Dirichlet process Bayesian clustering with the R package PReMiuM

Dr Silvia Liverani
Brunel University London

July 2015
Outline

- Motivation
- Method
- R package PReMiuum
- 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)
- ...
Goal of epidemiological studies is to investigate the joint effect of different covariates / risk factors on a phenotype...

... but highly correlated risk factors create collinearity problems!
Motivation

Multicollinearity

- Goal of epidemiological 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 ($y = \text{BP}$, in mm Hg) and:
- weight ($x_1 = \text{Weight}$, in kg)
- body surface area ($x_2 = \text{BSA}$, in sq m)
- duration of hypertension ($x_3 = \text{Dur}$, in years)
- basal pulse ($x_4 = \text{Pulse}$, in beats per minute)
- stress index ($x_5 = \text{Stress}$)
BP = y, Weight = x_1, BSA = x_2
Highly correlated risk factors create collinearity problems, causing instability in model estimation.

<table>
<thead>
<tr>
<th>Model</th>
<th>$\hat{\beta}_1$</th>
<th>SE $\hat{\beta}_1$</th>
<th>$\hat{\beta}_2$</th>
<th>SE $\hat{\beta}_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$y \sim x_1$</td>
<td>2.64</td>
<td>0.30</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>$y \sim x_2$</td>
<td>-</td>
<td>-</td>
<td>3.34</td>
<td>1.33</td>
</tr>
<tr>
<td>$y \sim x_1 + x_2$</td>
<td>6.58</td>
<td>0.53</td>
<td>-20.44</td>
<td>2.28</td>
</tr>
</tbody>
</table>
Highly correlated risk factors create collinearity problems, causing instability in model estimation.

<table>
<thead>
<tr>
<th>Model</th>
<th>$\hat{\beta}_1$</th>
<th>SE $\hat{\beta}_1$</th>
<th>$\hat{\beta}_2$</th>
<th>SE $\hat{\beta}_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$y \sim x_1$</td>
<td>2.64</td>
<td>0.30</td>
<td>$-,$</td>
<td>$-,$</td>
</tr>
<tr>
<td>$y \sim x_2$</td>
<td>$-,$</td>
<td>$-,$</td>
<td>3.34</td>
<td>1.33</td>
</tr>
<tr>
<td>$y \sim x_1 + x_2$</td>
<td>6.58</td>
<td>0.53</td>
<td>-20.44</td>
<td>2.28</td>
</tr>
</tbody>
</table>

- **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

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

\[ f(x_i, y_i | \phi, \theta, \psi, \beta) = \sum_c \psi_c f(x_i | z_i = c, \phi_c) f(y_i | z_i = c, \theta_c, \beta, w_i) \]
Statistical Framework

- Joint covariate and response model

\[
f(x_i, y_i|\phi, \theta, \psi, \beta) = \sum_c \psi_c f(x_i|z_i = c, \phi_c) f(y_i|z_i = c, \theta_c, \beta, w_i)
\]

- For example, for discrete covariates

\[
f(x_i|z_i = c, \phi_c) = \prod_{j=1}^{J} \phi_{z_i,j,x_i,j}
\]

- For example, for Bernoulli response

\[
\text{logit}\{p(y_i = 1|\theta_c, \beta, w_i)\} = \theta_c + \beta^T w_i
\]
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 \( \psi_c \)
  - stick-breaking priors (constructive definition of the Dirichlet Process)

\[ \mathbb{P}(Z_i = c|\psi) = \psi_c \quad \psi_1 = V_1 \]

\[ \psi_c = V_c \prod_{l < c} (1 - V_l) \quad V_c \sim \text{Beta}(1, \alpha) \]

- Larger concentration parameter \( \alpha \) the more evenly distributed is the resulting distribution.
- Smaller concentration parameter \( \alpha \) the more sparsely distributed is the resulting distribution, with all but a few parameters having a probability near zero.
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
Applications and features of the model

Survival response with censoring: sleep study

![Graph showing survival response with censoring]

- Time
- Survival

Silvia Liverani (Brunel University London)
Spatial correlated response: deprivation in London
References


