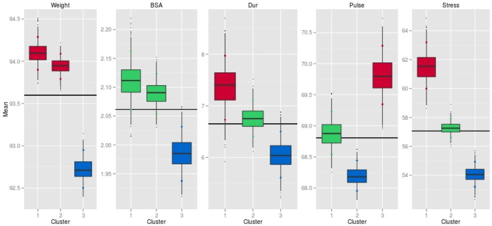
Dirichlet process Bayesian clustering with the R package PReMiuM

Dr Silvia Liverani
Brunel University London



July 2015

Outline

- Motivation
- Method
- R package `PReMiuM`
- Examples

Many collaborators

- John Molitor (University of Oregon)
- Sylvia Richardson (Medical Research Centre Biostatistics Unit)
- Michail Papathomas (University of St Andrews)
- David Hastie
- Aurore Lavigne (University of Lille 3, France)
- Lucy Leigh (University of Newcastle, Australia)
- ...

Multicollinearity

- Goal of epidemilogical studies is to investigate the joint effect of different covariates / risk factors on a phenotype...
- ... but highly correlated risk factors create collinearity problems!

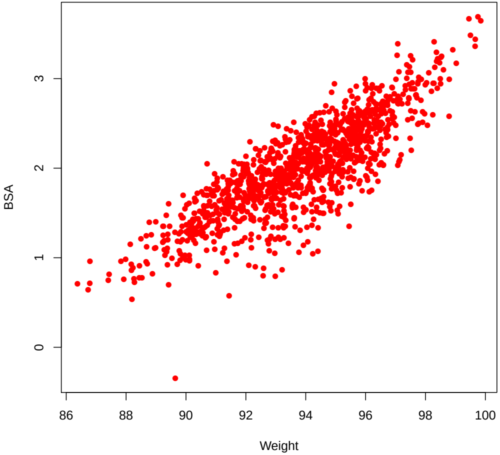
Multicollinearity

- Goal of epidemilogical studies is to investigate the joint effect of different covariates / risk factors on a phenotype...
- ... but highly correlated risk factors create collinearity problems!

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 (x_3 = Dur, in years)
- basal pulse (x_4 = Pulse, in beats per minute)
- stress index (x_5 = Stress)



BP=y, Weight = x_1, BSA = x_2

- Highly correlated risk factors create collinearity problems, causing instability in model estimation

Model	$\hat{\beta}_1$	SE $\hat{\beta}_1$	$\hat{\beta}_2$	SE $\hat{\beta}_2$
$y \sim x_1$	2.64	0.30	–	–
$y \sim x_2$	–	–	3.34	1.33
$y \sim x_1 + x_2$	6.58	0.53	-20.44	2.28

- Highly correlated risk factors create collinearity problems, causing instability in model estimation

Model	$\hat{\beta}_1$	SE $\hat{\beta}_1$	$\hat{\beta}_2$	SE $\hat{\beta}_2$
$y \sim x_1$	2.64	0.30	–	–
$y \sim x_2$	–	–	3.34	1.33
$y \sim x_1 + x_2$	6.58	0.53	-20.44	2.28

- Effect 1: the estimated regression coefficient of any one variable depends on which other predictor variables are included in the model.
- Effect 2: the precision of the estimated regression coefficients decreases as more predictor variables are added to the model.

Issues caused by

- correlated risk factors
- interacting risk factors

Issues caused by

- correlated risk factors
- interacting risk factors

$$\Downarrow$$

Profile regression

- **partitions the multi-dimensional risk surface into groups having similar risks**
- investigation of the joint effects of multiple risk factors
- jointly models the covariate patterns and health outcomes
- flexible but tractable Bayesian model

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

Statistical Framework

- Joint covariate and response model

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

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 discrete covariates

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

- For example, for Bernoulli response

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

Statistical Framework

- Joint covariate and response model

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

- Prior model for the mixture weights ψ_c
  - stick-breaking priors (constructive definition of the Dirichlet Process)

$$\mathbb{P}(Z_i = c|\boldsymbol{\psi}) = \psi_c \qquad \psi_1 = V_1$$

$$\psi_c = V_c \prod_{l<c}(1 - V_l) \qquad V_c \sim \mathrm{Beta}(1, \alpha)$$

  - larger concentration parameter α the more evenly distributed is the resulting distribution.
  - smaller concentration parameter α the more sparsely distributed is the resulting distribution, with all but a few parameters having a probability near zero

Implementation: R package PReMiuM

We have implemented profile regression in C++ within the R package PReMiuM.

- Binary, binomial, categorical, Normal, Poisson and survival outcome
- Allows for spatial correlation
- Fixed effects (global parameters) including also spatial CAR term
- Normal and/or discrete covariates
- Dependent or independent slice sampling (Kalli et al., 2011) or truncated Dirichlet process model (Ishwaran and James, 2001)
- Fixed alpha or update alpha, or use the Pitman-Yor process prior
- Handles missing data

Implementation: R package PReMiuM

We have implemented profile regression in C++ within the R package PReMiuM.

- Allows users to run predictive scenarios
- Performs post processing
- Contains plotting functions

Currently working on:

- Quantile profile regression
- Enriched Dirichlet processes

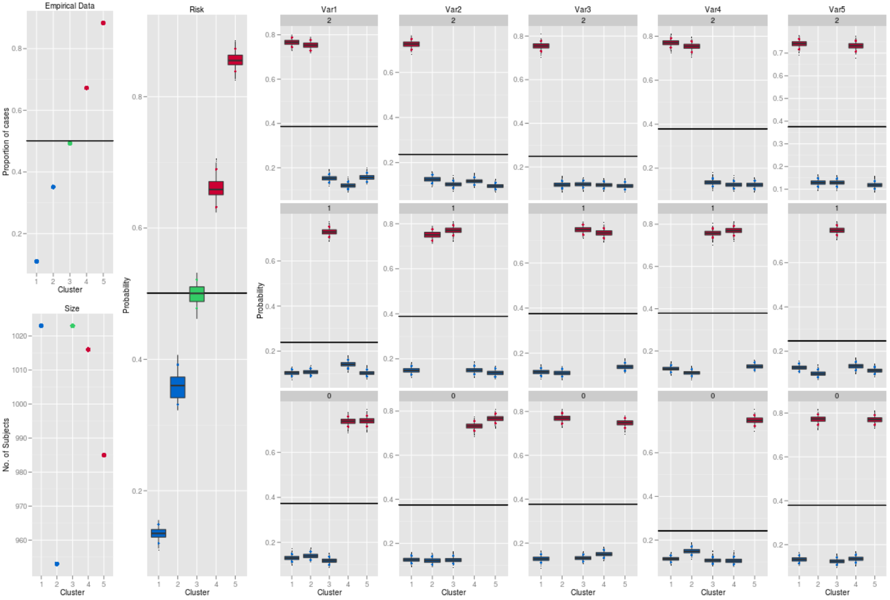
Example: Simulated data

The profiles are given by

y : outcome, Bernoulli

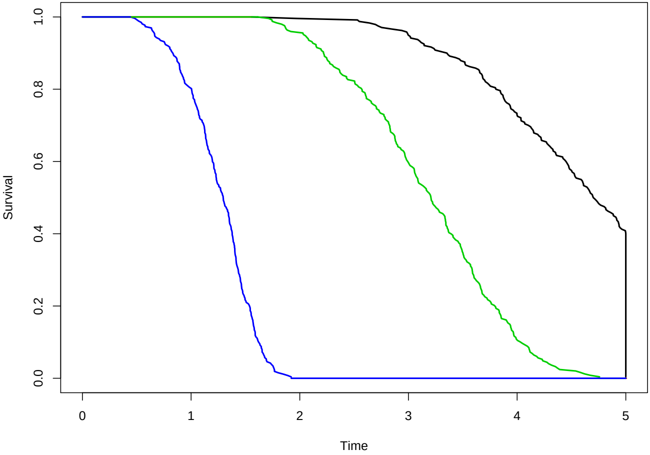
x : 5 covariates, all discrete with 3 levels

w : 2 fixed effects, continuous or discrete

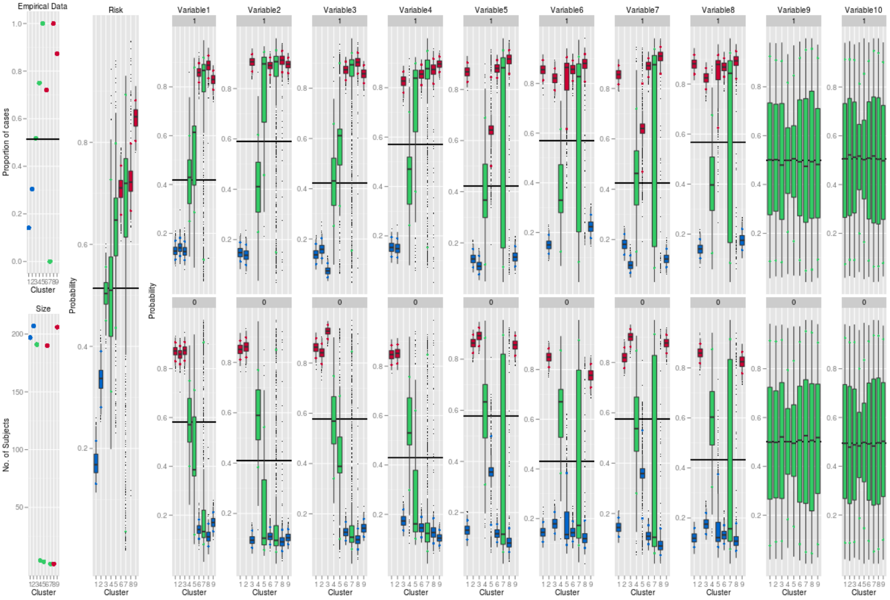


Empirical Data
Proportion of cases
Cluster
Size
No. of Subjects
Risk
Probability
Var1
Var2
Var3
Var4
Var5

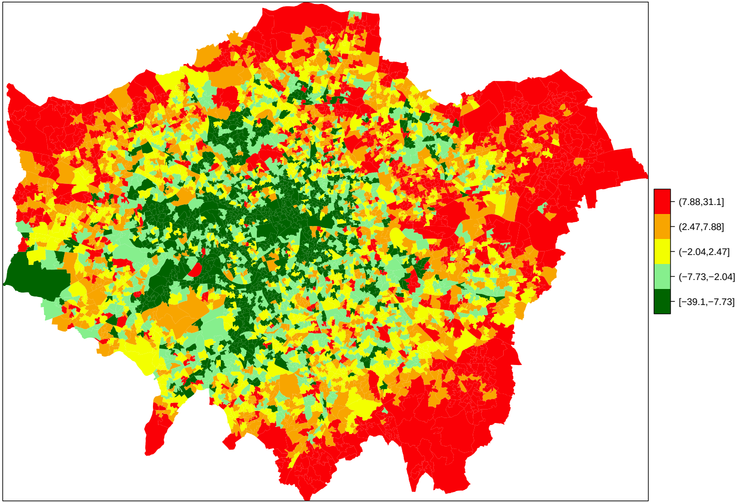
Survival response with censoring: sleep study



Variable selection



Spatial correlated response: deprivation in London



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 for Statistical Software, 64(7), 1-30.**
- J. T. Molitor, M. Papathomas, M. Jerrett and S. Richardson (2010) Bayesian Profile Regression with an Application to the National Survey of Childrens Health, Biostatistics, 11, 484-498.
- Molitor, J., Brown, I. J., Papathomas, M., Molitor, N., Liverani, S., Chan, Q., Richardson, S., Van Horn, L., Daviglus, M. L., Stamler, J. and Elliott, P. (2014) Blood pressure differences associated with DASH-like lower sodium compared with typical American higher sodium nutrient profile: INTERMAP USA. Hypertension.
- 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. To appear in Statistics and Computing.
- 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 .