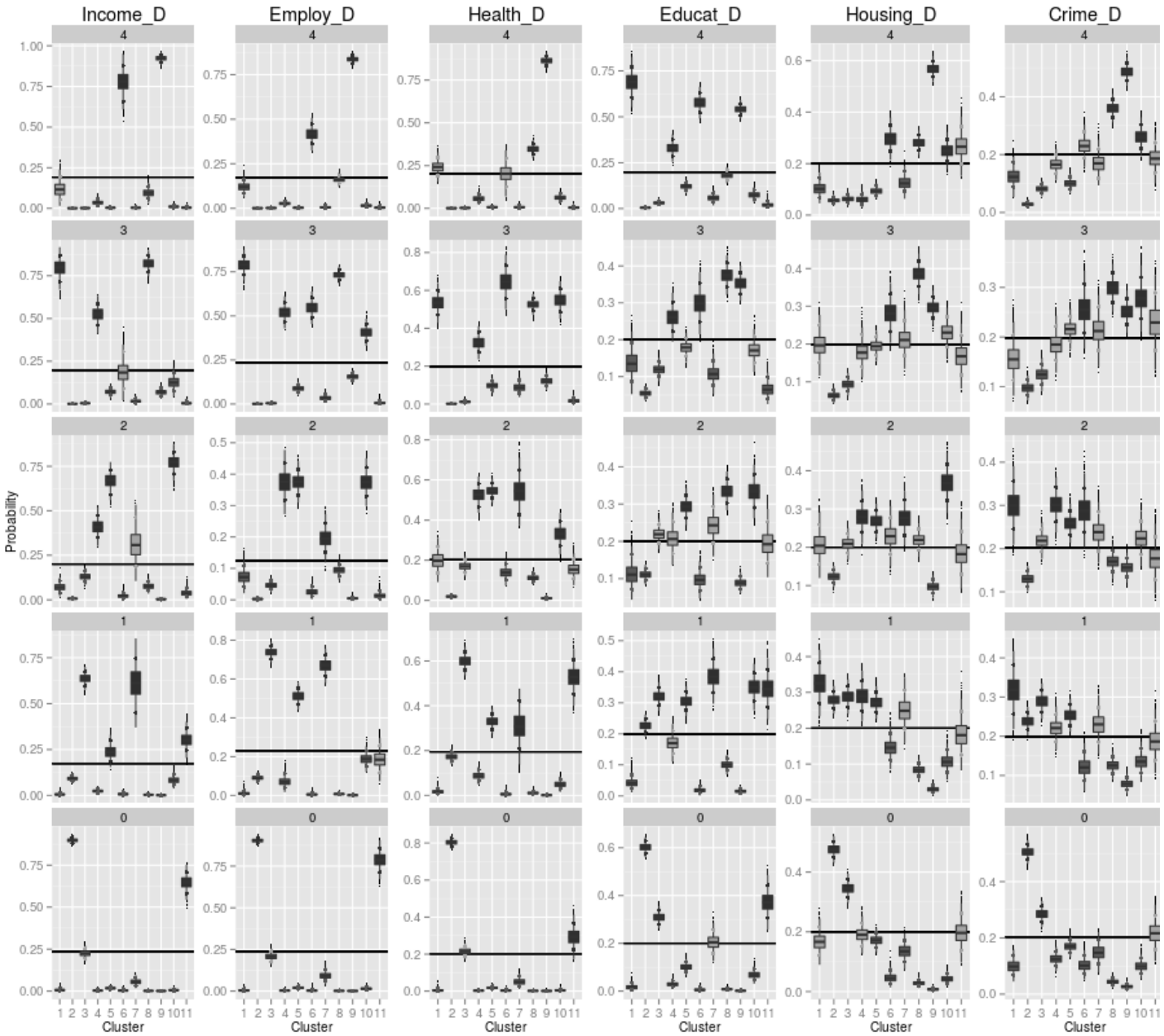
$$P_{ij} = \begin{cases} n_i & \text{if } i = j \\ -I\{i \sim j\} & \text{if } i \neq j \end{cases}$$

where n_i is the number of neighbours of subject i, I is the indicator function and $i \sim j$ indicates that regions i and j are neighbours. The prior of τ is given by

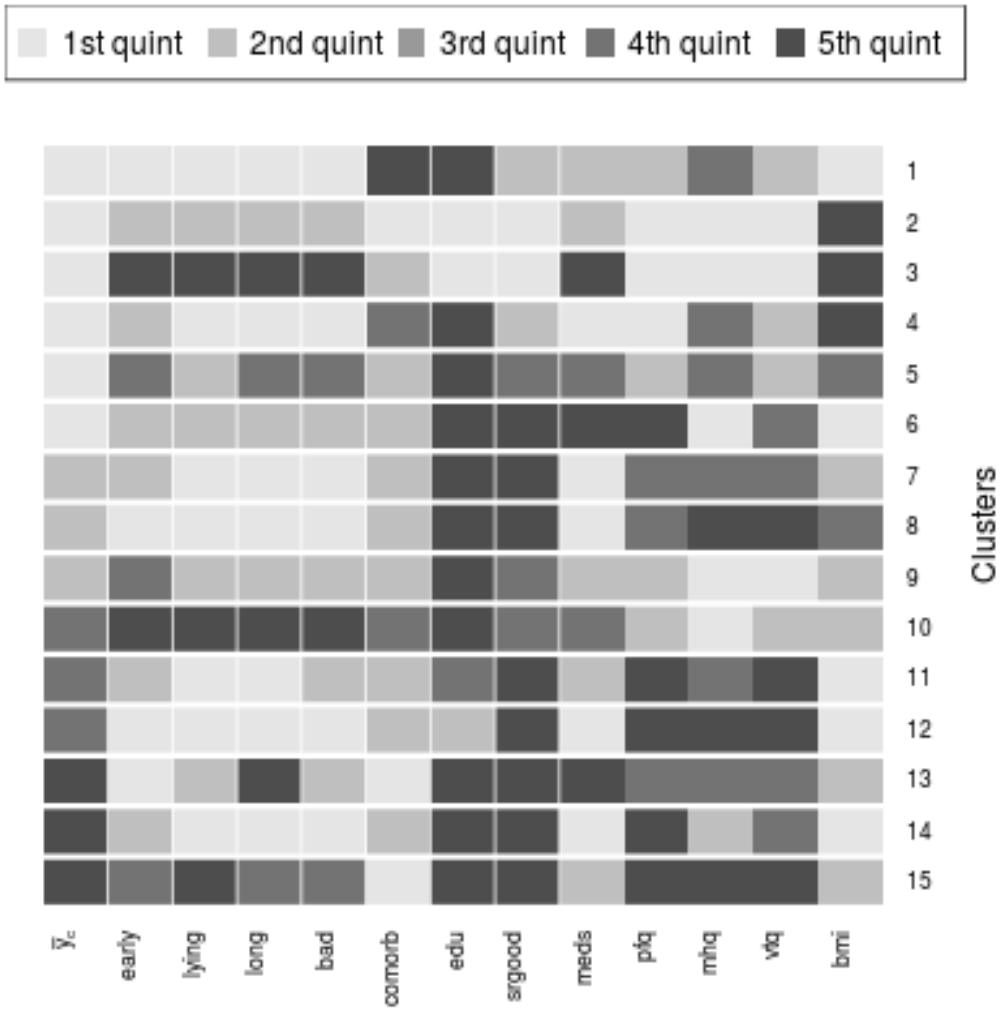
$$\tau \sim \text{Gamma}(a_\tau, b_\tau)$$

Study of the relationship between social deprivation and air pollution in 4,767 small areas in Greater London

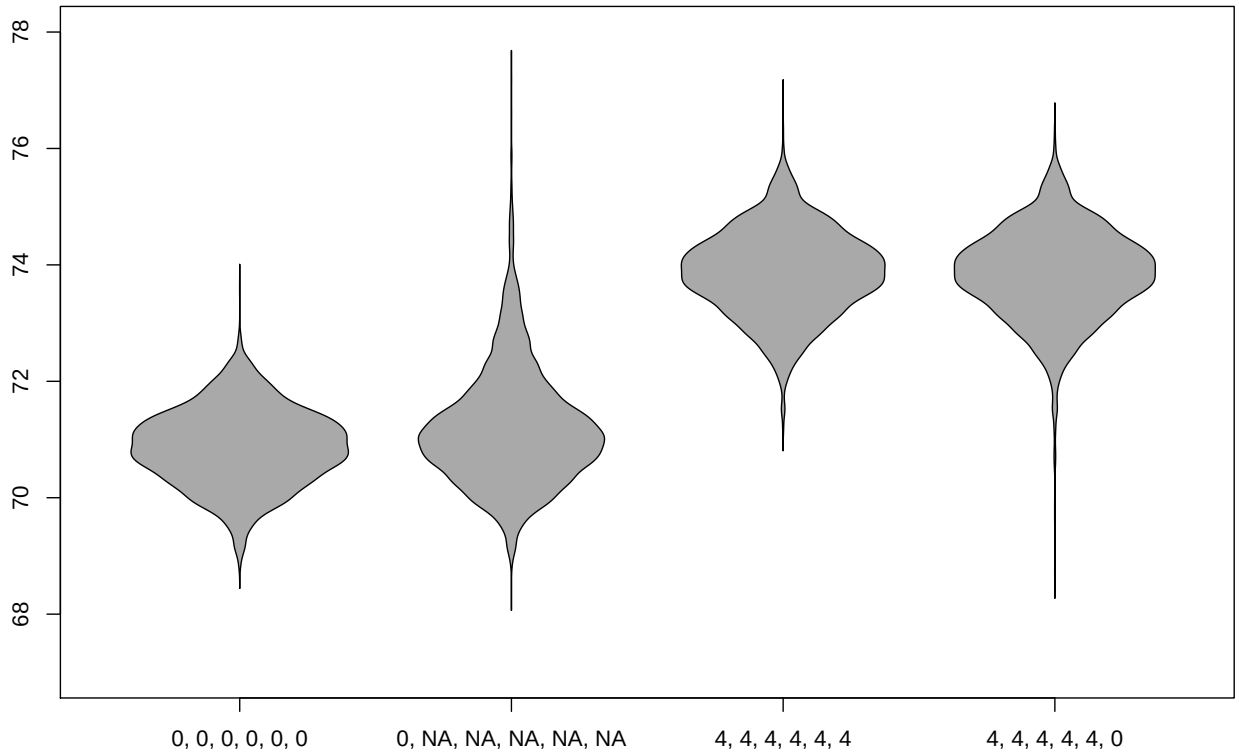



Income_D
Employ_D
Health_D
Educat_D
Housing_D
Crime_D
4
3
2
1
0
Probability
Cluster
1 2 3 4 5 6 7 8 9 10 11

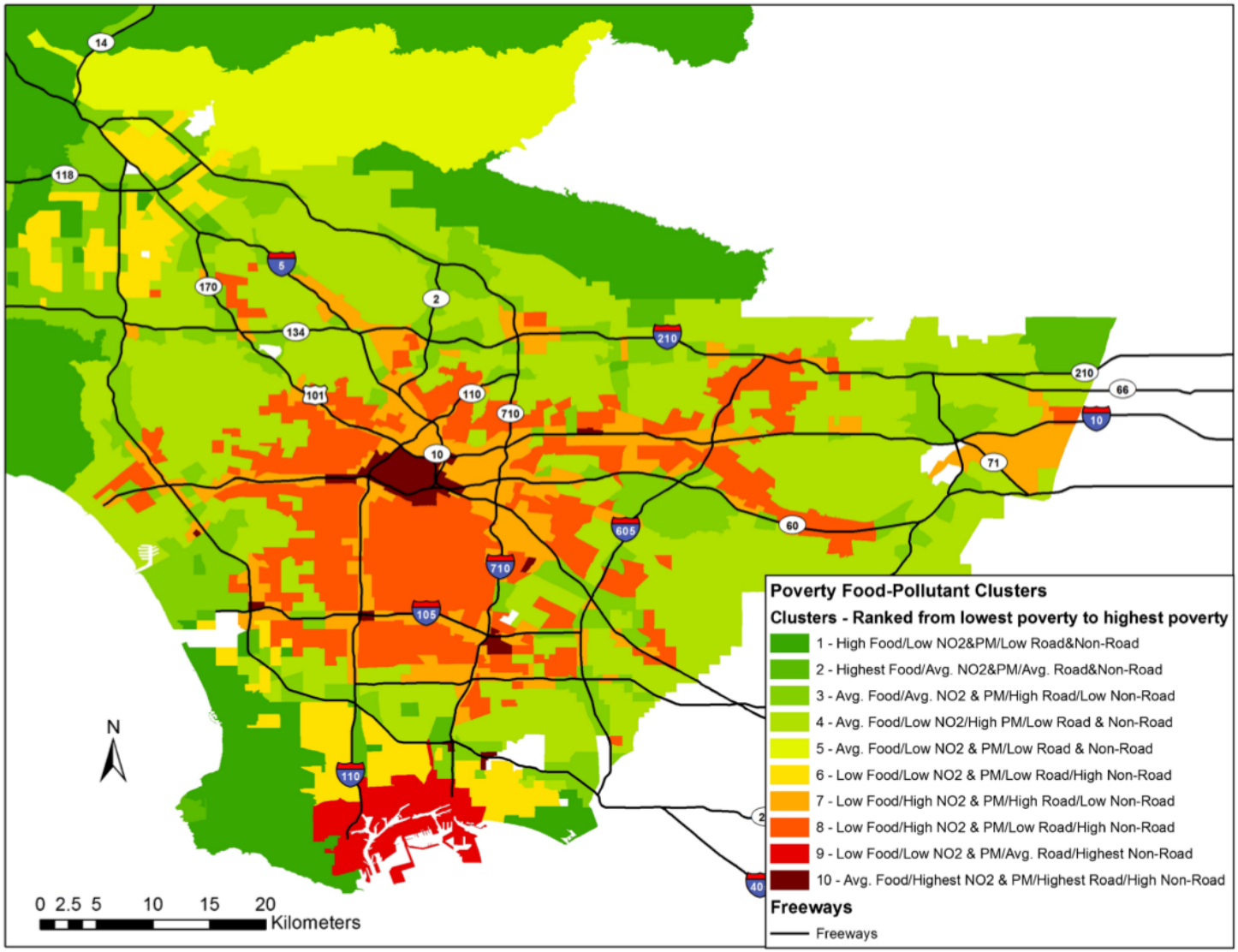
Heatmap for the clusters



Posterior predictive distributions



Air Pollution Exposures, Food Index, and Poverty in Los Angeles County



Study of the relationship between smoking and lung cancer

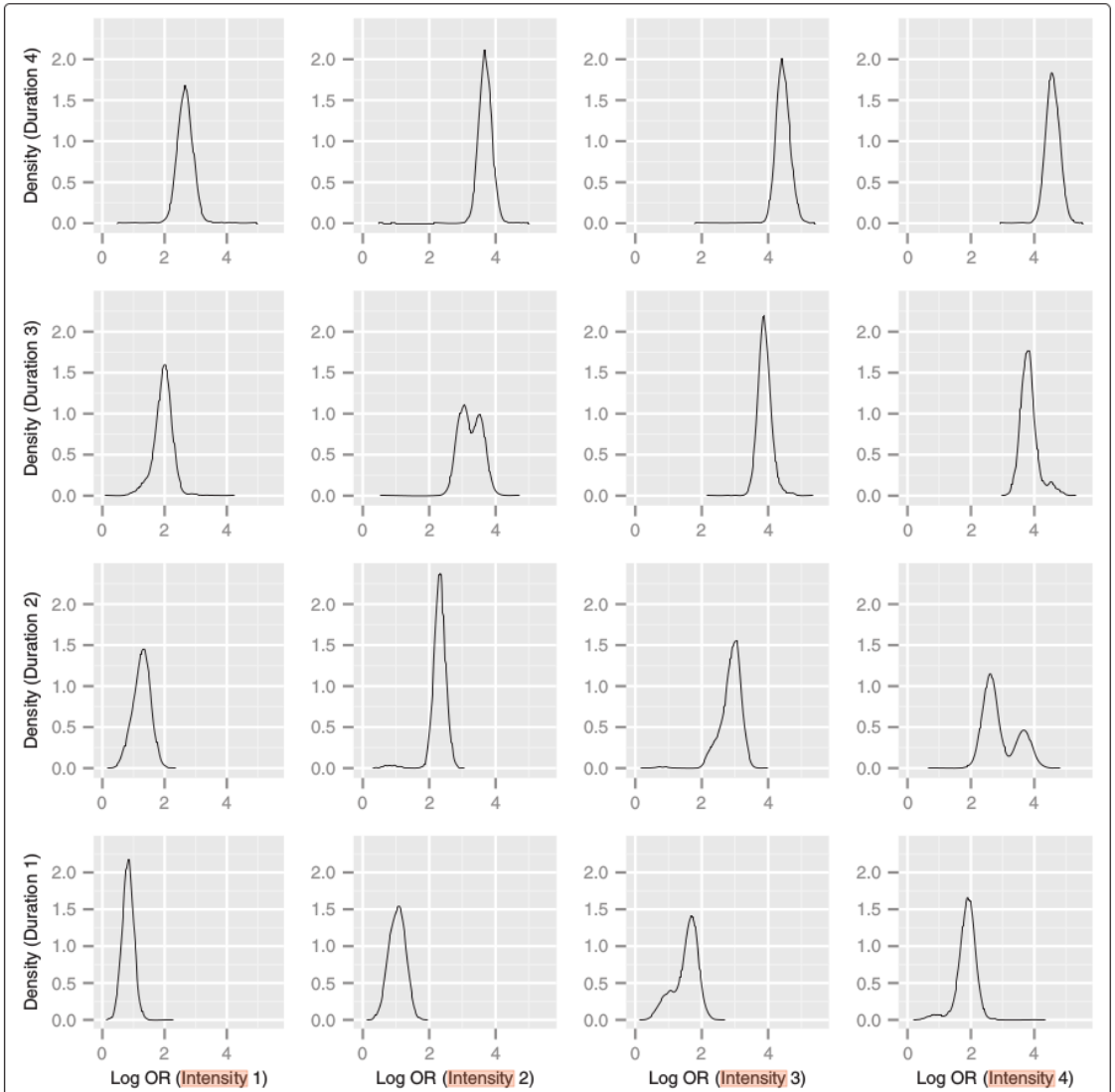


Figure 2 Log odds ratios for intensity and duration combinations. Density estimates of predicted log odds ratios relative to a non-smoking profile, for different intensity and duration combinations. Time since cessation and pack years treated as missing in pseudo-profiles.

Survival Analysis

$$f(\mathbf{x}_i, y_i|\boldsymbol{\phi}, \boldsymbol{\theta}, \boldsymbol{\psi}, \boldsymbol{\beta}, \boldsymbol{z}, \boldsymbol{w}_i) = \sum_{c=1}^{\infty} \psi_c f(\mathbf{x}_i|z_i = c, \phi_c) f(y_i|z_i = c, \theta_c, \boldsymbol{\beta}, \mathbf{w}_i)$$

Survival Analysis

$$f(\mathbf{x}_i, y_i|\boldsymbol{\phi}, \boldsymbol{\theta}, \boldsymbol{\psi}, \boldsymbol{\beta}, \boldsymbol{z}, \boldsymbol{w}_i) = \sum_{c=1}^{\infty} \psi_c f(\mathbf{x}_i|z_i = c, \phi_c) f(y_i|z_i = c, \theta_c, \boldsymbol{\beta}, \mathbf{w}_i)$$

- Survival response Weibull with global shape parameter

$$f_Y(y_i|\boldsymbol{\Theta}, \Lambda, W, Z) = h_Y(y_i|\theta_{Z_i}, \nu, \beta, W_i)^{d_i} S_Y(y_i|\theta_{Z_i}, \nu, \beta, W_i)$$

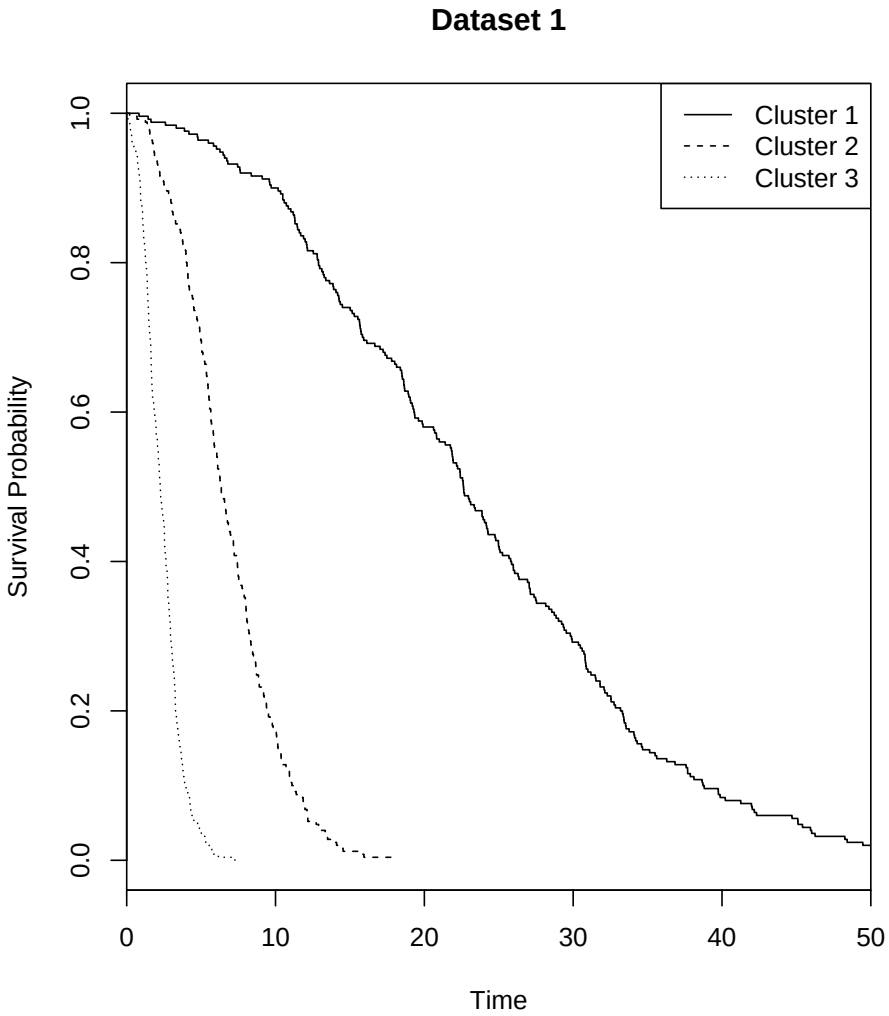
- Survival response Weibull with cluster-specific shape parameter

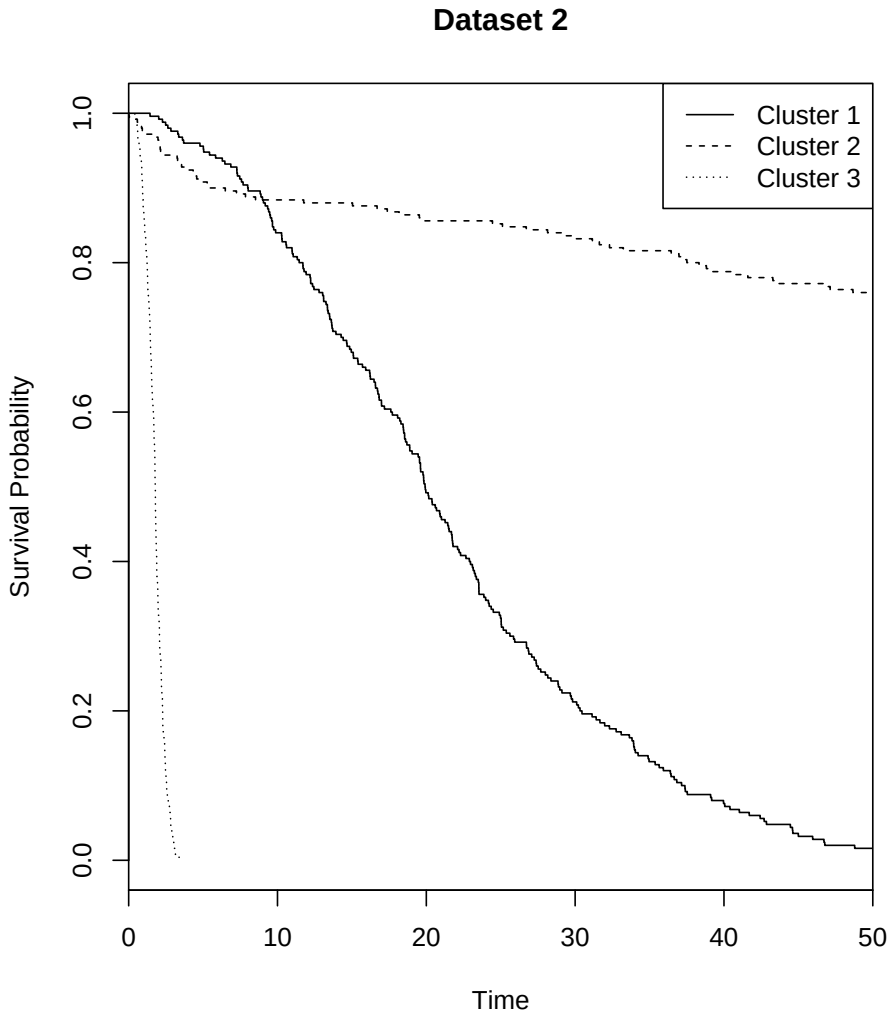
$$f_Y(y_i|\boldsymbol{\Theta}, \Lambda, W, Z) = h_Y(y_i|\theta_{Z_i}, \nu_{Z_i}, \beta, W_i)^{d_i} S_Y(y_i|\theta_{Z_i}, \nu_{Z_i}, \beta, W_i)$$

where h is the hazard function, S is the survival function and Y is the lifetime of an individual. The indicator d_i is defined as follows.

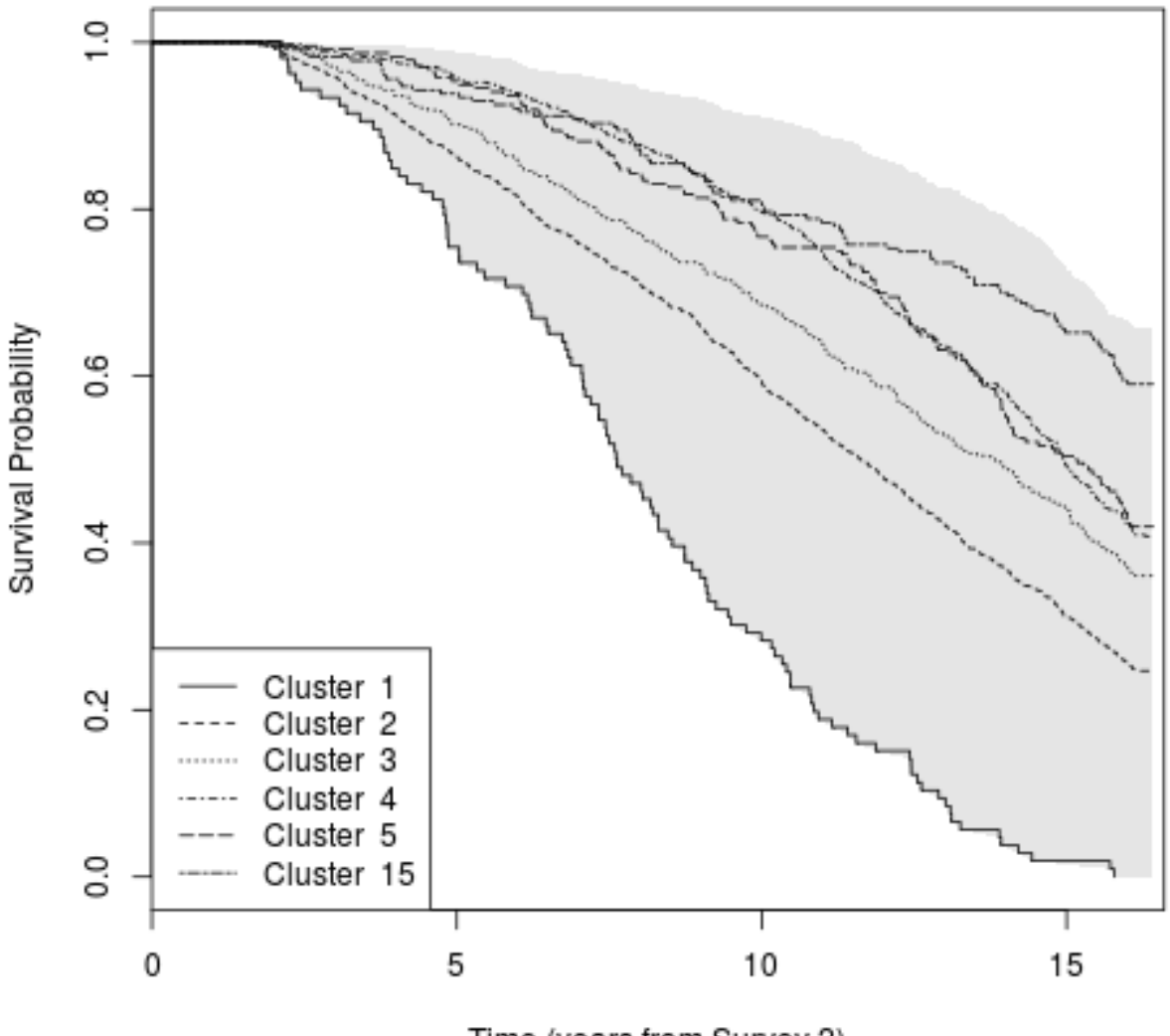
$$d_i = \begin{cases} 0 & \text{if the individual is censored} \\ 1 & \text{if the individual experiences the event} \end{cases}$$

Application to simulated data

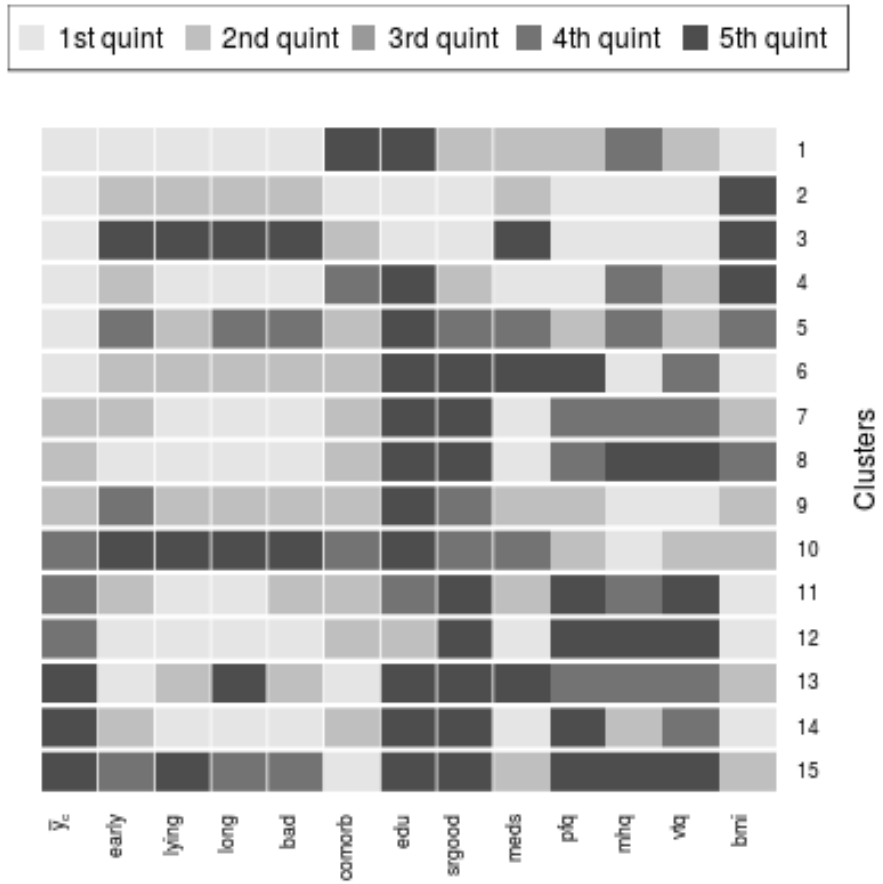




Application to real data (sleep survey)



Application to real data (sleep survey)



References

Liu X., Liverani S., Yu K. (2019). Modelling Tails for Collinear Data: Quantile Profile Regression. Submitted

Liverani S., Lavigne A. and Blangiardo M. (2016) Modelling collinear and spatially correlated data. Spatial and Spatio-temporal Epidemiology 18, 63-73.

Mattei F., Liverani S., et al. (2016) A multidimensional analysis of the effect of occupational exposure to organic solvents on lung cancer risk: the ICARE study. Occupational and Environmental Medicine 73 (6), 368-377.

Pirani, M. et al (2015) Analysing the health effects of simultaneous exposure to airborne particles. Environment International, 79, 56-64.

Liverani, S. et al (2015) PReMiuM: An R package for Profile Regression Mixture Models using Dirichlet Processes. Journal of Statistical Software, 64(7), 1-30.

Hastie, D. I., Liverani, S. and Richardson, S. (2015) Sampling from Dirichlet process mixture models with unknown concentration parameter: Mixing issues in large data implementations. Statistics and Computing 25 (5), 1023-1037.

Hastie, D. I., Liverani, S. et al. (2013) A semi-parametric approach to estimate risk functions associated with multi-dimensional exposure profiles. BMC Medical Research Methodology. 13 (1), 129.