Stats 225: Bayesian Analysis

More on Dirichlet Process Mixtures

Babak Shahbaba
UC Irvine

Overview

- In the previous lecture, we discussed Dirichlet process mixture (DPM) models for nonparametric clustering and density estimation.
- In this lecture, we will discuss some advanced Dirichlet process models.
- We specifically discuss the applications of DPM models in genomics and diagnostics.
- We will also discuss its extensions to nonlinear predictive models and biclustering.

Genomics

In Collaboration with Wes Johnson

Shahbaba, B., Johnson, W.O. (2013), Bayesian Nonparametric Variable Selection as an Exploratory Tool for Discovering Differentially Expressed Genes, *Statistics in Medicine*, 30(12), 2114-26.

Introduction

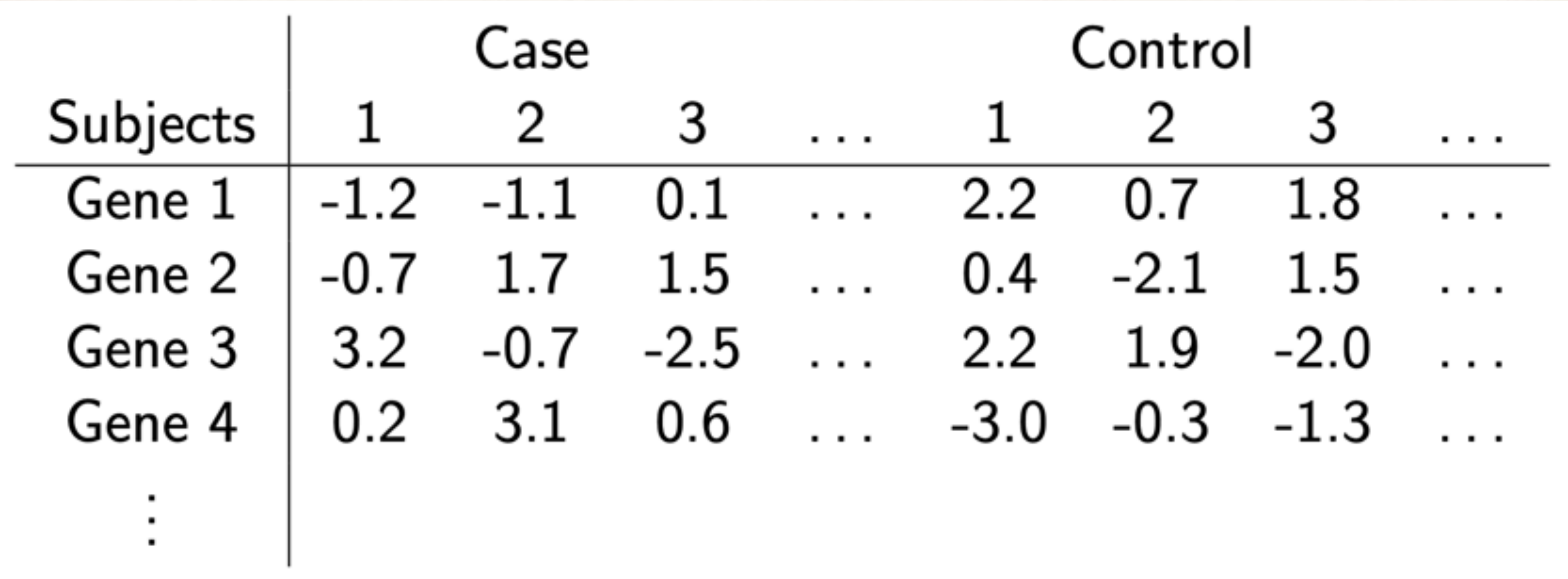
- Large-scale genomic studies examine thousands of genes simultaneously

- Objective is to identify a small number of genes for follow-up studies

- We divide the set of genes into several subgroups according to their degrees of "relevance," or potential effect, in relation to the outcome of interest (e.g., disease status)

- This could lead to a better identification of the underlying structure in our data and ultimately, genes that ``matter''

Data



	Case				Control			
Subjects	1	2	3	. . .	1	2	3	. . .
Gene 1	-1.2	-1.1	0.1	. . .	2.2	0.7	1.8	. . .
Gene 2	-0.7	1.7	1.5	. . .	0.4	-2.1	1.5	. . .
Gene 3	3.2	-0.7	-2.5	. . .	2.2	1.9	-2.0	. . .
Gene 4	0.2	3.1	0.6	. . .	-3.0	-0.3	-1.3	. . .
⋮								

Multiple hypothesis testing

- Most current methods applied to high-throughput experiments are extensions of the classical hypothesis testing approach (i.e., when there is a single hypothesis).

- For each gene, $\mathscr{G}_i$, where $i = 1, \ldots, N$, there is a corresponding [null] hypothesis, H_i, stating that there is no change in gene expression between two biological conditions (i.e., diseased vs. healthy).

- The observed expression values $\{Y_{ijk} : j = 1,\ldots,n_{ik}, k = 0,1\}$ are used to compute a simple test statistic T_i for gene i.

- Statistics above a certain cutoff are deemed significant, after adjustment to control the family-wise Type I error rate or false discovery rate (FDR).

FDR

- FDR is one of the most widely used measures for coping with multiplicity.

- Suppose we observe values for $T_1, T_2, \ldots, T_N$ and obtain the corresponding p-values:

$$p_j = P(T_j \geq t_j \,|\, H_j)$$

- Reject H_j if $p_j < \lambda$

$$FDR(\lambda) = E(\text{Proportion of true } H_j \,|\; \text{ rejected })$$

FDR

- Instead of *p*-values, it is convenient to work with

$$z_j = \Phi^{-1}[P(T_j \geq t_j | H_j)]$$

- Under H_j, $z_j \sim N(0,1)$.

- Large-scale testing situations however permit estimation of the null distribution.

- The following mixture density is assumed for the transformed p-values:

$$f(z) = p_0 f_0(z) + (1 - p_0) f_1(z)$$

FDR

- Under this model, if all inputs were known, then the Bayesian approach based on zero/one loss for just a single hypothesis rejects H_j if

$$\mathrm{fdr}(z_j) \equiv p_0 f_0(z_j)/f(z_j) < \lambda$$

- Efron et.~al. (2001) use empirical Bayes approach to estimate fdr(z).

- Their approach is referred to as *locFDR*.

Optimal discovery procedure

- Storey (2007) proposed the *optimal discovery procedure* (ODP): minimizing *missed discovery rate* (false negative) for each fixed FDR (false positive rate)

- Suppose $z_j \sim f(z_j; \mu_j)$, where f is some distribution indexed by an unknown parameter μ_j.

- The ODP for testing $H_j : \mu_j \in A$ is then based on a single significance thresholding statistic,

$$S_{ODP}(z_j) = \frac{\sum_{\mu_j \notin A} f(z_j; \mu_j)}{\sum_{\mu_j \in A} f(z_j; \mu_j)}$$

- We reject the null hypothesis $H_j$if $S_{ODP}(z_j) \geq \lambda$ for some $0 \leq \lambda < \infty$.

Bayesian discovery procedure

- Guindani et.~al.(2009) showed that the ODP could be interpreted as approximate Bayes rule under a semiparametric model.

- They proposed a Bayesian discovery procedure (BDP) that improves the approximation and allows for multiple shrinkage in clusters implied by a Dirichlet process mixture model:

$$\begin{aligned} z_i \,|\, \mu_i &\sim f(z_i \mid \mu_i), i = 1, \ldots, N \\ \mu_i \,|\, G &\sim G \\ G &\sim \mathscr{D}(G_0, \gamma) \\ G_0 &= p_0 h_{\{0\}}(\,.\,) + (1 - p_0) h_{\{0\}^c}(\,.\,) \end{aligned}$$

- Here, $f(z_i \mid \mu_i)$ is typically considered to be a normal distribution, $N(z_i \mid \mu_i, \sigma^2)$.

- The distribution $h_{\{0\}}$ is point mass at zero and $h_{\{0\}^c}$ is set to a continuous distribution such as $N(0,\sigma^2)$.

Bayesian discovery procedure

- Latent cluster membership indicators, s_i, partition the observations into clusters such that

$$s_i = s_k \qquad \text{if } \mu_i = \mu_k$$

- The label $s_i = 1$ is reserved for the null distribution; that is, $s_i = 1$ when $\mu_i = 0$.

- Guindani et al. (2009) showed that thresholding based on the measure

$$v_i = 1 - \sum_{b=1}^{B} I(s_i^{(b)} = 1)/B$$

can be approximated by $\hat{S}_{ODP}$.

Nonparametric relevance determination

- We (Shahbaba and Johnson) developed an alternative model:

$$z_j \,|\, \tau_j^2 \sim N(0, \tau_j^2)$$

$$\tau_j^2 \,|\, G \sim G$$

$$G \sim \mathscr{D}(G_0, \gamma)$$

- We refer to our model as Bayesian Relevance Determination: *BRD*.

Nonparametric relevance determination

- Alternatively, let y_{ijk} denote the j^{th} observed gene expression value in group k for gene i.

$$y_{ijk} \mid \alpha_i, \beta_i \sim N(\alpha_i + \beta_i x_{ijk}, \sigma_i^2)$$

- Our model for the regression coefficients is hierarchical where the first level assigns independent normal priors to the β_is with distinct variances, namely

$$\beta_i \mid \tau_i^2 \sim N(0, \tau_i^2)$$

- We assume a Dirichlet Process prior for τ_i^2

$$\tau_i^2 \mid G \sim G$$

$$G \sim \mathcal{D}(G_0, \gamma)$$

Nonparametric relevance determination

- We could define a relevance measure similar to that of Guindani et.~al.(2009)

- To this end, we denote $min_j\{\tau_j^2\}$ at each iteration as ϕ_0^2

- For gene*i*, we create a binary indicator, s_i, which is set to 1 when $\tau_i^2 = \phi_0^2$, and zero otherwise

- Similar to the measure proposed by Guindani et.~al.(2009), we can use B posterior Monte Carlo samples to calculate

$$v_i = 1 - \sum_{b=1}^{B} I(s_i^{(b)} = 1)/B$$

Nonparametric relevance determination

- Both methods use a Dirichlet process mixture of normals for modeling gene expression data
- For BDP, the DP prior is assumed for the means of the normal distributions (all mixture components share the same variance)
- An alternative variation of BDP mixes on the means and the variances
- We use the DP prior on the variances, τ^2, and fix means at zero
- Our model provides a natural framework for ranking mixture components, and in turn, for ranking the genes assigned to each component with respect to their potential importance

Nonparametric relevance determination

- This approach is related to *robust Bayesian inference.*

- It can be regarded as Dirichlet Process Scale Mixture of Normals (e.g., Andrews & Mallows, 1974; West 1984; Carvalho et al., 2009):

$$Y_i \sim N(0,\sigma_i^2)$$

$$\sigma_i^2 \sim g(\sigma_i^2)$$

- When σ^2 has Inv-Gamma($\nu/2,\nu/2$) distribution, Y has a t-distribution with ν degrees of freedom.

- The distribution of Y will become *Laplace* or *horseshoe* (Carvalho et al. 2010) if instead of Inv-Gamma we use exponential or half-Cauchy respectively.

Diagnostics

Akhavan, S., Holsclaw, T., Shahbaba, B., Gillen, D. (2018), A Flexible Joint Longitudinal-Survival Model for Analysis of End-Stage Renal Disease Data (2018),arXiv:1807.02239 .

Capturing Albumin Volatility

Recall that we used the following model for longitudinal albumin measurement:

$$Y^L = X\beta + \mathbf{W}(\mathbf{t}) + \epsilon$$

where $\mathbf{W}(\mathbf{t})$ are realizations from a Gaussian Process with mean zero and covariance function

$$C(t,t') = \kappa^2 e^{-\lambda|t-t'|^2}$$

Here, κ^2 controls the height of the oscillations and λ controls the correlation length between realizations.

Capturing Albumin Volatility

Recall that larger values of κ^2 produce higher volatility around the mean function

Values of κ^2 near 0 produce nearly linear trajectories

This makes the GP model a natural choice for the scientific problem being considered as we can focus on κ^2 as the functional of interest:

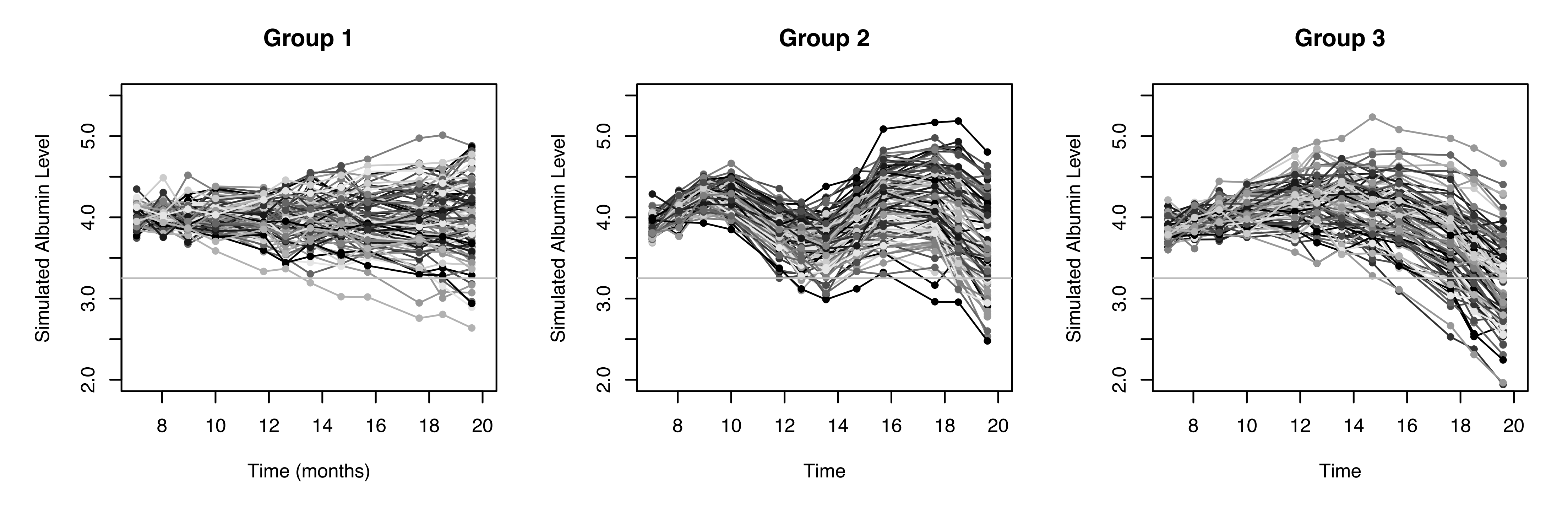
$$Y_i^L \,|\, \theta \sim N_{J_i}(X_i\beta_i, \kappa_i^2 K(\lambda) + \sigma^2 I)$$

Note that λ is shared by all subjects.

Capturing Albumin Volatility

For κ_i^2, to ensure model flexibility we specify a prior distribution with a Dirichlet process (DP) mixture prior:

$$\begin{aligned}
&\pi(\kappa_i^2 \mid G) \sim G \\
&\pi(G) \sim DP(G_0 = \Gamma^{-1}(A, B), \gamma) \\
&\pi(\alpha) \sim \Gamma(2,4) \\
&\pi(A) \sim \Gamma(2,1) \\
&\pi(B) \sim \Gamma(1,1)
\end{aligned}$$



Capturing Albumin Volatility

The proposed model provides a flexible framework for modeling nonlinear trajectories while allowing for linear patterns for some clusters

However, it does not automatically identify a cluster of subjects for whom the longitudinal patterns are approximately linear

In a manner similar to Guindani et al (2009), we can account for this by considering a spike-and-slab prior for κ_i^2 of the form:

$$\pi(\kappa_i^2 \mid p, G) \sim pU_D + (1-p)G_{D^c}$$

$$\pi(G_{D^c}) \sim DP(G_0^*, \gamma)$$

$$\pi(p) \sim \text{Beta}(p_a, p_b)$$

where U_D is a Uniform distribution with small support near zero (taken to be Unif(0,.1) here).

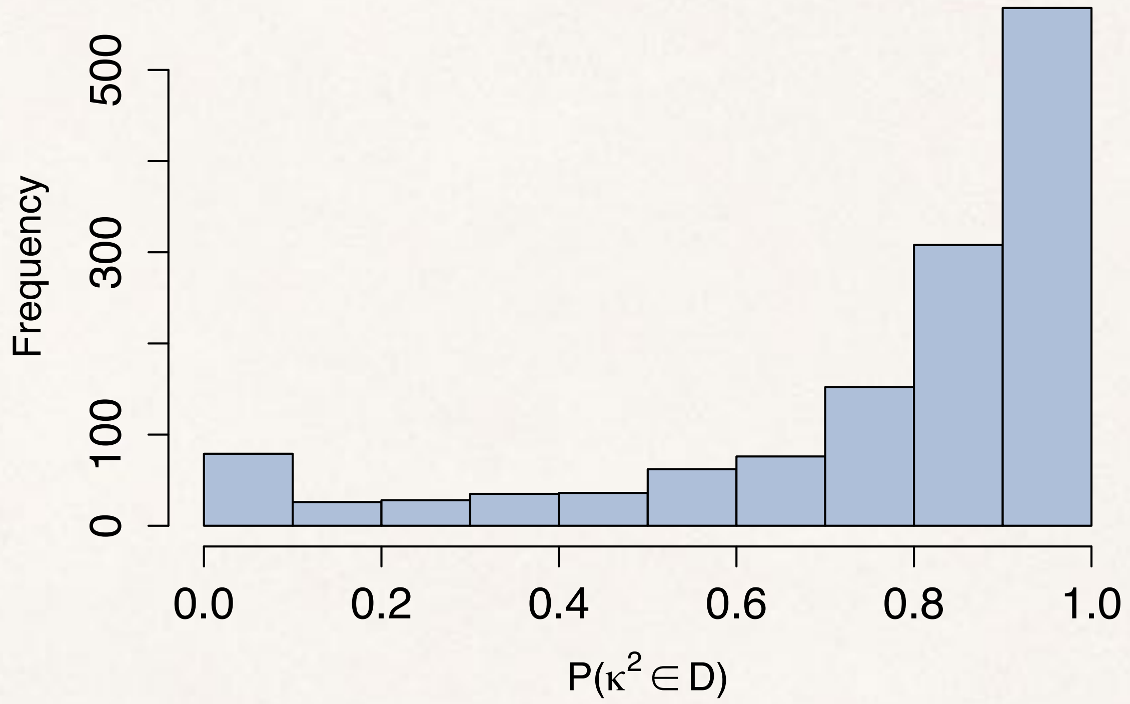
Capturing Albumin Volatility

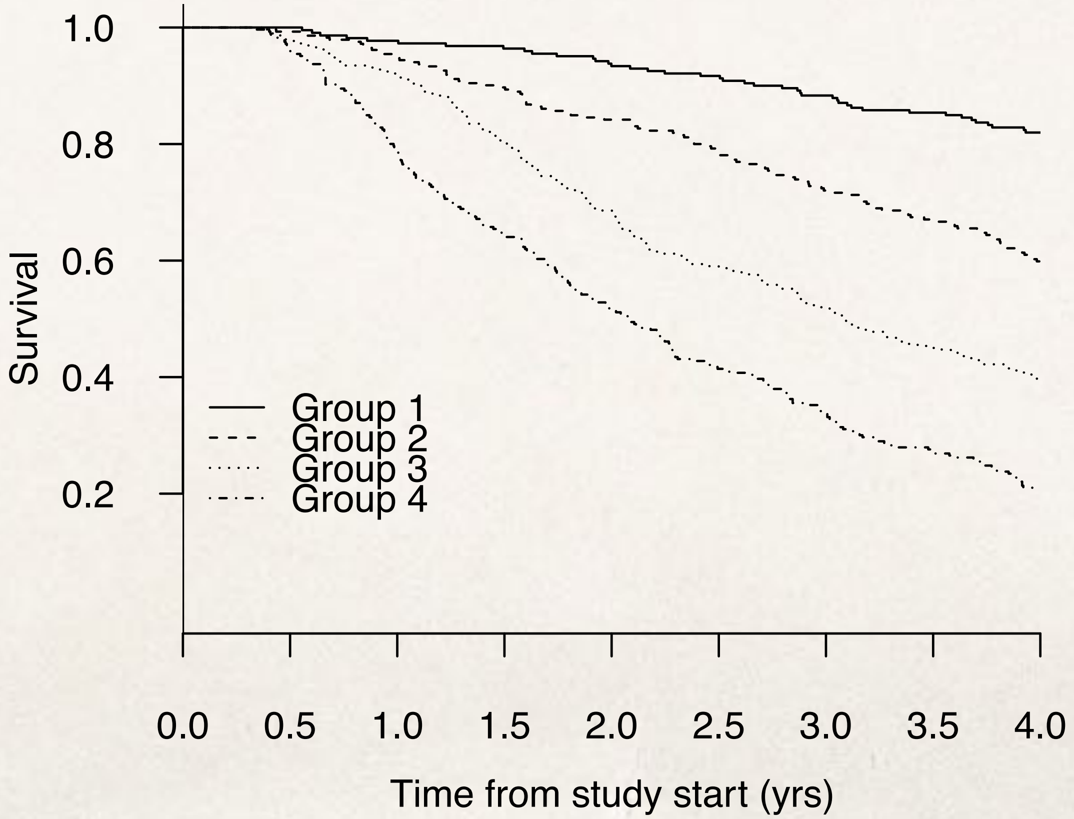
As before, G_{D^c} has a Dirichlet process prior, but this time, the support of its base distribution G_0^* is $[0.1, \infty)$ (shifted Inverse-Gamma)

This separates the support of the spike-and-slab distributions, so there is no overlap

Sampling from the posterior distribution of κ_i^2 can be achieved using "algorithm 8" in Neal (2000).

Capturing Albumin Volatility





Nonlinear Regression and Classification

In Collaboration with Radford Neal

Shahbaba B, Neal RM (2009), Nonlinear models using Dirichlet process mixtures, *Journal of Machine Learning Research*, Volume 10, 1829-1850.

Nonlinear regression models using DPM

- We (Shahbaba and Neal, 2009) introduced a new nonlinear Bayesian model, which non-parametrically estimates the joint distribution of the response variable,*y*, and covariates, *x*, using Dirichlet process mixtures:

$$y_i, x_{i1}, \ldots, x_{ip} \,|\, \theta_i \sim F(\theta_i)$$

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

$$G \sim \mathscr{D}(G_0, \gamma)$$

- Within each component, assume the covariates are independent, and model the dependence between *y* and *x* using a linear model.

Nonlinear regression models using DPM

- When both x and y are continuous, define F as follows:

$$x_l \sim N(\mu_l, \sigma_l^2)$$

$$y \,|\, x, \alpha, \boldsymbol{\beta} \sim N(\alpha + x\boldsymbol{\beta}, \epsilon^2)$$

- In this model, $\theta = \{\mu, \sigma, \epsilon, \alpha, \boldsymbol{\beta}\}$.

Nonlinear classification models using DPM

- Now consider a classification problem with continuous covariates, $x = (x_1, \ldots, x_p)$, and a categorical response variable, y.

- Define F as follows:

$$x_{il} \sim N(\mu_l, \sigma_l^2)$$

$$P(y = j \,|\, x, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{\exp(\alpha_j + x\boldsymbol{\beta}_j)}{\sum_{j'=1}^{J} \exp(\alpha_{j'} + x\boldsymbol{\beta}_{j'})}$$

- In this model, $\theta = \{\mu, \sigma, \boldsymbol{\alpha}, \boldsymbol{\beta}\}$.

Nonlinear classification models using DPM