NONPARAMETRIC BAYESIAN ANALYSIS OF SOME CLUSTERING PROBLEMS

A Dissertation

by

SHUBHANKAR RAY

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

August 2006

Major Subject: Statistics

NONPARAMETRIC BAYESIAN ANALYSIS OF SOME CLUSTERING PROBLEMS

A Dissertation

by

SHUBHANKAR RAY

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Approved by:

Co-Chairs of Committee,	Bani K. Mallick
	Raymond J. Carroll
Committee Members,	Daren B. Cline
	Edward R. Dougherty
Head of Department,	Simon J. Sheather

August 2006

Major Subject: Statistics

ABSTRACT

Nonparametric Bayesian Analysis of Some Clustering Problems. (August 2006)

Shubhankar Ray, B.Tech., Indian Institute of Technlogy, Guwahati;
M.S., Texas A&M University

Co–Chairs of Advisory Committee: Dr. Bani K. Mallick
Dr. Raymond J. Carroll

Nonparametric Bayesian models have been researched extensively in the past 10 years following the work of Escobar and West (1995) on sampling schemes for Dirichlet processes. The infinite mixture representation of the Dirichlet process makes it useful for clustering problems where the number of clusters is unknown. We develop nonparametric Bayesian models for two different clustering problems, namely functional and graphical clustering.

We propose a nonparametric Bayes wavelet model for clustering of functional or longitudinal data. The wavelet modelling is aimed at the resolution of global and local features during clustering. The model also allows the elicitation of prior belief about the regularity of the functions and has the ability to adapt to a wide range of functional regularity. Posterior inference is carried out by Gibbs sampling with conjugate priors for fast computation. We use simulated as well as real datasets to illustrate the suitability of the approach over other alternatives.

The functional clustering model is extended to analyze splice microarray data. New microarray technologies probe consecutive segments along genes to observe alternative splicing (AS) mechanisms that produce multiple proteins from a single gene. Clues regarding the number of splice forms can be obtained by clustering the functional expression profiles from different tissues. The analysis was carried out on the

Rosetta dataset (Johnson *et al.*, 2003) to obtain a splice variant by tissue distribution for all the 10,000 genes. We were able to identify a number of splice forms that appear to be unique to cancer.

We propose a Bayesian model for partitioning graphs depicting dependencies in a collection of objects. After suitable transformations and modelling techniques, the problem of graph cutting can be approached by nonparametric Bayes clustering. We draw motivation from a recent work (Dhillon, 2001) showing the equivalence of kernel k-means clustering and certain graph cutting algorithms. It is shown that loss functions similar to the kernel k-means naturally arise in this model, and the minimization of associated posterior risk comprises an effective graph cutting strategy. We present here results from the analysis of two microarray datasets, namely the melanoma dataset (Bittner *et al.*, 2000) and the sarcoma dataset (Nykter *et al.*, 2006).

To my parents.

ACKNOWLEDGMENTS

I would like to express my gratitude to my advisor, Dr. Bani K. Mallick, for his encouragement and the many helpful discussions which led to this work. I also want to thank the co-chair, Dr. Raymond J. Carroll, my dissertation committee members, Dr. Daren B. Cline and Dr. Edward R. Dougherty, for their support and understanding. I which to express my gratitude to my parents for their love and support.

TABLE OF CONTENTS

CHAPTER Page

LIST OF TABLES

LIST OF FIGURES

FIGURE Page

FIGURE Page

CHAPTER I

BAYESIAN WAVELET METHODS FOR FUNCTIONAL CLUSTERING

A. Introduction

Functional data arise in a wide variety of applications and is often clustered to reveal differences in the sources or to provide a concise picture of the data. For instance, clustered gene expression profiles from microarrays may point to underlying groups of functionally similar genes. Model-based clustering relies largely on finite mixture models to specify the cluster-specific parameters (Banfield and Raftery, 1993; Yeung *et al.*, 2001) assuming that the number of clusters is known in advance. This approach is unreasonable in practice, as it relies on one's ability to determine the correct number of clusters. Medvedovic and Sivaganesan (2002) use the Dirichlet process based infinite mixture model to overcome these deficiencies. Nonetheless, all these approaches use multivariate normal distributions for modeling and disregard the functional form of the data.

This 'functional' approach was pursued only recently in a mixed-effects spline model by James and Sugar (2003) and in the context of yeast cell-cycle data analysis using periodic basis modelling by Wakefield *et al.* (2003). However, shifts in global and local characteristics in the functional data may not be detectable in these frameworks. As an example, the gene expression profiles of yeast cell-cycles may occasionally depart from the usual cyclic behavior and these shifts will be overlooked, in general, by the periodic basis model.

The Bayesian wavelet modelling used in this paper manages to overcome these

The format and style of this dissertation follows that of *Journal of the Royal Statistical Society, Series B.*

limitations as wavelets have nice approximation properties over a large class of functional spaces (Daubechies, 1992), that can accommodate almost all the functional forms observed in real life applications. Indeed, this richness of the wavelet representation provides the backbone for the popular frequentist wavelet shrinkage estimators of Donoho and Johnstone (1994,1995), which are the precursors of the more recent Bayesian wavelet estimation models (Abramovich *et al.*, 1998, Vidakovic, 1998, Clyde and George, 2000, Clyde *et al.*, 1998). Wavelet representations are also sparse and can be helpful in limiting the number of regressors. Dimension reduction is not inherent in other models, for example, this is done in James and Sugar (2003) by attaching an additional dimension reducing step on the spline coefficients. In Bayes wavelet modelling this is effortlessly achieved by a selection mixture prior to isolate a few significant coefficients for the collection of functions.

The nonparametric Bayes clustering model presented here is based on the mixture of Dirichlet Processes (Ferguson, 1973, Antoniak, 1974). The Dirichlet process (DP) provides a rich two-parameter conjugate prior family and much of the posterior inference for a particular parametric conjugate family applies, when the same prior becomes the base measure for a DP prior, used instead. The base prior modelling of the wavelet coefficients can be motivated by traditional hierarchical wavelet models and allows the specification of the smoothness and regularity properties of the functional realisations from the DP. There are several advantages in the context of clustering. The computation is straightforward owing to the Gibbs sampling schemes proposed in the mid-90s (Escobar and West, 1995) and the number of clusters are automatically determined in the process. In addition, the Bayesian methodology provides a direct way to predict any missing observations extending the applicability of model to incomplete datasets.

The paper is organized as follows. In Section B, we overview the parametric

Bayesian models for wavelet shrinkage. This is later extended to the Dirichlet process based nonparametric model in Section C and the posterior inference is detailed in Section D. Some properties of the clustering model are discussed in Section E. Finally, Section F addresses the simulations with a discussion of the model selection and the missing data case.

B. Hierarchical Wavelet Model

Consider a collection of unknown functions $\{f_i\}$, $i \in \{1, \ldots, n\}$ on the unit interval that are observed with white Gaussian noise at m equi-spaced time points as

$$y_{i,k} = f_i(k/m) + \varepsilon_{i,k},\, \varepsilon_{i,k} \sim N(0, \sigma_i^2)$$

where $k \in \{1, \ldots, m\}$ and m is a power of 2. In a gene microarray, for example, the observed curve $y_{i,k}$ is the response profile at the k^{th} time point for the i^{th} gene. In nonparametric estimation, the functions are analysed in the sequence space of coefficients in an orthonormal wavelet basis for $L_2([0,1])$. Restriction of the functions to the unit interval introduces discontinuities at the edges. Boundary artefacts can be avoided by periodised bases when the functions are periodic (Daubechies, 1992), otherwise boundary folding/reflection extensions are used to improve the behavior at the boundaries.

Wavelet representations are sparse for a wide variety of function spaces and their multiresolution nature enables us to combine results from different resolutions and make conclusions for the estimation problem. In particular, the sparseness implies that when the wavelet basis is orthogonal and compactly supported (Daubechies, 1992), the independent and identically distributed (i.i.d.) normal noise affects all the wavelet coefficients equally, while the signal information remains isolated in a few coefficients. In shrinkage estimation, these small coefficients which are mostly noise are discarded to retrieve an effective reconstruction of the function. The expansion of f_i in terms of periodised scaling and wavelet functions (φ, ψ) has the dyadic form

$$f_i(t) \approx \beta_{i00}\varphi_{00}(t) + \textstyle\sum_{j=1}^{J} \sum_{k=0}^{2^{j-1}} \beta_{ijk}\psi_{jk}(t) \tag{1.1}$$

where β_{i00} is the scaling coefficient, β_{ijk}'s are the detail coefficients and $J = \log_2 m$ is the finest level of the wavelet decomposition. Wavelets also provide a direct way to study the functional smoothness in that the wavelet representation usually contains all the information that can tell whether f_i lies in a smoothness space.

1. Wavelet estimation Models

Wavelet shrinkage estimation was popularised by Donoho and Johnstone (1994,1995), who showed that thresholding rules on the empirical detail coefficients provide optimal estimators for a broad range of functional spaces. Bayesian wavelet shrinkage proceeds by eliciting mixture priors over the detail coefficients β_{ijk}, $(j > 0)$ with a degenerate mass at zero (Abramovich *et al.*, 1998; Clyde and George, 2000) for selection,

$$\beta_{ijk} \sim \pi_j N(0, g_j\sigma^2) + (1 - \pi_j)\delta_0. \tag{1.2}$$

The scaling coefficients β_{i00} on the other hand, are usually modeled by a vague prior. The selection probabilities π_j and the scaling parameters g_j allow us to place our prior belief by level, producing a simple estimation strategy that is more adaptive than the classical analogues of hard or soft thresholding.

In linear model notation, if $\mathbf{Y}_i = (y_{i,1}, \ldots, y_{i,m})$ is the vector of m observations from the i^{th} unit, the regression model is

$$\mathbf{Y}_i = \mathbf{X}\boldsymbol{\beta}_i + \boldsymbol{\varepsilon}_i,\ i = 1, \ldots, n \tag{1.3}$$

where $\boldsymbol{\beta}_i = (\beta_{i00}, \beta_{i10}, \beta_{i20}, \beta_{i21,} \ldots)^t$ are the wavelet coefficients for f_i after the discrete wavelet transformation $\mathbf{X}$ and $\boldsymbol{\varepsilon}_i \sim N(0, \sigma_i^2 \mathbf{I}_m)$. The selection priors (1.2) are conveniently incorporated as a scale-mixture with latent indicator variables γ_{jk} that equal 1 with probability π_j (Clyde & George, 2000; Clyde *et al.*, 1998; De Canditiis & Vidakovic, 2004). The effective joint prior for the coefficients and the model variance is

$$\boldsymbol{\beta}_i, \sigma_i^2 | \mathbf{V} \sim NIG(\mathbf{0}, \mathbf{V}; u, v)$$

where NIG denotes the normal-inverse gamma prior – the product of the conditionals $\boldsymbol{\beta}_i | \sigma_i^2, \mathbf{V} \sim N(0, \sigma_i^2 \mathbf{V})$ and $\sigma_i^2 \sim IG(u, v)$ with u, v as the usual hyperparameters for the inverse gamma (IG) prior.

The diagonal matrix $\mathbf{V}$ can be used to obtain a scale mixture prior, we let $\mathbf{V} = \text{diag}(\boldsymbol{\gamma})\text{diag}(\mathbf{g})$ where $\boldsymbol{\gamma} = (\gamma_{00}, \gamma_{10}, \gamma_{20}, \gamma_{21}, \ldots)$ is a vector of latent indicator variables for selection of each coefficient and $\mathbf{g} = (g_0, g_1, g_2, g_2, \ldots)$ comprise the corresponding scaling parameters given by

$$\gamma_{j,k} \sim Bernoulli(\pi_j) \text{ and } g_j \sim IG(r_j, s_j)$$

where (r_j, s_j) are hyperparameters specified levelwise. This hierarchical layer is especially useful for modelling sparse wavelet representations, which otherwise requires Laplace-like sharp non-conjugate priors (Vidakovic, 1998). In particular, there is the flexibility of controlling our prior belief about the scaling coefficient β_{i00} by letting $\pi_0 = 1$ and tuning (r_0, s_0) to vary $\text{var}(g_0)$.

To summarise the hierarchical wavelet model, we have

$$\begin{aligned} \mathbf{Y}_i | \boldsymbol{\beta}_i, \sigma_i^2 &\sim N(\mathbf{X}\boldsymbol{\beta}_i, \sigma_i^2 \mathbf{I}_m), \\ \boldsymbol{\beta}_i, \sigma_i^2 | \boldsymbol{\gamma}, \mathbf{g} &\sim NIG(0, \mathbf{V}; u, v), \end{aligned}$$

$$
\begin{aligned}
g_j &\sim IG(r_j, s_j),\\
\gamma_{j,k} &\sim Bernoulli(\pi_j),
\end{aligned}
$$

for $i \in \{1, \ldots, n\}$, $j \in \{0, \ldots, J\}$ and $k \in \{0, \ldots, 2^{j-1}\}$.

C. Wavelet Clustering Model

In the clustering model, the parameters $\boldsymbol{\theta}_i = (\boldsymbol{\beta}_i, \sigma_i^2)$ for the underlying functions f_i are elicited by Dirichlet process priors (DP). Dirichlet processes are almost surely discrete and comprise a certain partitioning of the parameter space needed for clustering. More precisely, the sequence of parameters $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_n$ come from a random distribution F that is a realisation from a Dirichlet process $D(\alpha, H_{\boldsymbol{\phi}})$ depending on a precision parameter α and base prior $H_{\boldsymbol{\phi}} = \mathrm{E}F$ with $\boldsymbol{\phi}$ as the parameters of H. The nonparametric hierarchical model is completed by mixing the base prior for the DP with the hyperpriors of Section B.1 and is described as

$$
\begin{aligned}
\mathbf{Y}_i|\boldsymbol{\beta}_i, \sigma_i^2 &\sim N(\mathbf{X}\boldsymbol{\beta}_i, \sigma_i^2\mathbf{I}_m), && (1.4)\\
\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_n &\sim F,\\
F &\sim D(\alpha, NIG(0, \mathbf{V}; u, v)), && (1.5)\\
g_j &\sim IG(r_j, s_j), && (1.6)\\
\gamma_{j,k} &\sim Bernoulli(\pi_j),\\
\alpha &\sim G(d_0, \eta_0).
\end{aligned}
$$

Here $\boldsymbol{\phi} = \{\mathbf{g}, \boldsymbol{\gamma}\}$.

The underlying clustering properties of the DP are easier to appreciate in its Pölya urn representation (Blackwell and MacQueen, 1973). This connection is also used later to perform sequential flexible Gibbs sampling of the clustering parameters $\boldsymbol{\theta}_i$ as in Escobar and West (1995). In a sequence of draws $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots$ from the Pölya urn the n^{th} sample is either distinct with a small probability $\alpha/(\alpha+n-1)$ or is tied to a previous sample with positive probability to form a cluster. Let $\boldsymbol{\theta}_{-n} = \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n\} \backslash \boldsymbol{\theta}_n$ and d_{n-1} = number of preexisting clusters of tied samples in $\boldsymbol{\theta}_{-n}$ at the n^{th} draw, then we have

$$\boldsymbol{\theta}_n | \boldsymbol{\theta}_{-n}, \alpha, \boldsymbol{\phi} = \frac{\alpha}{\alpha+n-1} H_{\boldsymbol{\phi}} + \textstyle\sum_{i=1}^{d_{n-1}} \frac{n_i}{\alpha+n-1} \delta_{\bar{\boldsymbol{\theta}}_i} \tag{1.7}$$

where $H_{\boldsymbol{\phi}} = NIG(\mathbf{0}, \mathbf{V}; u, v)$ is the base prior, $\boldsymbol{\phi} = \{\mathbf{g}, \boldsymbol{\gamma}\}$ and the i^{th} cluster has n_i tied samples that are commonly expressed by $\bar{\boldsymbol{\theta}}_i = (\bar{\boldsymbol{\beta}}_i, \bar{\sigma}_i^2)$ and $\sum_{i=1}^{d_{n-1}} n_i = n-1$. In the long term of sequential draws, the number of clusters d_n is much less than n and is determined by the precision α. We also use the set $\mathcal{C}_n$ containing the clustering profile at the n^{th} draw, such that $\mathcal{C}_n(i)$ is the set of indices of all the curves in the i^{th} cluster.

The modelling of the base prior $H_{\boldsymbol{\phi}}$ is important and is reflected in all the realisations from the DP which are centered at $H_{\boldsymbol{\phi}} = NIG(0, \mathbf{V}; u, v)$. As before, the detail coefficients are conveniently modelled by a selection prior. The scaling coefficients are modelled using priors for which the hyperparameters are empirically estimated and in Section F, we show that this in general works better than vague priors.

We consider two models that differ in the way the error terms are modelled. Model 1 is a heteroscedastic model with $\boldsymbol{\theta}_i = (\boldsymbol{\beta}_i, \sigma_i^2)$ and can be useful to handle potential fluctuations in variability across clusters or to check the normality assumptions in (1.3). Model 2 is an oversmoothed homoscedastic model with $\sigma_i^2 = \sigma^2$ for all i, which is computationally more straightforward. The fact that the d_n separate draws of variance in Model 1 are replaced by a single draw, imparts more stability to the Markov chain. This makes it suitable for cases where the population is not overly heteroscedastic.

The nonparametric model does not allow direct control over the number of clusters, but it offers the parameter α that determines the expected and the asymptotic number of clusters: $d_n/\log n \rightarrow \alpha$ almost surely (Korwar and Hollander, 1973). Minor subjective reservations aside, we think this vagueness does not affect the inference in practice.

1. Choice of hyperparameters

The specification of (g_j, π_j) represents our prior belief about the collection of curves at each level. A variety of scale mixture or shrinkage priors (Vidakovic, 1998; Clyde and George, 2000) have been proposed for robust and heavy-tailed modelling of wavelet coefficients. All these specifications comprise different ways of modelling the decay of wavelet coefficients relating them to functional smoothness. In particular, Abramovich *et al.* (1998) show that these parameters can be specified such that the functions fall in Besov spaces – a valuable backbone for modelling a broad range of smoothness and spatial adaptation properties. For the Besov space $\mathcal{B}^{\ell}_{p,q}$ $(\ell > 0, 1 < p, q \leq \infty)$, ℓ gives the order of smoothness in $L_p([0,1])$, p controls the spatial inhomogeneity and the parameter q allows fine distinctions in the smoothness of fixed order ℓ.

In general, for a function with an almost surely finite wavelet representation the smoothness is the same as that of the mother wavelet ψ. Suppose $\psi \in \mathcal{B}^{\ell}_{p,q}$ $(1 \leq p, q \leq \infty)$ has ζ vanishing moments satisfying $\max(0, 1/p - 1/2) < \ell < \zeta$. For $j > 0$, fixing $g_j = c_1 2^{-aj}$ and $\pi_j = \min(1, c_2 2^{-bj})$ with $c_1, c_2, a > 0$ and $b > 1$ gives an a.s. finite wavelet series and f_i also belongs to $\mathcal{B}^{\ell}_{p,q}$. Abramovich *et al.* (1998) extend the equivalence for $b \in [0, 1]$, if (a, b) are chosen to satisfy

$$(b-1)/p + (a-1)/2 \geq \ell - 1/p \tag{1.8}$$

with equality when $q = \infty$ and $1 \leq p < \infty$. For convenience, we shall refer to (a, b) as the Besov parameters.

For any (a, b) the realisations lie in a family of Besov spaces described by the relation (1.8). By fixing a, we see a direct relation between b and the positive range of p. In effect b controls the sparseness and the spatial inhomogeneity of the functional realisations. In a similar manner, a defines the positive range of ℓ and therefore, controls the overall decay of the wavelet coefficients and the smoothness. This analogy can be drawn directly by looking at the decay of the wavelet coefficients realised by the prior distributions. Loosely, by increasing a or b we realise functions with higher *effective* smoothness, since in (1.8), $\ell > 1/p$ implies functions in B^{ℓ}_{pq} are continuous. Apart from the interesting connection with smoothness, in practice, greater flexibility is achieved by updating the probabilities π_j. These parameters can also affect the model's ability to identify clusters. For instance, smoother functions on average cluster more tightly and should result in a drop in the number of clusters.

Recall that in our hierarchical model, the scaling parameters g_j were mixed by a conjugate prior (1.6) and in general, this imparts greater flexibility than subjective deterministic specifications. It also allows modelling within the Besov spaces through restrictions imposed on the IG hyperparameters (r_j, s_j). For $j > 0$ and a fixed integer $p \geq 1$, if the hyperparameters satisfy

$$\mathrm{E}(g_j^{p/2})^{1/p} = \frac{{r_j}^{1/2}}{\{(s_j-2)(s_j-4)\ldots(s_j-p)\}^{1/p}} = c_1 2^{-aj} \text{ and } \pi_j = \min(1, c_2 2^{-bj})$$

with (a, b) satisfying (1.2), then the Besov correspondence still holds. The proof (see Appendix B) is similar to that of Abramovich *et al.* (1998), except the use of the marginal t-distribution after averaging out g_j. A simple way to choose the IG hyperparameters is to fix $s_j = p + 1$ and calculate r_j to satisfy the foregoing relation for given values of c_1 and a.

Abramovich *et al.* (1998) use fixed values of (a, b) based on the prior knowledge about the functions regularity, followed by method of moments estimators to calculate the constants c_1, c_2. We maximize the marginal likelihood (Clyde and George, 2000) with respect to the base prior $H_{\boldsymbol{\phi}}$ to estimate the constants while fixing a and b,

$$L(c_1, c_2) \propto -\log|\mathbf{V}_n^*| - (v + mn)\log\left\{u + \sum_{i=1}^{n} \mathbf{Y}_i'\mathbf{Y}_i - \boldsymbol{\mu}_n^{*\prime}(\mathbf{V}_n^*)^{-1}\boldsymbol{\mu}_n^*\right\}$$

where $\boldsymbol{\mu}_n^* = \mathbf{V}_n^* \sum_{i=1}^{n} \mathbf{X}'\mathbf{Y}_i$ and $\mathbf{V}_n^* = (\mathbf{V}^{-1} + n\mathbf{I}_m)^{-1}$. For fixed values of a and b, a gridded maximization procedure can lead to estimates of c_1, c_2. Moreover, an objective selection of the hyperparameters (a, b) can be performed by extending the maximization procedure to a grid in (a, b), as illustrated in Section F.5. Ideally, the marginal likelihood can be calculated for the DP priors using the collapsed sequential importance sampling methods of Basu and Chib (2003). This procedure can be computationally intensive depending on the size of the dataset. In general, for small datasets, it leads to estimates that are comparable to the marginal maximum likelihood estimates using $H_{\boldsymbol{\phi}}$.

D. Posterior Inference

Adopting base priors that are conjugate to the likelihood expedites the posterior sampling of the clustering parameters $\boldsymbol{\theta}_i$ from the Pölya urn. We also retain the computational advantage of the scale mixture form of the base prior and the conditional posteriors for all the indicator variables γ_{jk} and the scale parameters g_j are in standard form.

The conditional posterior distributions for the heteroscedastic case are derived here. The corresponding conditionals for the special homoscedastic case easily follow from these derivations.

1. Conditional distributions for clustering

To update the clustering parameters $\boldsymbol{\theta}_n = (\boldsymbol{\beta}, \sigma^2)_n$, we combine the likelihood (1.4) with the Pölya urn prior (1.7),

$$(\boldsymbol{\beta}, \sigma^2)_n | (\boldsymbol{\beta}, \sigma^2)_{-n}, \alpha, \boldsymbol{\phi}, \mathbf{Y}_n \propto q_{n0} H^*_{\boldsymbol{\phi},n}(\beta, \sigma^2) + \textstyle\sum_{i=1}^{d_{n-1}} q_{ni} \delta_{(\bar{\boldsymbol{\beta}}, \bar{\sigma}^2)_i} \tag{1.9}$$

where $H^*_{\boldsymbol{\phi},n} = NIG(\boldsymbol{\mu}^*, \mathbf{V}^*; u^*_n, v^*_n)$ is $H_{\boldsymbol{\phi}}$–a posteriori with

$$\begin{aligned} \mathbf{V}^* &= (\mathbf{V}^{-1} + \mathbf{I}_m)^{-1},\ \boldsymbol{\mu}^* = \mathbf{V}^* \mathbf{X}' \mathbf{Y}_n, \\ u^*_n &= u + \mathbf{Y}'_n \mathbf{Y}_n - \boldsymbol{\mu}^{*\prime} (\mathbf{V}^*)^{-1} \boldsymbol{\mu}^* \text{ and } v^*_n = v + m. \end{aligned}$$

The weights $q_{n\cdot}$ follow from the likelihood and its marginals in the NIG conjugate family (Escobar and West, 1995) and determine the posterior inclination towards new distinct samples. We have $q_{ni} \propto n_i \phi(\mathbf{Y}_n | \mathbf{X} \bar{\boldsymbol{\beta}}_i, \bar{\sigma}^2_i \mathbf{I}_m)$ as the conditional distribution of $\mathbf{Y}_n$ given the i^{th} cluster or distinct sample $(\bar{\boldsymbol{\beta}}, \bar{\sigma}^2)_i$ and $q_{n0} \propto \alpha t_m\{\mathbf{0}, u(\mathbf{I}_m + \mathbf{X}\mathbf{V}\mathbf{X}')\}$ is the marginal distribution.

Similarly, setting $H_{\boldsymbol{\phi}} = N(\mathbf{0}, \sigma^2 \mathbf{V})$ for the homoscedastic model gives

$$\begin{aligned} H^*_{\boldsymbol{\phi},n} &= N(\boldsymbol{\mu}^*, \sigma^2 \mathbf{V}^*) \text{ with} \\ q_{n0} &\propto \alpha \phi\{\mathbf{0}, \sigma^2(\mathbf{I}_m + \mathbf{X}\mathbf{V}\mathbf{X}')\} \text{ and } q_{ni} \propto n_i \phi(\mathbf{Y}_n | \mathbf{X}\bar{\boldsymbol{\beta}}_i, \sigma^2 \mathbf{I}_m). \end{aligned}$$

The posterior distribution of σ^2 is simply $(\sigma^2 | \mathbf{Y}, \boldsymbol{\phi}) \sim IG\,(u^*, v^*)$ where

$$\begin{aligned} u^* &= u + \sum_{i=1}^{d_n} \left(\sum_{j \in \mathcal{C}_n(i)} \mathbf{Y}'_j \mathbf{Y}_j - \boldsymbol{\mu}^{*\prime}_i (\mathbf{V}^*_i)^{-1} \boldsymbol{\mu}^*_i \right), \\ v^* &= v + mn, \end{aligned}$$

$$
\begin{aligned}
\boldsymbol{\mu}_i^* &= \mathbf{V}_i^* \textstyle\sum_{j \in \mathcal{C}_n(i)} \mathbf{X}'\mathbf{Y}_j, \text{ and,} \\
\mathbf{V}_i^* &= (\mathbf{V}^{-1} + n_i \mathbf{I}_m)^{-1}.
\end{aligned}
$$

Sampling requires computation of the mixture probabilities q_{ni} for the distinct preexisting parameter values, which are small compared to n. The sequential update of model parameters from the Pölya urn model can be randomized to preclude any ordering related bias. This is followed by a resampling step (Bush and MacEachern, 1996) that expedites model mixing by literally shaking up the converged mixture model.

2. Posterior sampling of the mixing parameters

For notational convenience, the parameters are grouped by the dyadic levels j of the wavelet decomposition – $\boldsymbol{\gamma}_j = \{\gamma_{jk} : \forall k\}$ and $\bar{\boldsymbol{\beta}}_{ij} = \{\bar{\beta}_{ijk} : \forall k\}$ for $j \geq 0$. To obtain the conditional posteriors for the scaling parameters g_j and the indicator variables γ_{jk}, we exploit the conditional independence of the distinct cluster parameters $\{(\bar{\boldsymbol{\beta}}_i, \bar{\sigma}_i^2)\}_{i=1}^{d_n}$ given the clusters $\mathcal{C}_n$. Korwar and Hollander (1973) show that conditional on $\mathcal{C}_n$ the distinct parameters are i.i.d. $H_{\boldsymbol{\phi}}$.

For the heteroscedastic case, the scaling parameters g_j are updated levelwise by combining the base prior and (4),

$$
g_j | \boldsymbol{\gamma}_j, \mathcal{C}_n, \{\bar{\sigma}_i^2, \bar{\boldsymbol{\beta}}_{ij}\}_{i=1}^{d_n} \sim IG(r_j^*, s_j^*),
$$

where $s_j^* = d_n \sum_{k=0}^{2^{j-1}} \gamma_{jk} + s_j$, $r_j^* = \sum_{i=1}^{d_n} \bar{\sigma}_i^{-2} \sum_{k=0}^{2^{j-1}} \gamma_{jk}^2 \bar{\beta}_{ijk}^2 + r_j$. For Model 2, the $\bar{\sigma}_i^2$'s are replaced by a single σ^2. Observe that these updates combine the information at the j^{th} resolution across all the distinct functions and the average shrinkage is conservative and depends on the total variation at any resolution.

Similarly, for each level, the indicators $\boldsymbol{\gamma}_j$ are updated conditional on the indicators at other levels $\boldsymbol{\gamma}_{-j}$

$$f(\boldsymbol{\gamma}_j|\boldsymbol{\gamma}_{-j}, \mathbf{g}, \mathcal{C}_n, \mathbf{Y}) \propto \textstyle\prod_{i=1}^{d_n} f(\{\mathbf{Y}_j\}_{j\in\mathcal{C}_n(i)}|\boldsymbol{\gamma}, \mathbf{g}, \mathcal{C}_n)\pi_j$$

where

$$f(\{\mathbf{Y}_j\}_{j\in\mathcal{C}_n(i)}|\boldsymbol{\gamma}, \mathbf{g}, \mathcal{C}_n) \propto C_i \frac{|\mathbf{V}_i^*|^{1/2}}{|\mathbf{V}|^{1/2}} \left\{u + \sum_{j\in\mathcal{C}_n(i)} \mathbf{Y}_j'\mathbf{Y}_j - \boldsymbol{\mu}_i^{*\prime}(\mathbf{V}_i^*)^{-1}\boldsymbol{\mu}_i^*\right\}^{-(v+mn_i)/2} \tag{1.10}$$

is the marginal likelihood for the i^{th} cluster with $C_i = \Gamma((v+mn_i)/2)/\pi^{n_i m/2}$ and $\mathbf{Y}$ depicts the collection of all responses $\{\mathbf{Y}_i\}_{i=1}^n$. The update is similar in form for Model 2. Again, the selection of γ_{jk} is more conservative and is decided by the proportion of variation explained by the coefficients $\bar{\beta}_{ijk}$ at location (j,k) for all i.

3. Posterior sampling of the precision parameter

We update the precision α as in Escobar & West (1995). The posterior distribution is derived by combining a gamma prior for α with the distribution of d_n:

$$f(d_n|\alpha, n) = c_n(d_n)\frac{\Gamma(\alpha)}{\Gamma(n+\alpha)}\alpha^{d_n} n! \tag{1.11}$$

that resembles the Cauchy formula for counting permutations and has been calculated independently by several authors including Antoniak (1974). Combining (1.11) with $f(\alpha)$, the prior for α, it can be shown that posterior $f(\alpha|d_n, n) \propto f(\alpha)\alpha^{d_n-1}(\alpha + n)\int_0^1 x^\alpha(1-x)^{n-1}dx$. Equivalently, there is a beta random variable η such that $f(\alpha|d_n, n) = \int f(\alpha, \eta|d_n, n)d\eta$. Thus α can be updated in two steps. First, conditional on α and d_n, update η. Second, conditional on the last sampled value of η and d_n, draw α. When $\alpha \sim G(d_0, \eta_0)$, both conditional distributions are in standard form, are given by

$$(\alpha|\eta, d_n) \;\sim\; \rho_n G(d_0 + d_n, \eta_0 - \log\eta) + (1-\rho_n)G(d_0 + d_n - 1, \eta_0 - \log\eta)$$

$$(\eta|\alpha, d_n) \quad \sim \quad Beta(\alpha + 1, n)$$

where $\rho_n/(1-\rho_n) = (d_0 + d_n - 1)/(n(\eta_0 - \log \eta))$.

E. Properties of the Clustering Model

When sequentially sampling from the Pölya urn (1.9) the mixture probabilities $q_{n\cdot}$ are based on the ℓ_2-distance between the n^{th} observed function and the d_{n-1} previously sampled functions. Most classical clustering algorithms such as k-means, neural network clustering and so on, or even statistical models with normal likelihoods (and priors) inherently use the ℓ_2-distance as a measure of distance between two curves. Likewise, in James and Sugar (2003) the decision of choosing a cluster is based on the squared distance between spline coefficients.

In this section, we ascertain if in the long term of sequential draws, there is a minimum squared-distance that would ensure a distinct sample. This distance may be viewed as the eventual minimum separation between clusters and is referred to as the *sampling resolution.* As n gets larger the collection of sampled functions becomes populated and we expect the resolution to grow, i.e. in order for a new sample to be distinct it must distinguish itself more clearly from increasing population as n gets large. The rate of this growth gives an idea about the adaptation of the clustering model. The following theorem proved in Appendix C characterizes the resolution of the Dirichlet process without any mixing, i.e. while fixing the parameters $\mathbf{g}, \boldsymbol{\gamma}$ and α.

Theorem 1 *For the homoscedastic model with the base prior $H_{\boldsymbol{\phi}}$ such that $||\boldsymbol{\beta}_i||_2 < \infty$ almost surely $\forall i = 1, 2, \ldots$, the posterior sampling resolution is $\sigma^2 O(\log(\log n)^{(1+\delta)})$, for any $\delta > 0$.*

Remarks. 1. The slow rate of increase suggests good adaptation properties of the model that does not change drastically as n becomes large.

2. Note that the condition of bounded ℓ_2 norm can be achieved by choosing hyperparameters from (1.2).

3. The error variance directly affects the separation between clusters in that a higher variability means the clusters are more spread out.

F. Examples

We analyse one synthetic and two real datasets to illustrate the practical potential of the functional clustering model. The virtues of wavelet modelling are emphasised by using datasets that exhibit different degrees of smoothness and spatial inhomogeneity. The Besov class of priors mentioned in Section C.1 easily accommodate the three extremes listed below in that b controls the inhomogeneity and the overall smoothness can be attributed to a.

1. **Smooth but inhomogeneous.** The synthesized Doppler signals although intrinsically continuous, the finite sampling along with the changing intensity of oscillations makes it spatially inhomogeneous.

2. **Not smooth and inhomogeneous.** Yeast cell cycle gene expression profiles in the second example, may be far from continuous and are characterized by varying degrees of temporal fluctuation.

3. **Not smooth but homogeneous.** The third example analyses a meteorological precipitation dataset that exhibits a certain homogeneity in the prevailing bumpiness.

For performance evaluation the number of clusters and the misclassification rate (in supervised conditions) are reported. In addition, the robustness to missing observations is checked for synthetic datasets with different amounts of missing points and a yeast cell cycle microarray data. All the reported results are averaged over 100 simulations with 10,000 iterations per simulation and a burn-in period of 1000. In general, these specifications must vary depending on n, m and the prior estimate of the model variance. The MCMC mixes fairly well in all these examples and the chains seem to converge much before the alloted burn-in time. Excepting the microarray dataset analysed below, a near-symmetric Symmlet wavelet basis with 8 vanishing moments was used in all the experiments.

1. Mixed effects spline model

In some cases, the results are compared with the mixed-effects spline model of James and Sugar (JS) (2003) given by

$$\mathbf{Y}_i = \mathbf{S}\boldsymbol{\beta}_{1i} + \mathbf{S}\boldsymbol{\beta}_{2i} + \boldsymbol{\varepsilon}_i,\ \boldsymbol{\varepsilon}_i \sim N(0, \sigma_i^2 \mathbf{I}_m) \tag{1.12}$$

where $\mathbf{S}(m \times p)$ is a natural cubic spline design with suitably chosen knots, $\boldsymbol{\beta}_{1i}$ are cluster specific coefficients (i.e., all responses in a cluster have the same $\boldsymbol{\beta}_{1i}$) and coefficients $\boldsymbol{\beta}_{2i}$ account for individual variations in the functions within each cluster. We do not consider the additional dimension reducing transformation used by JS, but instead compensate with a smaller number of knots to fit higher order splines. Since the original EM implementation of this model was unavailable, Bayesian modelling was used to expedite the comparison and the following conjugate priors were used

$$\begin{aligned} \boldsymbol{\beta}_{1i} &\sim P,\ P \sim DP(\alpha, N(0, \sigma_i^2 \boldsymbol{\Gamma}_1)),\ \boldsymbol{\Gamma}_1 \sim IW(\mathbf{R}_1, s_1) \text{ and} \\ \boldsymbol{\beta}_{2i} &\sim N(0, \sigma_i^2 \boldsymbol{\Gamma}_2),\ \boldsymbol{\Gamma}_2 \sim IW(\mathbf{R}_2, s_2) \end{aligned}$$

The posterior conditionals for $\boldsymbol{\beta}_{1i}$ and $\boldsymbol{\Gamma}_1$ follow closely from the derivations in Section D and inverse Wishart posteriors take the place of the IG posteriors. Conditional on the clusters – the distinct β_{1i}'s and $\mathbf{\Gamma}_1$, there are n additional conjugate normal draws of $\boldsymbol{\beta}_{2i}$ followed by another inverse Wishart draw of $\mathbf{\Gamma}_2$. In the simulations, $\mathbf{R}_1 = 10.0\mathbf{I}$ and $s_1 = 5$ to get a diffuse prior for $\mathbf{\Gamma}_1$. Averaging out $\boldsymbol{\beta}_{1i}$ with respect to the base prior, and $\boldsymbol{\beta}_{2i}$ with respect to its normal prior, the hyperparameters $\mathbf{R}_2$ and s_2 are estimated by Empirical Bayes.

2. Priors for scaling coefficients

Prior modelling of the scaling coefficients determines the sensitivity to the locational differences in the dataset. We compare the traditional choice of vague/diffuse priors with a prior for which the hyperparameters have been empirically estimated.

The prior scale parameter g_0 for the scaling coefficients β_{i00} follows an inverse gamma distribution with parameters (r_0, s_0). For diffuse priors these parameters are adjusted for large prior variance. An approximate empirical estimation of these hyperparameters can be carried out by marginalising the likelihood with respect to the base prior (of the Dirichlet process) individually for each curve and then applying moment matching on the population of estimated g_0's to estimate (r_0, s_0). The marginal likelihood follows by combining the model likelihood (1.4) $\tilde{\beta}_{ijk} \sim N(\beta_{ijk}, \sigma^2)$, written in terms of the empirical wavelet coefficients $\tilde{\beta}_{ijk}$, with the prior $\beta_{ijk} \sim N(0, \gamma_{jk} g_j \sigma^2)$. For the scaling coefficients, we have $\tilde{\beta}_{i00} \sim N(0, (1+g_0)\sigma^2)$ and g_0 for the i^{th} curve is simply estimated as $\tilde{\beta}^2_{i00}/\sigma^2 - 1$, where σ^2 can be estimated from the finer levels of wavelet decomposition. The median and the one-third range of the population of n such estimates is matched with

$$\begin{aligned} \mathrm{E}g_0 &= r_0/(s_0-2), \text{ and,} \\ \mathrm{var}(g_0) &= 2r_0^2/\{(s_0-2)^2(s_0-4)\}, \end{aligned}$$

respectively, to generate rough estimates of (r_0, s_0).

From a comparative study of the two priors applied to various example datasets, the empirically estimated prior seem to be a more reliable choice and lead to better cluster estimates than the diffuse prior. In all the examples furnished below, empirically estimated priors were used for the scaling coefficients. For illustrative purposes, a comparative study is also discussed in one of the examples.

3. Model choice

The state space of possible clustering combinations can be very large and some model selection criterion is required to decide between the best models. Recently, Quintana and Iglesias (2003) provide a search algorithm to approach the best model by minimizing a penalized risk, however, for large datasets the computational constraints can be prohibitive. Traditional approaches, such as using model marginal likelihoods for model comparison seem to be more practicable considering the large datasets commonly encountered in clustering problems.

The marginal likelihoods conditional on the specific clustering configurations follows directly from the calculations in Section D.2. For a fixed cluster configuration $\mathcal{C}$, simple Monte-Carlo averaging of the marginal distributions (1.10), gives

$$f(\mathbf{Y}|\mathcal{C}) \approx \frac{1}{N} \sum_{k=1}^{N} \prod_i f(\{\mathbf{Y}_j\}_{j \in \mathcal{C}(i)} | \boldsymbol{\gamma}^{(k)}, \mathbf{g}^{(k)}, \mathcal{C}), \tag{1.13}$$

where N is the total number of MCMC samples. Here a large state space of the indicators $\boldsymbol{\gamma} \in \{0,1\}^m$ would ideally require a very large number of MCMC samples. In general, it is observed that the Markov chain of the indicator variables mixes well like most hierarchical wavelet models, with little change in the convergent states across simulations. This is essentially due to the sharp localisation of features on the wavelet scale. In such cases, a reasonable estimate of the marginal likelihood (1.13) is achieved by averaging over the Markov chain for a previously estimated $\boldsymbol{\gamma}$.

4. Missing data interpolation

It is common to encounter missing values in clustering and to supplement the model with a Bayesian imputation step is useful. There has been some work for wavelet methods on unequispaced grids (Kovac and Silverman (2000); Pensky & Vidakovic, 2001), however, we limit ourselves to the case were points are missing from a fixed equispaced grid.

Missing data imputation is expedited by Gibbs sampling under the *a priori* independence of the d_n distinct parameters. Gibbs sampling starts with a random clustering configuration and randomly imputed missing points. Conditional on the imputed dataset, the Pölya urn samples the cluster parameters $\boldsymbol{\theta}_i$. After n such steps of sequential sampling the initial clustering configuration $\mathcal{C}_n$ is completely updated. Next, conditional on $\mathcal{C}_n$ the posterior predictive distribution is used to perform the imputation. For instance, for an incomplete response (say the j^{th} response) that was assigned to the i^{th} cluster in the preceding step of the Gibbs sampling, we use $\mathbf{Y}_j|\bar{\boldsymbol{\beta}}_i, \bar{\sigma}_i^2 \sim N(\mathbf{X}\bar{\boldsymbol{\beta}}_i, \bar{\sigma}_i^2\mathbf{I}_m)$ and $\bar{\boldsymbol{\beta}}_i \overset{iid}{\sim} N(0, \bar{\sigma}_i^2\mathbf{V})$, to write the marginal $\mathbf{Y}_j|\bar{\sigma}_i^2, \mathbf{g}, \boldsymbol{\gamma} \sim N(\mathbf{0}, \bar{\sigma}_i^2(\mathbf{I}_m + \mathbf{XVX}'))$. Let $\mathbf{Y}_{j1}$ and $\mathbf{Y}_{j2}$ be the known and missing parts of $\mathbf{Y}_j$ respectively, then from standard normal distribution theory the posterior predictive distribution is given by

$$(\mathbf{Y}_{j2}|\mathbf{Y}_{j1}, \bar{\sigma}_i^2, \mathbf{g}, \boldsymbol{\gamma}, \mathcal{C}_n) \sim N(\Lambda_{21}\Lambda_{11}^{-1}\mathbf{Y}_{j1}, \bar{\sigma}_i^2(\Lambda_{22} - \Lambda_{21}\Lambda_{11}^{-1}\Lambda_{12})),$$

where Λ_{ij} are obtained by partitioning the marginal covariance as

$$\mathbf{I}_m + \mathbf{XVX}' = \begin{bmatrix} \Lambda_{11} & \Lambda_{12} \\ \Lambda_{21} & \Lambda_{22} \end{bmatrix}.$$

We can use this technique to accurately predict unobserved portions of any curve with uncertainty intervals. The effectiveness of our method is shown in one of the examples.

5. Shifted Doppler signals

This dataset is motivated by similar examples in Donoho and Johnstone (1994, 1995) and consists of 200 shifted Doppler signals with the common form

$$f_{t_0}(t) = -0.025 + 0.6\sqrt{t(1-t)}\,\sin(2.10\pi/(t-t_0))$$

and the phase t_0 is continuously varied in eight disjoint intervals equally interspersed in $[0,1]$ to generate the 200 signals. In one of these intervals, three functions were perturbed to assess the model sensitivity to small local fluctuations. This created a total of nine distinct classes of functions that have been equi-sampled on a common grid of 128 points. For simulation, noisy data were generated by adding independent normal noise $\varepsilon_{i,k} \sim N(0,\sigma^2)$ to each of the 200 Doppler signals at 128 points. Later, with a fixed probability p_m, points were randomly selected and dropped from each function to evaluate the robustness to missing observations.

Table I shows the estimated number of clusters $\hat{d}_n$ and the percentage of misclassifications for different amounts of randomly missing data across $\sigma = 0.1, 0.06, 0.02$ when the data is fitted with Model 2. These figures were averaged over the best models from 100 simulations for each combination of σ and missing data probability p_m. For caomparison, similar results for the mixed effects model are shown in Table II.

Fig. 1 shows the model log-marginal likelihoods for different p_m when $\sigma = 0.06$. The most favoured models on the basis of log-marginal likelihoods had nine clusters followed by models with 7 to 11 clusters. First, the $p_m = 0.0$ case with no missing observations is discussed. Fig. 2 shows the nine clusters estimated in one of the simulations at $\sigma = 0.1$. In most of the situations the wavelet model performs better

Table I. Performance of the wavelet model at different SNRs and percentage of missing observations.

	$\hat{d}_n$				Misclassifications(%)			
σ	p_m =0.0	0.1	0.2	0.3	0.0	0.1	0.2	0.3
0.1	8.4	7.9	6.5	5.5	10.5	14.5	20.0	26.1
0.06	9.1	8.9	8.4	7.5	8.3	9.8	12.2	17.4
0.02	9.1	9.1	8.7	8.1	3.1	5.2	7.5	10.5

than the spline model and the estimated number of clusters $\hat{d}_n$ is consistently close to 9. The wavelet model also has lower misclassification rates than the spline model, indicating its robustness in high noise situations.

Table II. Performance of the mixed-effects spline model at different SNRs and percentage of missing observations.

	$\hat{d}_n$				Misclassifications(%)			
σ	p_m =0.0	0.1	0.2	0.3	0.0	0.1	0.2	0.3
0.1	7.3	6.8	5.2	4.8	18.4	22.5	24.9	36.3
0.06	9.5	8.9	7.5	6.3	11.5	16.1	20.4	27.4
0.02	9.1	8.6	7.6	6.9	4.0	7.4	11.0	17.0

The Besov parameters (a, b) are obtained by running the maximization procedure briefed in Section C.1 on a 2×2 grid of (a, b) pairs and we obtain the estimates as $a = 1.90$ and $b = 1.20$. We present the log Bayes factors to compare the model at $(a, b) = (1.90, 1.20)$ with some other neighbouring values of (a, b) in Table III.

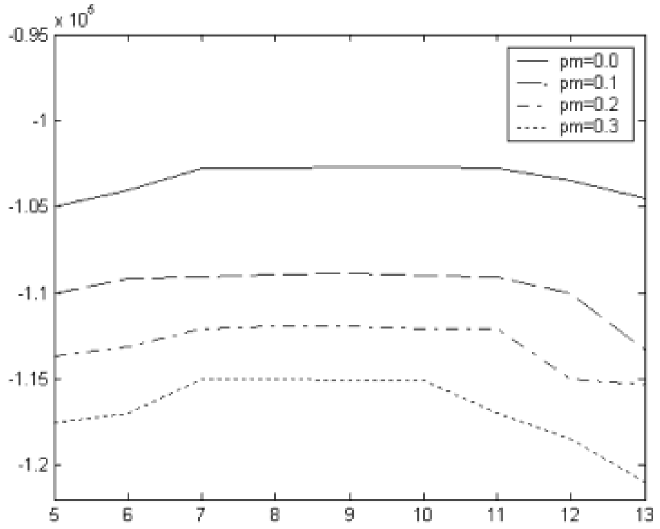

Fig. 1. Model log-marginal likelihoods vs number of clusters for shifted Dopplers data.

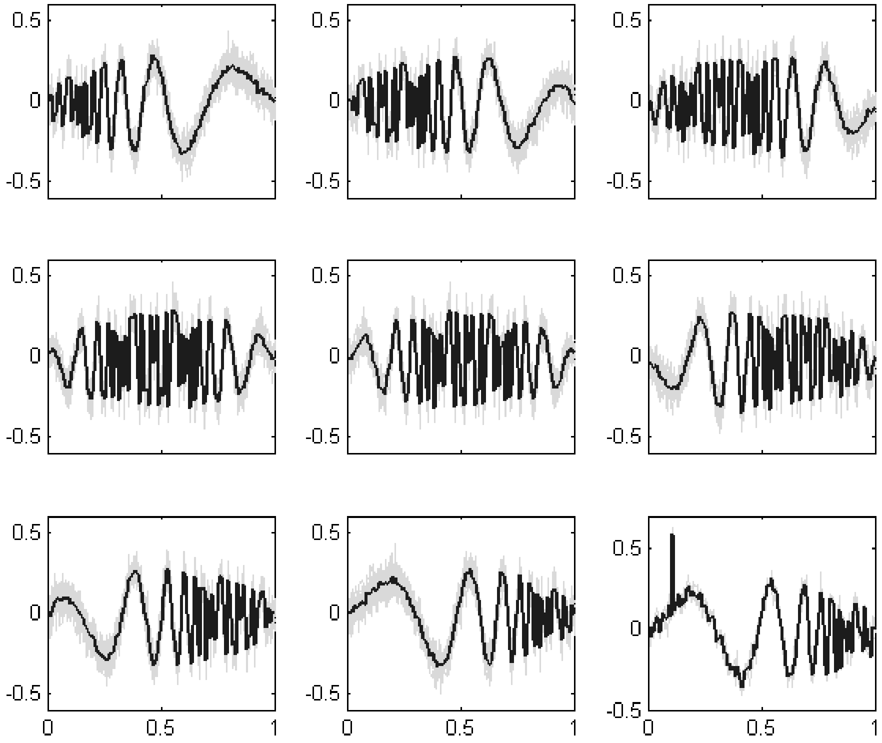

Fig. 2. Nine estimated clusters for the shifted Doppler signals.

Table III. Estimated log-Bayes factor with respect to the model for the shifted Dopplers data with Besov parameters 1.90 and 1.20.

	b				
a	1.00	1.10	**1.20**	1.30	1.40
1.75	16.78	13.24	9.53	12.63	15.95
1.85	14.07	12.64	3.79	5.98	14.47
1.90	14.52	10.80	**-**	6.44	13.74
1.95	19.45	17.98	15.84	12.17	15.61

a. Results for the missing data case

To evaluate the effects of missing data, points from each function were randomly selected and dropped with probabilities p_m. Tables I and II summarise the results from three separate simulations performed with $p_m = 0.1$, 0.2 and 0.3. At $p_m = 0.1$, our method performs similar to the complete data case, with a small deviation in the estimated size and a marginal increase in the misclassifications even at a $\sigma = 0.1$. At $p_m = 0.2$ and 0.3, the number of misclassifications increase and there is a drop in $\hat{d}_n$. Notably, the deterioration in the wavelet model (Table I) with higher amounts of missing data is less drastic than the spline model (Table II).

b. Predictive inference with missing data

A major advantage of the functional clustering procedure is that it can accurately predict unobserved portions of a curve. We demonstrate another contrived example where a portion of a curve is missing, we examine missing data prediction and its effects on clustering. A fixed portion of the tail from a curve in the shifted Doppler example is dropped and prediction bands are generated from the MCMC samples.

We expect the clustering algorithm to be reasonably stable when a few curves are partially observed owing to the shrinkage within clusters and this is confirmed in the prediction bands plotted in Fig. 3. With 12.5% missing points the effect of missing data prediction is hardly visible and although the bands widen with increasing number of missing points, it is only when this number gets to 37.5% that the curve is occasionally thrown into a distinct cluster.

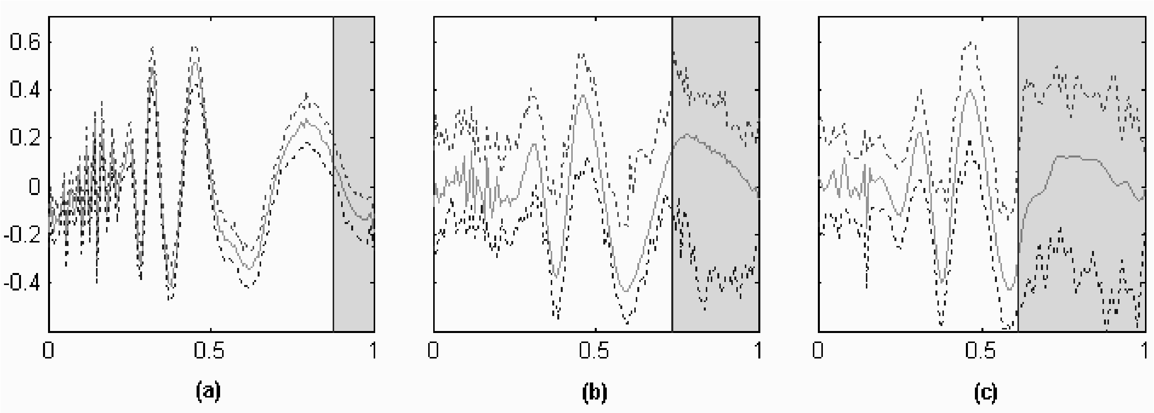

Fig. 3. MCMC prediction bands for a partially observed curve in the shifted Doppler example. Missing points in the three sub-plots (a),(b),(c) are 12.5%,25%,37.5% respectively.

c. Effect of scaling coefficients

Three types of shifted Doppler datasets are considered (see Fig. 4) for the analysis with the first two types ((a)–(b)) having curves that differ either in their scaling coefficients, or detail coefficients. Dataset (c) has differences in both the scaling and detail coefficients. Each curve is replicated five times in normal noise ($\sigma = 1$) so that the sample size n is 15 for each dataset.

For diffuse specifications in all the three cases, we set $r_0 = 2.2$ and $s_0 = 4.2$ with a prior mean, $\mathrm{E}g_0 = 1$ and variance, $\mathrm{var}(g_0) = 10$. The empirical estimates of (r_0, s_0) for the three datasets calculated in the aforementioned manner are $(7.991, 6.695)$,

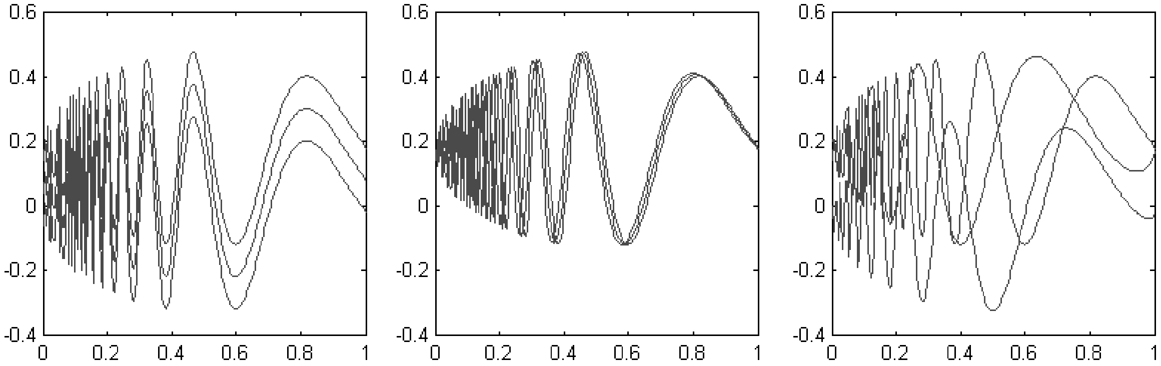

Fig. 4. Three different Doppler datasets with three classes used to assess the model sensitivity to scaling coefficients.

$(5.7961, 11.315)$ and $(5.110, 6.5823)$. The corresponding priors means of g_0 are 1.7019, 0.6222, 1.1153 and the prior variances are 2.1491, 0.10584, 0.96344.

Table IV summarises the performance of these priors applied to the three datasets. In dataset (a) only the scaling coefficients play a role in the clustering. The tighter empirically estimated prior produces better estimates of d_n with lower misclassification rates than the diffuse prior. The differences between the curves in dataset (b) are almost entirely encoded in the detail coefficients and the prior modelling of scaling coefficients does not play a role in the clustering, as shown in Table IV. In dataset (c), both the scaling and detail coefficients differ with clear evidence of three clusters in the latter. Although the role of scaling coefficients is diminished, the empirically estimated prior still manages to outperform the diffuse prior.

6. Yeast cell cycle data

Recently there has been a huge interest in the analysis of gene expression data from DNA microarray experiments. When microarray experiments are performed consecutively in time, we call this experimental setting a time course of gene expression

Table IV. Comparison of prior choices for the scaling coefficients for different Doppler signals. Actual number of clusters=3.

	Misclassifications(%)	
Set	Prior=Estimated	Diffuse
(a)	7.0	12.2
(b)	13.5	14.1
(c)	4.6	10.5

profiles. Clustering of the time course data gives insight about genes that behave similarly over the course of the experiment. By comparing genes of unknown function with profiles that are similar to genes of known function, clues to function may be obtained. Hence, co-expression of genes are of interest.

We analysed a similar dataset due to Spellman *et al.* (1998) which measures the relative levels of mRNA over time from 6178 genes in α-pheromone synchronized yeast cell cultures. Of interest are the connected genetic regulatory loops controlling the *Saccharomyces cerevisiae* pheromone response pathways (PrP) and whether the involved genes can be identified by their characteristic expression profiles in one or more clusters. In the sexual reproduction of yeast there is an essential role of pheromone response and mating pathways that ultimately target the protein STE12 and bind DNA as a transcriptional activator for a number of other genes. This is a natural choice for our methodology because the intergenic regions in the yeast genome that are bound to STE12 are known from genomic location analysis in presence of pheromone (Ren *et al.*, 2000). With the induction of the α pheromone, the expression levels for the PrP genes show a steep rise, until an internal stabilising mechanism is triggered and the fervour dies down. The result is a spiky event (or more contextually, a temporal singularity) in an otherwise smooth expression profile which for the most part, is comparable with the response of genes that do not participate in the pheromone response signaling and yeast mating.

The wavelet transforms give a time-frequency breakdown of information and the coefficients may reveal new patterns in frequency as well as time. For example, Klevecz (2000) performed a detailed wavelet analysis of the yeast cell cycle data and found significant high frequency artifacts isolated in time, contrary to the earlier notion that yeast cell cycle profiles are representative of slowly varying biological events. The hierarchical clustering model can easily delineate profiles with increased levels of gene expression occurring uniformly throughout the cycle from profiles characterized by sporadic bursts of spiky events (when relatively more messages are synthesised). The latter being a characteristic of the PrP genes is more important here. Note that this extremal behavior is easily accommodated within the Besov spaces.

In the experiments, 16 (of the 18) equi-sampled measurements over two cell cycles (lasting roughly 140 minutes) from 600 significantly expressed genes were considered. Some expression profiles were incomplete with a maximum of eight missing expressions per gene. To allow for possible deflections in the error variance in the population of gene expressions, we resorted to heteroscedastic model (Model 1) and found the estimated variance to vary between clusters. This may well be attributed to the significant deviations in the cluster sizes and the relatively short-sized expression profiles.

In general, the normality assumptions do not hold for microarray experiments and some pre-processing steps are necessary. For example, for the yeast data a log-transformation seems to suffice. This is confirmed by a Bayesian analysis (Chaloner & Brant, 1988) of the residuals. The residuals in the normal likelihood (1.3) are sampled from their posterior distribution conditional on the clustering configuration. From standard distribution theory, this posterior distribution is normal with mean

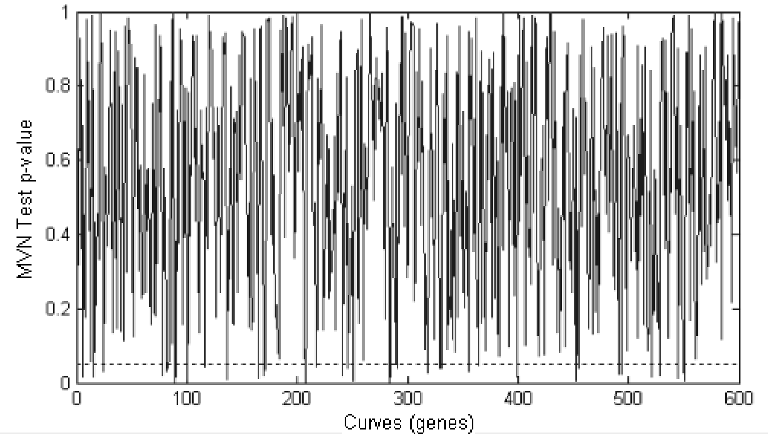

Fig. 5. p-values vs curves from multivariate test of normality for the yeast data.

$\mathbf{Y}_i - \mathbf{X}\boldsymbol{\mu}_i^*$ and covariance $\sigma^2\mathbf{V}^*$ ($\boldsymbol{\mu}_i^*$ and $\mathbf{V}^*$ defined in (1.9)). A multivariate χ^2 test is then performed to check the normality of the sampled residuals for each curve. The p-values (at a level $\alpha = 0.05$) from each curve are provided in Fig. 5 and show that most of the vector responses satisfy the normality assumptions.

The clustering algorithm showed maximum preference for models with 6 to 9 clusters (Fig. 6), of which two models with 8 clusters dominated the others in terms of the log model marginal likelihoods (plotted in Table V). Using a grid maximisation procedure, the Besov parameters (a, b) were set to $(1.45, 0.5)$ and this explains the spatial inhomogeneity in the expression profiles noticeable in the eight clusters (for one of the best models) plotted in Fig. 7. The plots can be divided between periodic (Figures 7.a-d) and non-periodic (Figures 7.e-h) patterns. The two clusters in the second category can be identified with the early on-off switch patterns and pertain to almost all the PrP genes mentioned above.

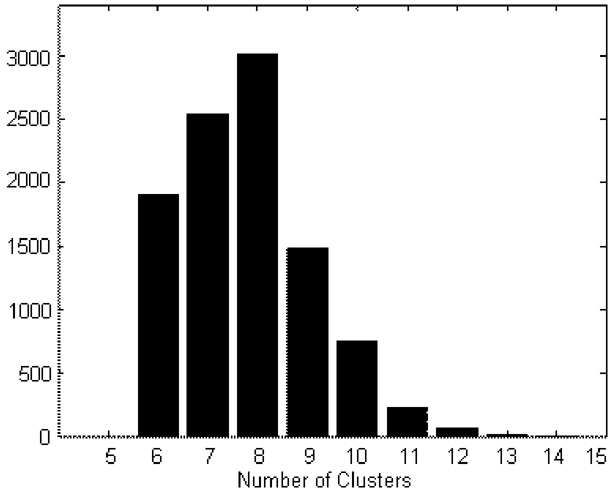

Fig. 6. Histogram showing the preferred number of clusters for the yeast cell-cycle data over 10,000 MCMC iterations.

Table V. Log marginal likelihoods of the best models for the yeast cell-cycle data.

d_n	Best models	Log marginal likelihood
8	1,2	-8.697×10^4
	3,4,5	-8.699×10^4
7	1	-8.700×10^4
	2,3	-8.701×10^4
6	1,2	-8.701×10^4
9	1	-8.703×10^4
	2	-8.704×10^4

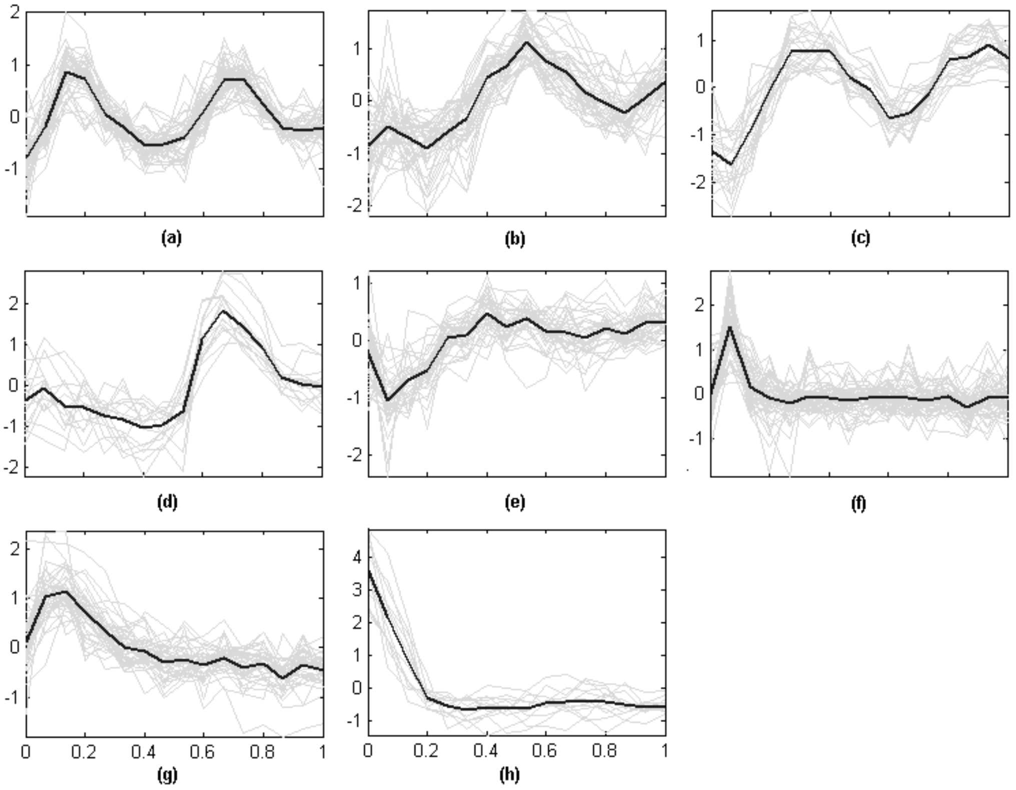

Fig. 7. Clustering of the yeast cell data. Eight clusters for the 600 expression profiles with 18 time points and a maximum of 8 missing points. Clusters (a)-(d) hold the periodic expression profiles and (e)-(h) hold the non-periodic on-off switching patterns.

a. Comparison with the spline model

The JS model is fitted with 4 quantile knots. The most preferred models vary in size from 5 to 7 and suggest over-smoothing and the models incapability to adjust to sharp fluctuations. The log Bayes factor of the best wavelet models compared to the best spline model was much larger than 10. In the results, not detailed here, there is a tendency for clusters (b)-(e) and (f)-(g)-(h) in Fig. 7 to merge together in one cluster. To emphasise this point, we fit three models differing in spatial adaptation to 3 (of the 8) clusters obtained from the wavelet model. The first cluster has periodic and smooth profiles, the second cluster is smooth but not periodic and finally, the third cluster is totally irregular and comes with a sharp on-off pattern of the PrP genes. The three rows in Fig. 8 plot the fits by a periodic (Fourier cosine series) basis, a spline basis and a wavelet basis divided in three columns corresponding to the 3 clusters. In the first column, we see that all three fit equally well, while in the second column, the periodic basis fit (Fig. 8.b) shows considerable bias at the boundaries. The situation deteriorates further as we move to the third column with a sharp fluctuation at $t = 0.1$ which is completely missed by both the periodic and spline models.

b. Further simulation study

The yeast cell cycle data is a typical cDNA microarray example where high noise levels makes inference difficult. Moreover, there is a marked heteroscedasticity in the data as indicated in Table VI tabulating the estimated variances of the eight clusters of Fig. 7. A follow-up simulation procedure is useful in such situations to validate the results. In order to simulate a realistic dataset for comparing the successful wavelet and spline models, we used the yeast cell cycle data as a prototype. As realistic values of the

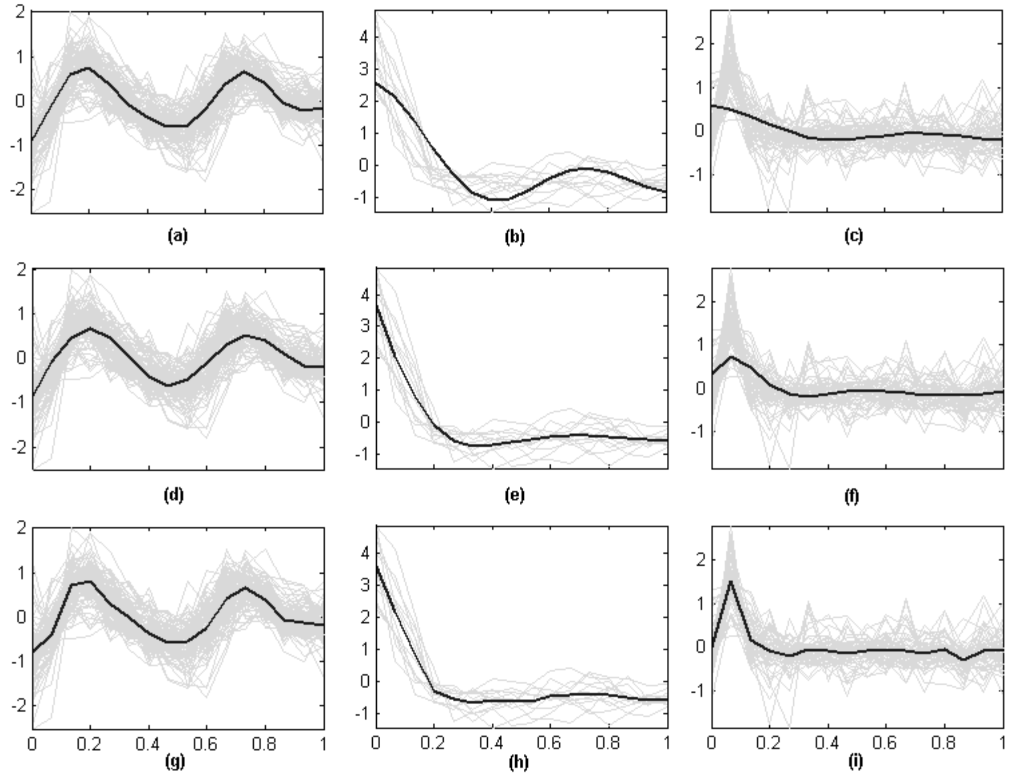

Fig. 8. Comparison of fits for three types of responses. Rows: (a)-(c) periodic basis fit, (d)-(f) spline fit and (g)-(i) wavelet fit. Columns: (a)-(g) periodic smooth function - all the bases fit well, (b)-(h) non-periodic but smooth function - the periodic basis has problems fitting at the boundary (c)-(i) Non-periodic and irregular function - only the wavelet basis captures the sharp fluctuation.

parameters, we use the representative curves (reproduced from the estimated wavelet coefficients) of the eight clusters plotted with thick lines in Fig. 7 and replicate them in i.i.d. normal noise with the estimated variances of Table VI following the structure of our model to generate the responses. In other words, this is an imitation of the original 600 gene expression profiles, to which we want to apply the algorithm and confirm our findings. This simulation is repeated 100 times to generate 100 different data sets, which are later analysed using wavelet as well as spline models to obtain the average misclassification rates.

The estimated number of clusters averaged over 100 simulations for the wavelet model is 8.21 with a very low average ‘misclassification rate’ (the deviation from the previously estimated clustering configuration) of 5.38%. In fact, the estimated clusters in these simulations almost always resemble Fig. 7 with differences due to occasional switchover of curves between clusters (f) and (g); or the formation of new clusters out of (or from combination of) clusters (f), (g) and (h). For the spline model, we see a lot of clusters merging due to over smoothing and the average number of clusters is 5.96 with a misclassification rate of 36.42%.

7. Precipitation spatial time series data

We consider a NCEP reanalysis spatio-temporal data that records the daily precipitation over Oregon and Washington between 1949 and 1994 (Widmann and Bretherton, 2000). The gridded observations represent area-averaged precipitation on a 50km grid. Bi-weekly averages of the daily observations from 179 locations are used, with a total of 512 time points over a time span of roughly 5 years between 1989 and 1994.

The example illustrates the potential application of functional clustering for the topographical categorisation of meteorological factors such as precipitation, temperature, snowfall, etc. It is usually difficult to generate topographical contour maps

Table VI. The size and the estimated variance of the eight clusters for the yeast cell-cycle data.

Clusters	# of curves	Estimated variance
(a)	124	0.3519
(b)	163	0.2996
(c)	21	0.8154
(d)	33	0.8862
(e)	59	0.2893
(f)	98	0.3006
(g)	24	0.2694
(h)	78	0.5360

of precipitation vs elevation although these are of great interest in climate analysis. A functional clustering model provides a natural way to group similar precipitation patterns viewed as functions and associate them with elevation. This can also deal with the problem of missing points in precipitation analysis. Missing points are typically interpolated with information from satellite observations and analysis amidst the differences in the measurement errors from two sources can be problematic.

Precipitation maps are formed by using the slope of a simple regression of the average local precipitation and the elevation. The clustered data plotted in Fig. 9 show a clear need for nonlinear modelling. This could in fact be used to delineate regions with occasional swings in rainfall patterns - patterns that can be overlooked by other geostatistical methods such as kriging (Rivoirard, 1994).

In many spatial models, a spatial random effects term viewed as a random intercept process is introduced to capture the spatial correlation. Previously, the intercept or scaling coefficients β_{i00} in our model were specified by nonparametric priors for clustering. Following Gelfand *et al.* (2003), we introduce a spatial random effect $\alpha(\mathbf{x}_i)$ that can be interpreted as a random spatial adjustment at a location $\mathbf{x}_i$ (in latitude and longitude) to the overall intercept β_{i00}. Thus for an observed set of locations $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$, we write $\mathbf{Y}_i = \alpha(\mathbf{x}_i)\mathbf{1} + \mathbf{X}\boldsymbol{\beta}_i + \boldsymbol{\varepsilon}_i$, where the overall intercept β_{i00} is an element of $\boldsymbol{\beta}_i$. We assume the prior distribution for $\boldsymbol{\alpha}$ is a zero mean Gaussian process with exponential correlation function $\tau^2\boldsymbol{\Phi}_\rho$. Here $\boldsymbol{\Phi}_\rho$ has a special structure in that its $(i,j)^{th}$ entry is $\exp(-\rho\|\mathbf{x}_i - \mathbf{x}_j\|)$ where $\rho > 0$ is a spatial decay parameter. Following Banerjee (2004), we assume a gamma prior for ρ so that the mean of the spatial prior range is half the maximum inter-site distance in the dataset, and τ^2 is a scaling parameter specified by a vague inverse-gamma prior distribution $IG(0.005, 0.005)$. The dependence between the $\alpha(\mathbf{x}_i)$ makes them identifiable from the other intercept terms without the need of replications.

For posterior inference, Gibbs sampling is used to alternately sample $\boldsymbol{\alpha}$ and the clustering parameters. More specifically, for a fixed $\boldsymbol{\alpha}$ let $\mathbf{Y}_i^* = \mathbf{Y}_i - \alpha(\mathbf{x}_i)\mathbf{1}$, then the posterior inference in Section D can be performed conditional on $\mathbf{Y}_i^*$. In Model 1, each α_i is updated separately by combining the implied conditional prior $\alpha_i|\alpha_{-i}, \tau^2, \rho$ (from the prior $\boldsymbol{\alpha}|\tau^2, \rho \sim N(0, \tau^2\boldsymbol{\Phi}_\rho)$) with the respective likelihood $\mathbf{Y}_i \sim N(\alpha_i\mathbf{1} + \mathbf{X}\boldsymbol{\beta}_i, \sigma_i^2\mathbf{I})$. For the homoscedastic case, we can directly work with the joint prior and the joint likelihood to draw multivariate samples of $\boldsymbol{\alpha}$. Finally, the parameters ρ and τ^2 are updated by separate Metropolis-Hastings steps by conditioning only on $\boldsymbol{\alpha}$.

The NCEP reanalysis dataset was fitted with Model 1 and as expected there were considerable differences in the estimated variance between clusters. The estimated value of ρ is 0.0182 and its posterior distribution from MCMC is shown in Fig. 10. This suggests a high spatial correlation between different locations. The overall homogeneity associated with the annual events and the local bumpiness due to fluctuations in rainfall is described well by the estimated Besov parameters, $(a, b) = (0.95, 0.3)$.

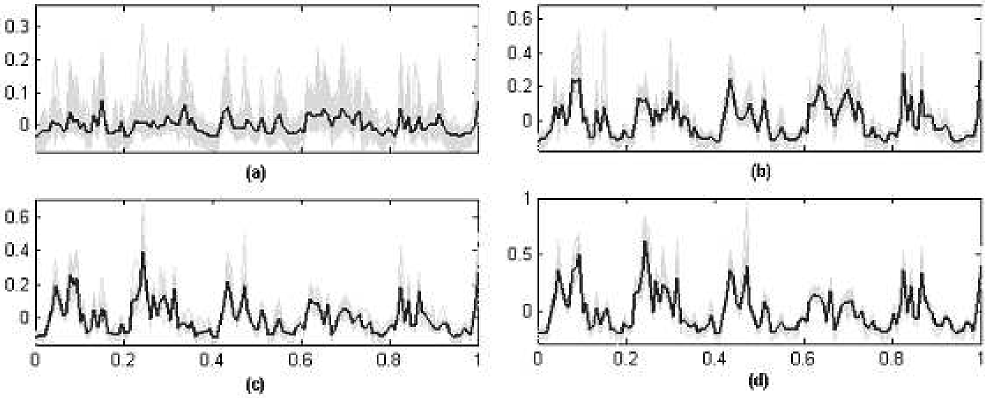

Fig. 9. Four clusters for the precipitation data.

The histogram of the MCMC samples for the number of clusters from one simulation in Fig. 11, shows clear preference for models with four clusters. The estimated clusters from one simulation are shown in Fig. 9 and their distribution on a geographical scale is contour-plotted in Fig. 12 at two different orientations. Fig. 9.a plots the largest cluster and corresponds to a large number of stations outlined by cluster 1 in Figures 9.a-b. The average annual rainfall in these areas has shown moderate fluctuations over the five year period and is notably less than the areas in clusters 2-4. Stations in cluster 3 (Fig. 9.c) have experienced heavier-than-usual rainfall between 1990-1991, but otherwise the average rainfall is comparable to cluster 2 (Fig. 9.b). Stations in cluster 4 (Fig. 9.d) although much wetter, share the same pattern as cluster 3, suggesting their geographical proximity which is confirmed from Fig. 12.

G. Discussion

The non-parametric Bayes model offers a flexible approach to functional clustering and has been shown to perform favorably against other functional clustering methods. Special stress is laid on the overall applicability of the methodology in that we rely on straightforward Gibbs sampling methods that are usable with high-dimensional

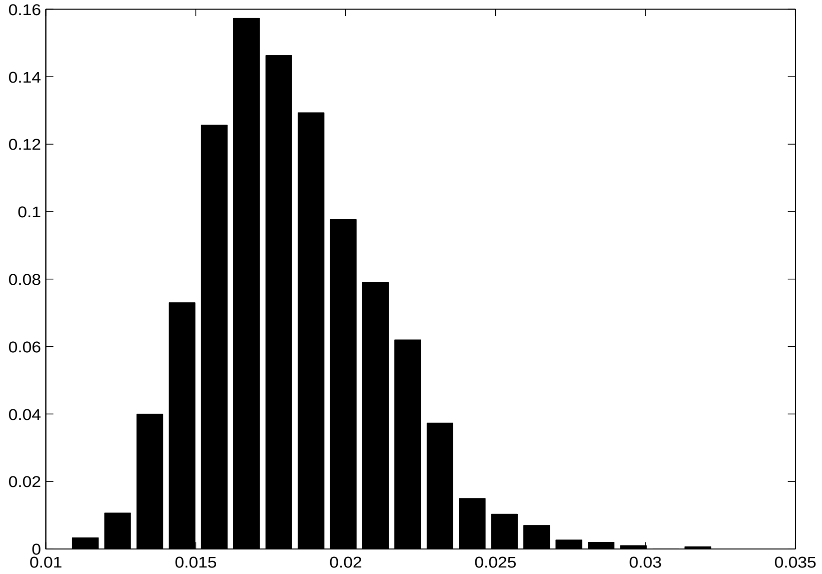

Fig. 10. Posterior distribution of the covariance parameter generated from 10,000 MCMC iterations.

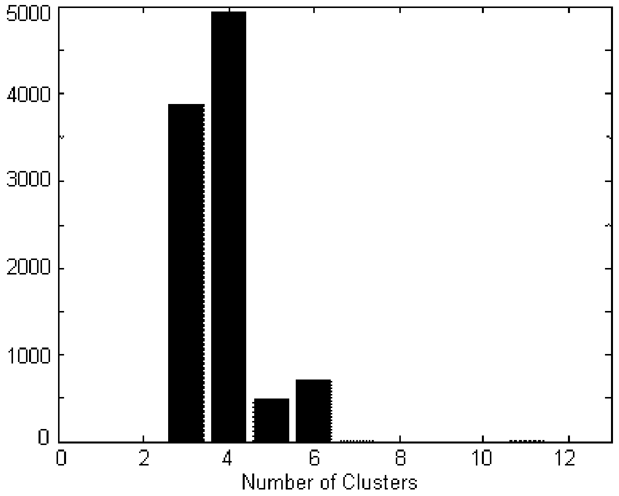

Fig. 11. Histogram showing the preferred number of clusters for the precipitation data over 10,000 MCMC iterations.

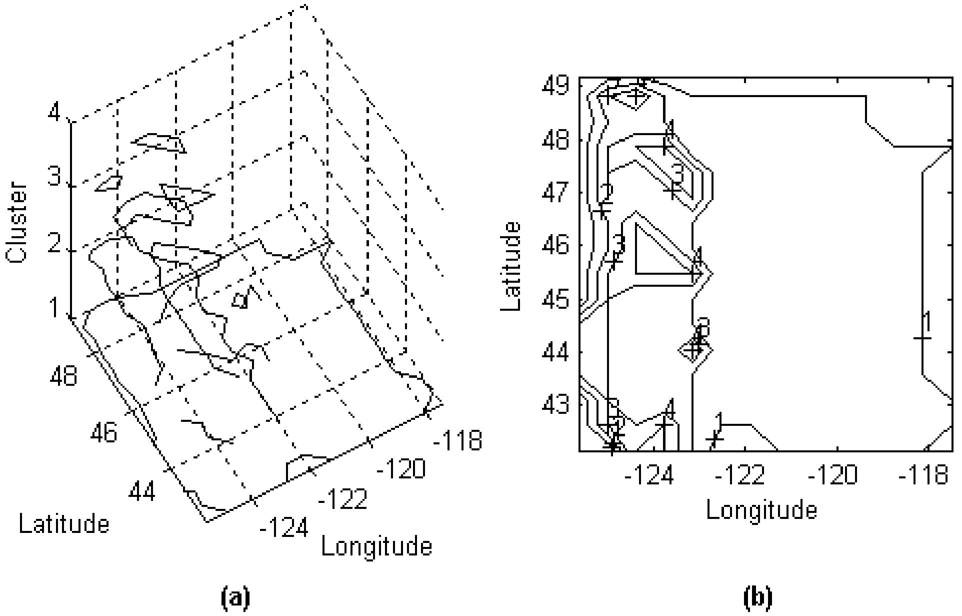

Fig. 12. Topographical distribution of the four clusters shown in two different orientations.

data, employ simple base prior modelling of the wavelet coefficients to encompass a large class of functions and address the missing data problem common in real-life applications. In addition, the method learns about the number of clusters in an automated manner unlike other clustering methods where a dimension change comes with a lot of computational burden.

In its ability to partition the predictor space into 'i.i.d.' regions, the Dirichlet process is comparable to product mixture models for clustering. (Indeed, Quintana and Iglesias (2003) show an equivalence under certain regularity conditions on the Dirichlet process.) This entails the use of two distinct approaches to Gibbs Sampling in this paper. First, the sampling of $(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n)$ from the Pölya urn allows the update and clustering of these parameters in a unified way; and replaces the reversible jump sampler (Green, 1995) used in product models for clustering, which has a reputation for being complicated. However, conditional on a sampled configuration of clusters, the remaining parameters are conveniently drawn from the product mixture provided by the Dirichlet process.

The discrete wavelet transform (1.1) requires that the number of sampled points m be a integer power of 2. The proposed model can be used with more flexible alternatives, such as the Lifting scheme (Sweldens, 1996), that does not place restrictions on the discrete support. This could additionally allow the extension of this model to un-equispaced data. Another interesting research problem would be to modify Quintana and Iglesias (2003) method in this high-dimensional functional clustering problem with the presence of missing data.

CHAPTER II

FUNCTIONAL CLUSTERING FOR IDENTIFICATION OF ALTERNATIVE SPLICE VARIANTS

A. Introduction

Alternative Splicing (AS) is a mechanism that increases the protein diversity in vertebrates. AS is considered to be an important source of functional complexity and disease; and about 40-60% of human genes have alternative splice forms. With the advent of new exon junction or splice microarrays, we can study the tissue distribution of splice forms for several thousand genes.

In this chapter, we propose a nonparametric Bayesian model for identifying the number of gene products produced from important genetic loci and to obtain clues regarding the regulation of Alternative pre-mRNA Splicing across different tissues, with the aim of aiding future exploratory analysis for isolation of splice forms and providing information on specific targets.

The work focuses on the analysis of the Rosetta gene expression dataset (Johnson *et al.*, 2003) from five Agilent chips covering over 10,000 multi-exon human genes in 52 tissues and cell lines. The results implicate more alternatively spliced (70-75%) human multi-exon genes than suggested by previous studies. In addition, there are a number of splice forms that appear to be unique to cancer. These new forms may be potential new targets for oncology drug discovery.

Section B provides a brief introduction to AS mechanisms and their contribuition to the protein diversity in humans. Abnormalities in AS can be associated with many types of cancer and other diseases. We discuss the importance of understanding these mechanisms in drug discovery and biology. In addition, we briefly review the

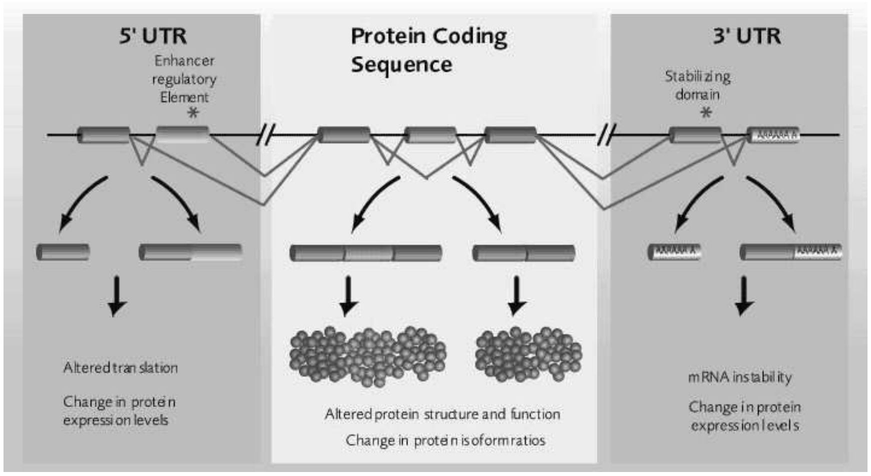
5' UTR
Enhancer regulatory Element
*
Protein Coding Sequence
3' UTR
Stabilizing domain
*
Altered translation
Change in protein expression levels
Altered protein structure and function
Change in protein isoform ratios
mRNA instability
Change in protein expression levels

Fig. 13. Alternative pre-mRNA splicing.

microarray based approches that have been used to study alternative splicing. In Section C, we extend the functional clustering model presented in Chapter I for the clustering the gene expression profiles from spice microarrays. The posterior specifications are mentioned in Section D. Finally, in Section E, we present results from the analysis of the Rosetta gene expression data (Johnson *et al.*, 2003).

B. Alternative Pre-mRNA Splicing

Vertebrates display a broad spectrum of functional and behavioural complexity that is a result of an increase in the number of components (structural, chemical etc) or their interaction during the course of evolution. The size of the proteome or the protein diversity has been expanded by increasing the number of genes and inventing new mechanisms such as varying the transcription start sites, alternative pre-mRNA splicing, polyadenylation and post-translational protein modifications. Of these, alternative pre-mRNA splicing is probability the most important source of protein diversity in vertebrates. Figure 13 gives a schematic illustration of the alternative pre-mRNA splicing mechanism.

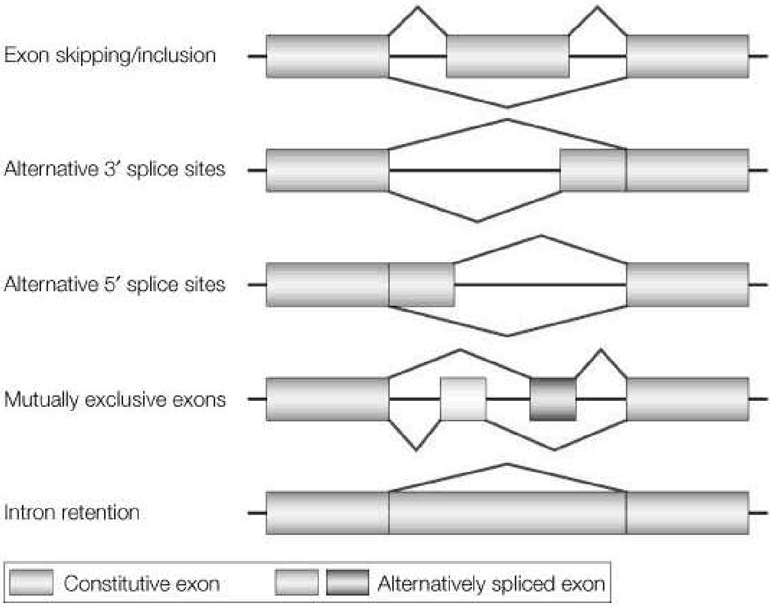
Exon skipping/inclusion
Alternative 3′ splice sites
Alternative 5′ splice sites
Mutually exclusive exons
Intron retention
Constitutive exon
Alternatively spliced exon

Fig. 14. Common modes of alternative splicing.

AS is a mechanism that generates multiple mRNAs from the same pre-mRNA by combinatorially joining 5' and 3' sites. For example, exons can be extended or shortened, skipped or included and increns may be removed or retained in the mRNA (see Figure 14). Therefore AS events that affect the protein coding region of the mRNA will give rise to proteins which differ in their sequence and their activities.

Enzymes known as spliceosomes are responsible for carrying out pre-mRNA splicing. Spliceosomes can identify 5' and 3' splice sites which are located on exon-intron boundaries. Exons are the functional portions of gene sequences that code for proteins and introns are the noncoding DNA sequences of unknown function that interrupt most mammalian genes and their number and size vary in different genes. The average size of a human exon is 150 nucleotides and the average size of intron is around 3,500 nucleotides. Thus, the splicing machinery has to recognize small exon sequences located within vast stretches of introns and exon recognition is a fundamental problem in pre-mRNA splicing. The accuracy of splicing is also monitored by RNA proof reading mechanisms that are able to target incorrectly spliced mRNA for destruction or can correct the error.

Given the complexity of the AS mechanisms it is not surprising that alterations in mRNA splicing can cause or be modified by a disease and characterization of these splice specific alterations can provide new therapeutic targets. Some examples of splicing defects that are associated with disease are:

1. **Spinal muscular atrophy** (SMA) is a disorder characterised by progressive loss of spinal cord motor neurons leading to paralysis. This disorder is a result of the deletion of the survivor of motor neuron gene (SMN1), which plays a role in the assembly of ribonucleoprotein complexes. This deletion prevents assembly of the U1 ribonucleoprotein complexes in the cytoplasm and results in global defects in pre-mRNA splicing.

2. **Duchenne muscular dystrophy** is characterized by the enlargement of muscles. It is one of the most prevalent types of muscular dystrophy and is characterized by rapid progression of muscle degeneration that occurs early in life. This dystrophy is caused by mutations in the dystrophin (DMD) gene that result in the insertion of a premature stop codon and the expression of a truncated inactive protein. An antisense technique has been used to induced exon skipping to restore the correct reading frame to frame-shifted mutant DMD genes.

3. **Colorectal Cancer.** The DCC (deleted in colorectal cancer) gene is regarded as a tumor suppressor gene and one DCC allele is deleted in roughly 70% of colorectal cancers and somatically mutated in others (Reale *et al.*, 1994). Several splicing aberrations have been observed between the first two exons in DCC due to which its expression is often reduced or absent in colorectal cancer tissues and

cell lines. Our analysis of the Rosetta dataset shows that the expression profile of DCC in the colorectal tissues is different than what is observed in other tissues. This points to a different protein product of DCC in the colorectal region that might be altered in some cases resulting in colorectal cancer.

1. Microarray approaches to splice form identification

Inkjet printing technologies allow rapid fabrication of customizable microarrays. Shoemaker *et al.* (2001) used this technology to monitor the coordinate expression of 8,183 exons annotated on chromosome 22q. Since alternative splicing of a given gene creates different exon-exon junctions, this technology can be adapted to detect alternative splicing by designing probes that span specific exon-exon junctions and measure the hybridization of mRNA samples from different tissues to these probes. Although the hybridization ratios of most exon-exon junction probes for a given gene will be constant, there will be some junctions for which the regulation will differ due to AS.

C. Nonparametric Model for Identifying Alternative Splice Forms

The Rosetta experiment (Johnson *et al.*, 2003) uses probes that spans consecutive exon junctions. In particular, 8 nt sequences are printed side-by-side as 16 nt sequences in separate probes of the microarray. For a given sample, 10,000 genes are probed in this fashion across 5 different microarrays. This is repeated for 54 different tissues or cell lines and the whole experiment generates $54 \times 5 = 270$ microarrays in total. Figure 15 gives a schematic illustration of the structure of the Rosetta dataset.

For our clustering analysis, we concentrate on the expression levels observed in consecutive exon junctions for one gene across all the 54 tissues or equivalently, 54

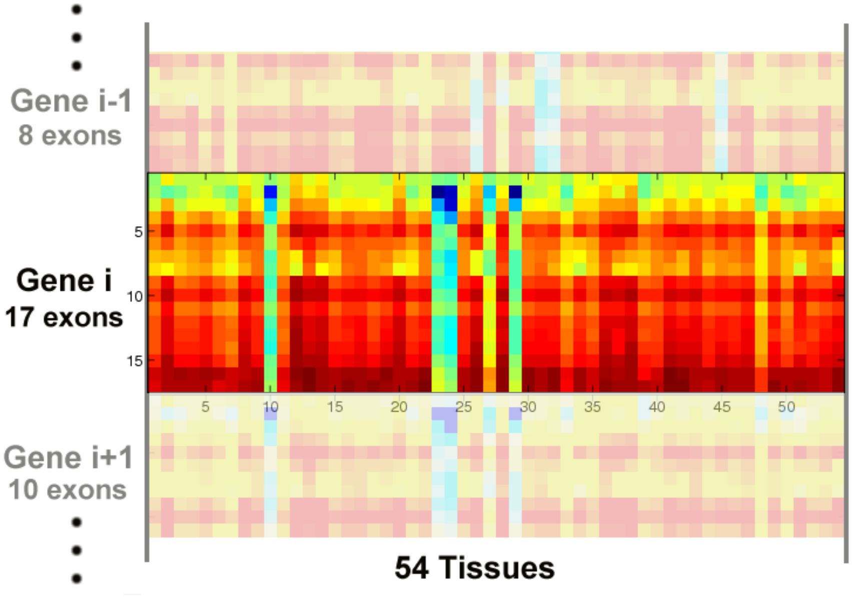

Fig. 15. Structure of the Rosetta dataset. The data can be viewed as a tall table with rows corresponding to consecutive exon junctions of all the genes. The diagram shows the 17 exon junctions corresponding for the ith gene.

differently microarrays. Assuming the microarrays are co-lognormalized, we consider a collection of $m \times 1$ vector responses, $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ where m is the number of exon junctions for the analyzed gene with the assumption that the number of tissues is n. The response $\mathbf{Y}_i$ from the ith tissue can be regarded as a curve that shows the regulation across consecutive exon junctions for a particular gene. Figure 16 shows the expression profiles for the Amyloid β-precursor protein (APP) gene from five tissues. Thus the functional clustering model described in Chapter I can be applied to the problem of identifying different splice forms. The only difference is that for each curve $\mathbf{Y}_i$ the consecutive samples $(y_{i1}, \ldots, y_{im})$ are not equispaced because the exons (or the exon-junctions) are separated by introns that may differ in size. As we will discuss later, almost orthogonal wavelet transforms for the unequispaced case can be produced by using the lifting scheme (Sweldens, 1996).

Given the co-lognormalized set of responses for any gene $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$, we use a

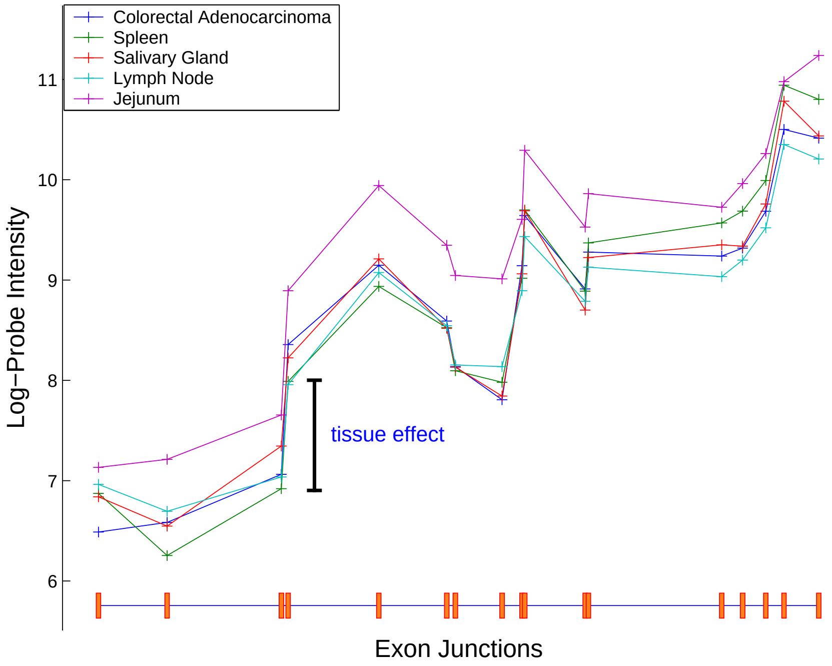

Fig. 16. An example set of log-expression profiles of the APP gene in five tissues. The red bars on the x-axis show the location of the exon junctions.

multiplicative measurement error model to express the response vector from the ith tissue as

$$\mathbf{Y}_i = \mu \mathbf{1}_m + \tau_i \mathbf{1}_m + \mathbf{X}\boldsymbol{\beta}_i + \boldsymbol{\varepsilon}_i,\ \boldsymbol{\varepsilon}_i \sim N(0, \sigma_i^2 \mathbf{I}_m) \tag{2.1}$$

where μ is the overall effect of the n microarrays, α_i is the constant tissue specific effect, $\mathbf{X}$ is an almost orthogonal wavelet transform defined over the unequispaced grid of exon junctions and $\boldsymbol{\beta}_i$ are the corresponding wavelet coefficients. We assume that the error terms are independent and all the correlation between the exon junctions is functionally encoded in the residual curve $\mathbf{X}\boldsymbol{\beta}_i$ after accounting for the overall background effects, μ in all microarrays and the tissue or individual effects, τ_i in absence of replications. Thus $\mathbf{X}\boldsymbol{\beta}_i$ can be regarded as the tissue specific differential expression profile of the analyzed gene.

The observational model (2.1) makes many simplifying assumptions. For convenience, we work with a multiplicative measurement error model and we study one gene at a time for computational feasibility, ignoring the biological dependence on other genes. The primary aim of this study was to estimate the number of splice forms for roughly 10,000 genes and from a computational viewpoint its a formidible task to work with a correlated model for a such a large dataset. In addition, each tissue or cell line is usually acquired from different subject and the above model fails to account for any subject or age-based variation. This brings us to the important question of whether it is reasonable to compare expression profiles from different subjects? We believe that such differences are largely reflected in the intensity, the overall shape of curves remains same (unless there are mutations along that chromosomal segment).

1. Prior elicitation

The prior distributions in the present model is largely follow from the prior elicitation described in Chapter I in that we need to specify μ and τ_i for all $i = 1, \ldots, n$. Both these paramters are treated as fixed effects viewing (2.1) as a mixed effects model.

In absence of any replications, the α_i's represent the confounded tissue and individual effects. The gene concentration in the ith tissue is a primary contributor to the tissue effect and can vary largely between tissues. Similarly, a large variability can be expected for individual effects. Therefore, μ and τ_i's are fitted by diffuse normal priors with a large variance. In addition, the tissue effects are assume to be a priori independent. More specifically, we let

$$\mu \sim N(0, s_0^2) \text{ and } \alpha_i \sim N(0, s_1^2\sigma_i^2), \forall i = 1, \ldots, n \tag{2.2}$$

where s_0^2 and s_1^2 take large values. Both these scaling or dispersion parameters are elicited by inverse gamma priors

$$s_0^2 \sim IG(u_0, v_0) \text{ and } s_1^2 \sim IG(v_1, v_1), \tag{2.3}$$

with hyperparameter specifications that ensure large values with higher probability.

Following the modelling of wavelet coefficients in Chapter I, we assume nonparametric Dirichlet process priors for $(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_n)$. The task of identifying the splice forms is accomplished by clustering the wavelet coefficients tissuewise and their number is seen as the number of clusters estimated in this manner.

Summing up the developments in this section, the hierarchical model becomes

$$\begin{aligned} \mathbf{Y}_i &\sim N(\mu\mathbf{1}_m + \tau_i\mathbf{1}_m + \mathbf{X}\boldsymbol{\beta}_i, \boldsymbol{\varepsilon}_i\mathbf{I}_m), \\ \mu \sim N(0, s_0^2),\ \tau_i &\sim N(0, s_1^2\sigma_i^2),\ (\boldsymbol{\beta}_i, \sigma_i^2) \sim NIG(0, \mathbf{V}; u, v), \\ s_0^2 \sim IG(u_0, v_0),\ s_1^2 &\sim IG(v_1, v_1),\ g_j \sim IG(r_j, s_j), \\ \gamma_{j,k} &\sim Bernoulli(\pi_j), \\ \alpha &\sim G(d_0, \eta_0) \end{aligned} \tag{2.4}$$

where $i \in \{1, \ldots, n\}$, $j \in \{0, \ldots, J\}$, $k \in \{0, \ldots, m_j\}$ and as one can recall from Chapter I, $\mathbf{V} = \text{diag}(\boldsymbol{\gamma})\text{diag}(\mathbf{g})$ where $\boldsymbol{\gamma} = (\gamma_{00}, \gamma_{10}, \gamma_{20}, \gamma_{21}, \ldots)$ is a vector of latent indicator variables for selection of each coefficient and $\mathbf{g} = (g_0, g_1, g_2, g_2, \ldots)$ are the corresponding scaling parameters. Also, m_j is the number of coefficients that is not necessarily a dyadic function of the level j due to the manner in which the wavelets are constructed.

D. Posterior Inference

As in Chapter I the use of conjugate base priors expedites the posterior sampling of $(\boldsymbol{\beta}_i, \sigma_i^2)$'s from the Pölya urn. The background and tissue or individual effect parameters are also sampled easily from the respective posterior distributions. The following subsections describe only the conditional posterior inference of these parameters and the relevant dispersion parameters such as s_0^2 and s_1^2.

1. Background and tissue or individual effects

The constant background effect μ has contributions from all the probes from all n microarrays. Therefore, the posterior conditional distribution of μ is calculated by combining its diffuse normal prior distribution and the likelihoods (2.4) for $i = 1, \ldots, n$; and is given by

$$\mu|\boldsymbol{\tau}, \boldsymbol{\beta}, \boldsymbol{\sigma}^2, \mathbf{Y}, s_0^2 \sim N\left(\sum_{i=1}^{n} \frac{s_0^2}{s_0^2 + m\sigma_i^2}(\mathbf{Y}_i - \tau_i \mathbf{1}_m - \mathbf{X}\boldsymbol{\beta}_i' \mathbf{1}_m), \sum_{i=1}^{n} \frac{m\sigma_i^2}{s_0^2 + m\sigma_i^2} s_0^2\right) \quad (2.5)$$

where as before $\mathbf{Y}$ is the collection of all response vectors $(\mathbf{Y}_1, \ldots, \mathbf{Y}_n)$ and in similar notation $\boldsymbol{\tau}$, $\boldsymbol{\beta}$ and $\boldsymbol{\sigma^2}$ represent the tissue effects, the wavelet coefficients and the observatioal dispersion parameters for all n tissues respectively.

The posterior distribution for tissue specific effects τ_i is calculated by with respect to its diffuse normal prior conditional on the responses for the ith tissue and is given by

$$\boldsymbol{\tau}|\mu, \boldsymbol{\beta}, \boldsymbol{\sigma}^2, \mathbf{Y}, s_1^2 \sim N\left(\frac{s_1^2}{s_1^2 + m}(\mathbf{Y}_i - \mu \mathbf{1}_m - \mathbf{X}\boldsymbol{\beta}_i)' \mathbf{1}_m, \frac{m}{s_1^2 + m} s_1^2\right). \quad (2.6)$$

2. The scaling parameters

For computation convenience, the scaling parameters s_0^2 and s_1^2 are directly drawn conditional on μ and $\boldsymbol{\tau}$ respectively. Both parameters have inverse gamma posterior distributions after combining (2.2) with the respective inverse gamma prior distributions (2.3). The exact form of these conditional posterior distributions is

$$s_0^2 \sim IG(u_0 + \mu^2, v_0 + 1) \text{ and } s_1^2 \sim IG\left(v_1 + \sum_{i=1}^{n} \tau_i^2/\sigma_i^2, v_1 + n\right). \quad (2.7)$$

It is reasonable to argue here that there is little objectivity in these updates, especially s_0^2. However, this is not the aim here as the priors (2.3) are diffuse and the hyperparameters involved in the above updates are chosen to obtain large values of s_0^2 (and s_1^2). In fact, the above updates comprise a convenient way to uncertainty and for practical purposes, it might suffice to assign fixed large values to both these parameters.

The reader is referred to Chapter I for details about the posterior inference for the remaining parameters $(\boldsymbol{\beta}, \boldsymbol{\sigma^2}, \mathbf{g}, \boldsymbol{\gamma})$. Briefly, the sampling of the cluster specific paramters $(\boldsymbol{\beta}, \boldsymbol{\sigma^2})$ and the indicators $\boldsymbol{\gamma}$ requires additional conditioning on $(\mu, \boldsymbol{\tau})$ in addition to the $\mathbf{Y}$. Some minor alterations are also required to accommodate the non-orthogonality of the wavelet transform $\mathbf{X}$.

3. Estimation of the clusters and other parameters

The Bayesian computation is carried out in two steps. The first step consists of Gibbs sampling all the model parameters are sampled from their respective posterior distributions. This includes sampling of the background, the tissue effects, the cluster locations and all the scaling or dispersion parameters. After collecting sufficient number of samples using MCMC, all the parameters except for the cluster specific parameters, $(\boldsymbol{\beta}, \boldsymbol{\sigma}^2)$ are estimated. For the second step, we repeat the MCMC conditional on the estimates derived from the first step. The samples from this second run are then used towards the estimation of $(\boldsymbol{\beta}, \boldsymbol{\sigma}^2)$ and the number of clusters d_n as detailed in Chapter I.

E. Examples

The clustering scheme was applied to all the 10,000 genes in the Rosetta dataset to estimate the number of splice forms. This is a computationally demanding task and extensive database programming was used to generate the results. For illustration, we present the results for three genes from the Rosetta dataset whose protein products have been implicated in various diseases. Following Chapter I, all the reported results were averaged over 100 simulations with 10,000 iterations per simulation and a burn-in period of 1000. The average number of exon-junctions for the genes in the dataset is roughly 20 and thus the length of the curves m is small compared to the examples in Chapter I. The MCMC mixed reasonably well for all the genes in the dataset and seemed to converge much before the alloted burn-in time. The employed wavelet basis was constructed using the Lifting scheme (Sweldens, 1996) as described in the next subsection.

1. Wavelets for unequispaced design

For non-equispaced design, such as in the next two examples analyzing proteomics datasets, we use lifted wavelet transforms (Sweldens, 1996). These, unlike the traditional wavelet transforms, do not require regularly spaced samples. Traditional wavelet transforms (designed for equispaced samples) can be factored into a sequence of simpler transforms using the lifting scheme (Daubechies and Sweldens, 1996); and each lifting step is a refinement over the previous steps and represents an increase in the smoothness (or order) of the wavelet bases. These features can be extended to non-equispaced designs by allowing more flexible basis functions that are not simply translates or dilates of one fixed function and using the Lifting scheme to perform the construction in the time domain. The wavelets resulting from the lifting scheme still have all the powerful properties of traditional wavelets such as localization and good approximation. Despite these properties, the lifting scheme has been largely overlooked in recent literature and many authors have resorted to using interpolation for generating equispaced samples for their analysis.

The lifted construction used in the following examples involves two separate steps. The first step involves an *unbalanced* Haar transform, that is the usual Haar transform with adjustments for unequal distance between two successive observations. The coefficients from this transform are used as the input for a second lifting step that is an unbalanced version of a biorthogonal Spline wavelet. Thus the degree of the spline functions determines the smoothness of the overall basis. More details about such constructions can be found in Delouille, (2002). The wavelet transforms built in this manner are not orthogonal as in the previous examples. This does not overly affect the posterior inference or the performance of our model as we ensure near orthogonality of transform within the lifting scheme. The posterior distributions as calculated in the appendix can be easily extended to the case where $\mathbf{X}$ is not orthogonal.

2. APP (amyloid beta-precursor protein) gene

The APP gene has 17 exons and is known to have three splice variants in that it encodes a cell surface receptor and a transmembrane precursor protein that is cleaved by enzymes to form a number of peptides. Some of these peptides promote transcriptional activation, while others form the protein basis of the amyloid plaques found in the brains of patients with Alzheimer disease. Splicing abberations in this gene have been implicated in autosomal dominant Alzheimer disease and various tumors in the nerveous system such as Neuroblastoma, Neuroglioma etc.

The tissuewise clustering of the expression profiles for APP helps us to corroborate the usefulness of our method. The analysis reveals three clusters implying different protein products in the brain and some tumorous tissues. Figure 17 plots all the curves for the APP genes observed in 54 tissues and the three clusters are plotted in Figure 18 with a clustered heatmap in Figure 19.

The first cluster contains all the tumorous tissues and confirms the previous findings that some APP products may be directly or indirectly involved in tumors. The second cluster contains a large number of tissues from different parts of the body and we believe that this corrresponds to the dominant transcriptional promoter or cell surface receptor proteins. More importantly, a distinct splice form, in the form of the third cluster, is observed in tissues from different regions for the brain. This suggests that APP has an important function in the brain and abnormalities in the protein product due to splicing aberration or other somatic mutation may promote Alzheimer's disease.

3. DCC (deleted in colon cancer) gene

Loss of heterozygosity (LOH) on chromosome 18 is frequently observed in nearly 70% cases of Colon carcinoma. This deleted stretch of the DNA encompasses a tumor suppressor gene known as the deleted in Colon cancer (DCC) gene. DCC is expressed well in most normal tissues, including colon-rectal region. Netrin is protein product of DCC that accurately guides the developing axons for the establishment of neuronal connections. Additionally, somatic mutations in within the DCC have been observed and several splicing aberrations have been implicated in colorectal cancers. The DCC gene has 29 exons of which junctions for 17 consecutive exons were probed in the Rosetta dataset. Due to missing data and saturation issues we only analyzed the data for 39 of the 54 tissues.

Tissuewise clustering of the expression profiles for DCC confirms two separate

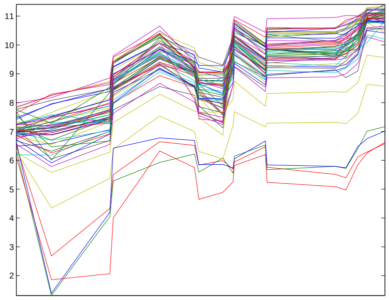

Fig. 17. All the expression profiles for APP gene from 54 tissues.

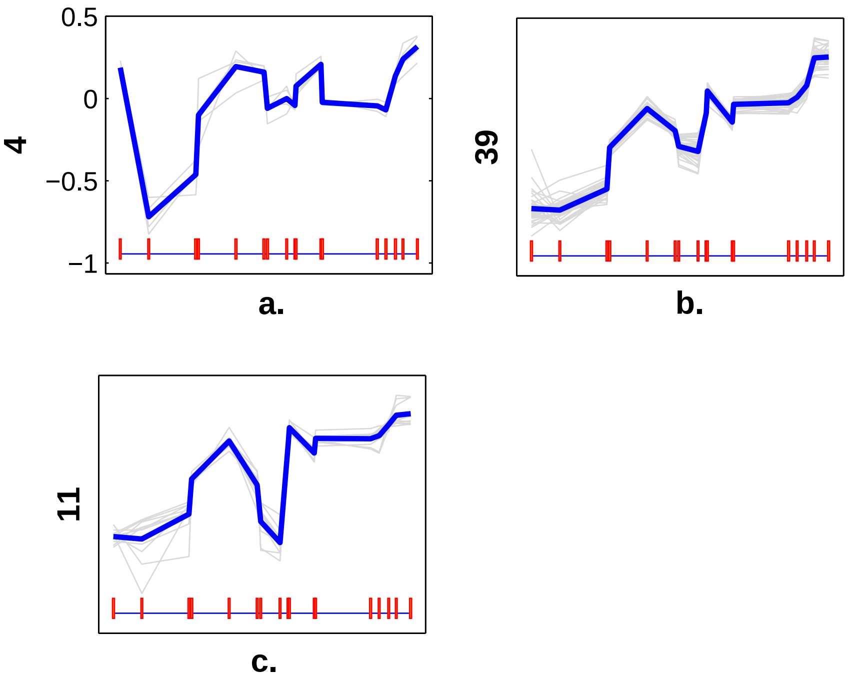

Fig. 18. Three estimated clusters for APP gene. a. Cluster 1 with 4 tumorous tissues, b. Cluster 2 with 39 tissues corresponding to the dominant form and c. Cluster 3 with 11 brain tissues.

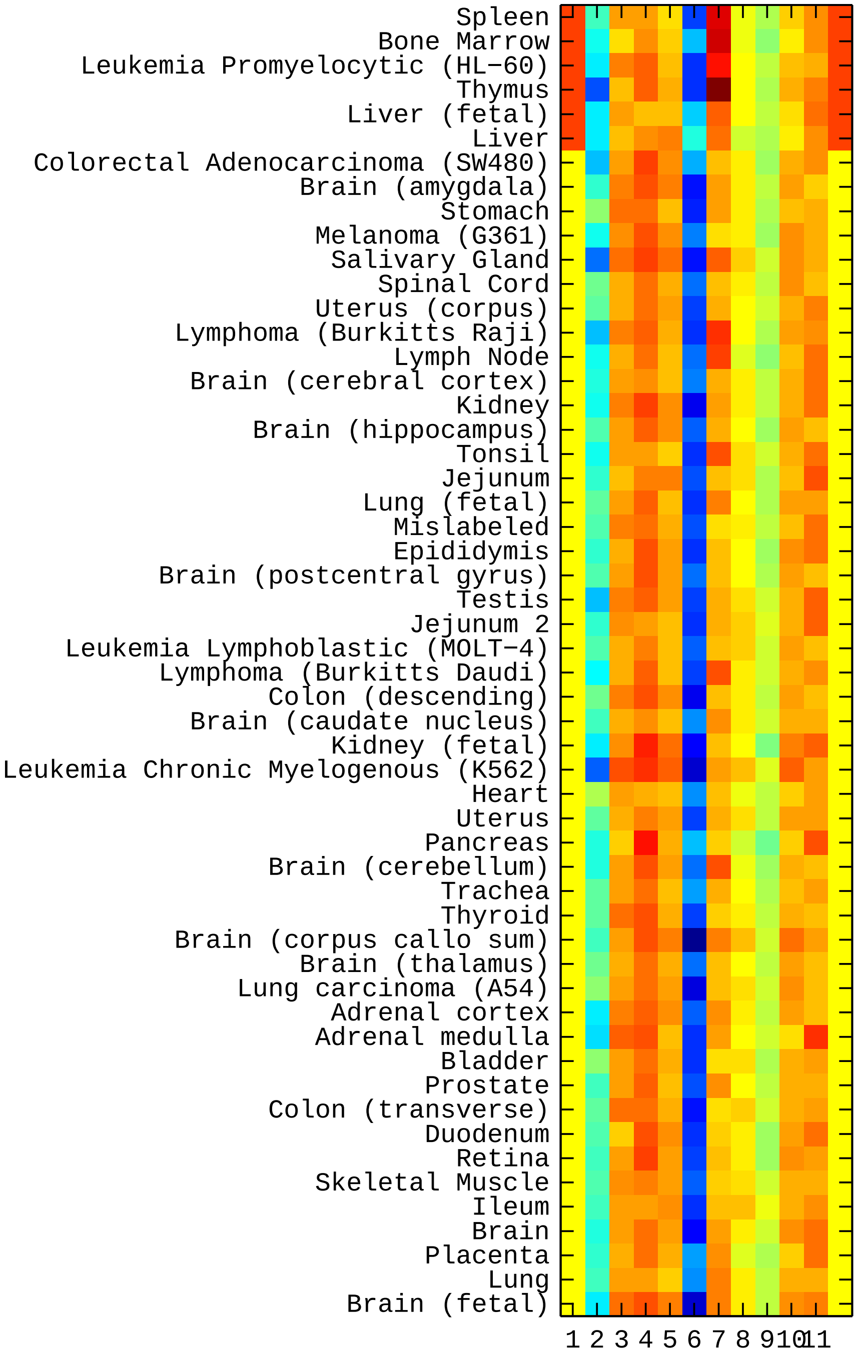

Fig. 19. Clustered heatmap of the three clusters shown in the previous figure.

splice forms in the nervous system and the colo-rectal region. Figure 20 plots all the curves for the DCC genes observed in 39 tissues, the four clusters are plotted in Figure 21 and a clustered heatmap of the clusters is shown in Figure 22.

The first cluster consists of several brain tissues comprise and conforms with the fact that DCC produces Netrin in the nervous system. The second cluster contains a large number of tissues from different parts of the body and may correspond to cumulative effect of different expression patterns of DCC. The third cluster is the most important cluster and it contains four tissues from the colo-rectal region namely, Colon, Ileum, Jejunum and cancerous colon-rectal mucosa. This strongly suggests a different role of the DCC in the colorectum. In addition, a comparison of the four curves within this cluster shows a reduced expression level in the first exon junctions. Therefore, a splicing aberration in the first two exons can be implicated in this case of colorectal cancer. The fourth estimated cluster is probably an outlier as it has a single tissue from the Testis with a very different expression pattern from the rest of the tissues.

4. Joint modelling of BCR-ABL

Leukemia is the uncontrolled proliferation of white blood cells and one of its common form is chronic myelogenous leukemia (CML). In many cases of CML, the leukemia cells share a chromosome abnormality not found in any nonleukemic white blood cells. This abnormality is a reciprocal translocation between one chromosome 9 and one chromosome 22 designated as t(9;22). It results in one chromosome 9 longer than normal and one chromosome 22 shorter than normal (known as the Philadelphia chromosome). The DNA removed from chromosome 9 contains most of the proto-oncogene ABL. The break in chromosome 22 occurs in the middle of the BCR gene. The fused ABL-BCR gene is a part of the Philadelphia chromosome and produces a

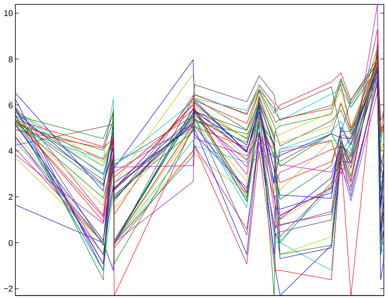

Fig. 20. Expression profiles for the DCC gene from 32 tissues.

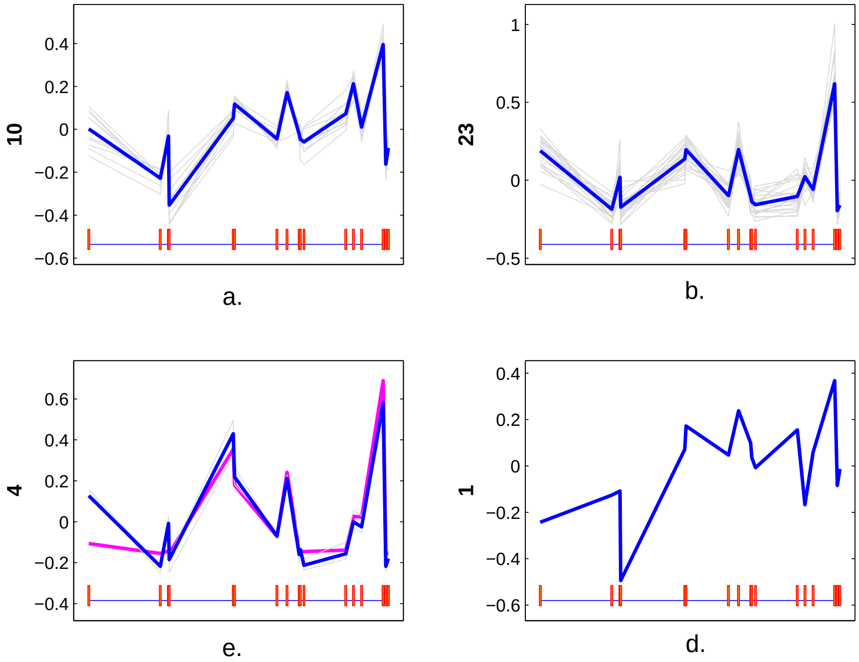

Fig. 21. Four estimated clusters for DCC gene. a. Cluster 1 has 10 brain tissues, b. Cluster 2 has 23 tissues from all over the body, c. Cluster 3 consists of tissues from the colon, ileum, jejunum and colorectal mucosa (plotted in purple) and d. Cluster 4 has a single tissue from the Testis.

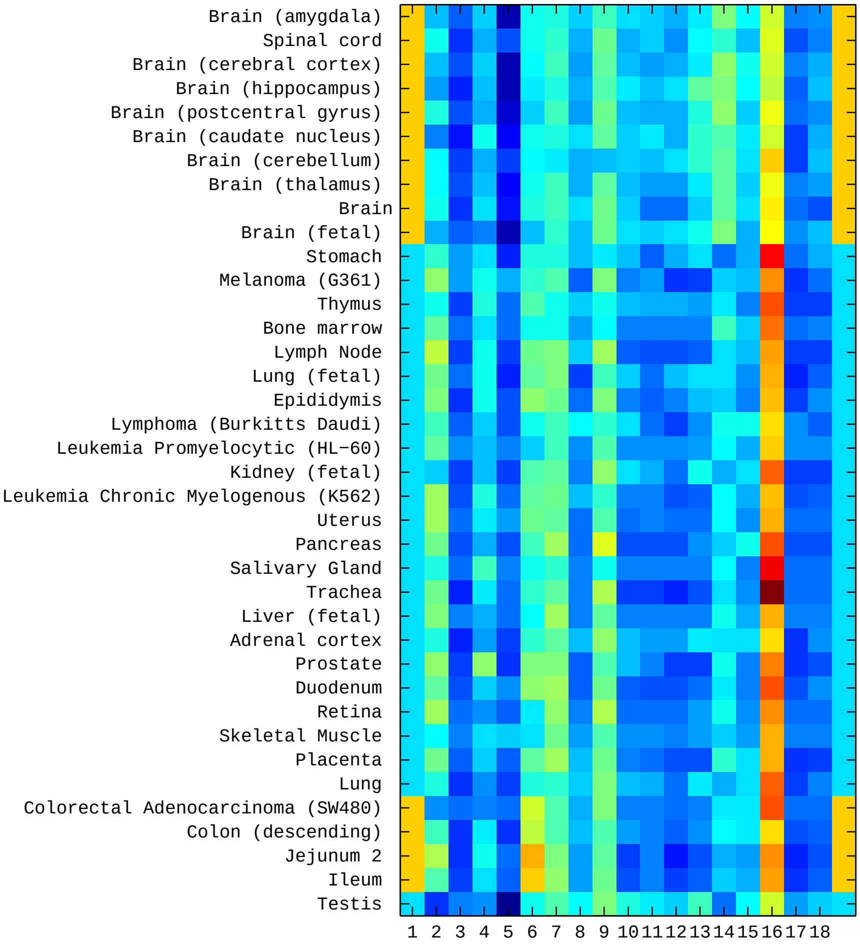

Fig. 22. Clustered heatmap of the four clusters shown in the previous figure.

abnormal protein that regulates or activates many cell processes that normally are turned on only when the cell is stimulated by a growth factor. This unrestrained activation leads ultimately to CML.

In this example, we want to determine whether the abnormal translocation can be identified by cluster. To this end we model the expression profiles for ABL and BCR jointly by stacking them one on top of another.

Tissuewise clustering of the combined expression profiles for ABL-BCR confirms an abnormality in a CML tissue that can be identified with the reciprocal translocation described above. Figure 23 plots together all the curves for the ABL-BCR genes as observed in 54 tissues, the four estimated clusters are plotted in Figure 24 and a clustered heatmap of the clusters is shown in Figure 25.

The first two clusters contain a large number of tissues from different parts of the body and may correspond to cumulative effect of different expression patterns of DCC. The second cluster has the CML tissue and the comparison with the expression profiles observed in other tissues reveals a very different pattern within BCR. The fourth cluster is probably an outlier as it has a single tissue from the bone marrkow with a very different expression pattern from the rest of the tissues.

F. Discussion

In this chapter, we report the tissuewise clustering of roughly 10,000 multi-exon genes. The results has been used to identify a number of splice forms that appear to be unique to cancer that can be future therapeutic targets. It also allows us to ascertain local intervention points for drugs. This method like all other in-silico methods, can help biologists by providing a hypothesis that must be verified later in the laboratory.

In future, we would like to improve this model to account for interactions or

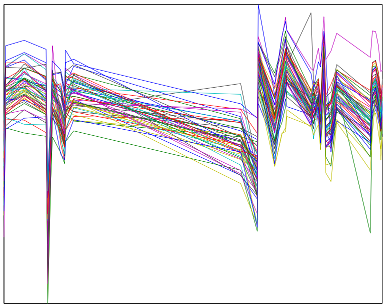

Fig. 23. Combined expression profiles for the ABL-BCR genes from 54 tissues.

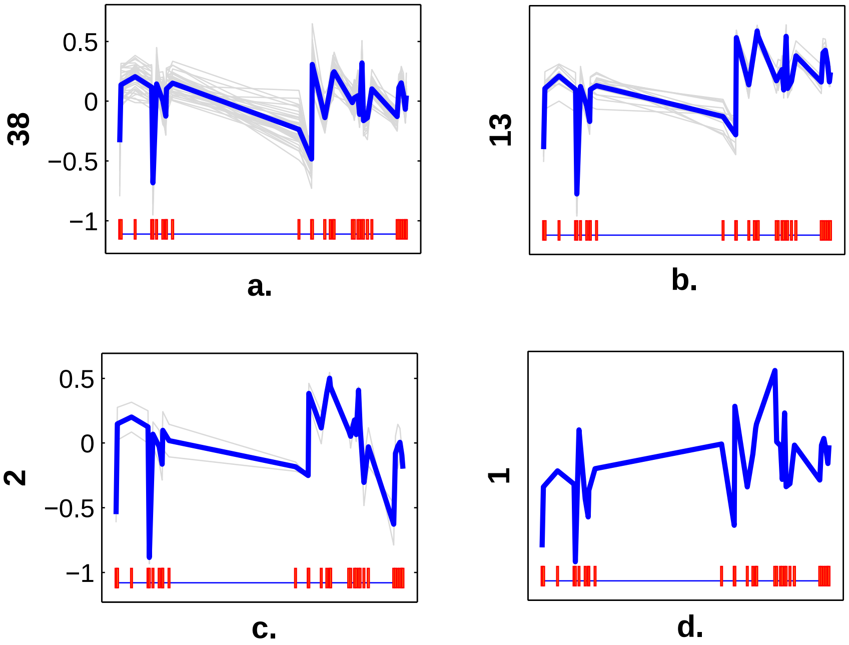

Fig. 24. Four estimated clusters from the joint modelling of ABL-BCR. a. Cluster 1 has tissues from all over the body, b. Cluster 2 has 13 tissues from all over the body, c. Cluster 3 consists of a CML tissue and d. Cluster 4 has a single tissue from the bone marrow.

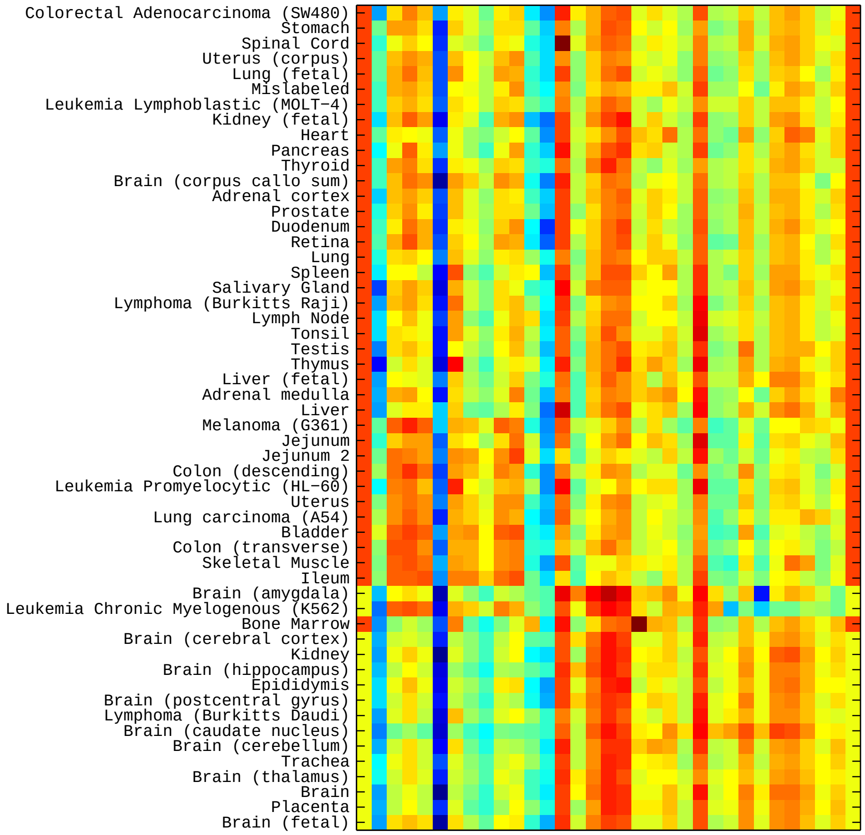

Fig. 25. Clustered heatmap of the four clusters shown in the previous figure.

dependencies between genes. This might enable us to study the cumulative effect of genes that are part of important genetic pathways.

CHAPTER III

BAYESIAN APPROACHES TO GRAPH PARTITIONING

A. Introduction

Graph cutting, partitioning or clustering as it is variously known is the decomposition of a graph into roughly equal sized pieces while minimizing the number of edges between those pieces (Chung, 1997). The model based graph partitioning approach proposed in this chapter draws motivation from the kernel k-means algorithm (Scholkopf *et al.*, 1998) that maps data points to a high-dimensional space in order to delineate non-linearly separable clusters and that has been shown to be mathematically equivalent to certain graph cutting algorithms (Dhillon *et al.*, 2004). This connection allows graph cutting to be performed with the simplicity of the k-means algorithm. Nevertheless, these graph cutting approaches are subject to the drawbacks of the classical k-means algorithm such as the specification of the number of clusters.

In this work, we show that the spectral graph cutting techniques can be related to certain nonparameteric Bayesian clustering methods which have the ability to learn the number of clusters from data. The proposed methodology enables the simultaneous estimation of the subgraphs as well as their number from the data. In addition, the set of all possible decompositions of the graph can be efficiently explored via MCMC. Optimal graph cuts in graph theory are obtained by minimizing intuitively defined loss functions. This optimization, however, is not straightforward due to the combinatorial nature of the problem and spectral graph theory is employed for approximate solutions.

In addition, the problem of graphically connecting a set of data points can be motivated within the statistically framework of graphical models. Let $G = (V, E)$

be an undirected graph describing the association between the vertices in vertex set V through edges in the edge set E. A graphical model is a family of probability distributions which is *Markov* in G (Lauritzen, 1996). The idea is to be able to fit a high-dimensional distribution in sets of univariate conditional distributions depicting patterns of association in a set or *clique* of variables. The vertex set V may consist of a vector of random variables, $\mathbf{Y}$ and the absence of an edge (i, j) indicates that $\mathbf{Y}_i$ and $\mathbf{Y}_j$ are conditionally independent. In other words, $\mathbf{Y}_i$ and $\mathbf{Y}_j$ are independent conditional on the other entries in $\mathbf{Y}$ whenever i and j are not neighbors under G.

When $P(G) = N(\mu, \Sigma)$ is a multivariate distribution with $\Sigma > 0$, we have a Gaussian graphical model. The covariance structure Σ is often unknown and there is a need for statistical methods for selecting the models that best fit the data. Bayesian approaches are well suited for graphical modelling as they provide posterior probabilities for comparing different graphical models.

In the graphical model literature, it is common to model the precision matrix $\Omega = \Sigma^{-1}$ rather than Σ itself as it allows better interpretation. For instance, $\omega_{i,j} = 0$ implies that i and j are conditionally independent or the edge (i, j) is absent in the graph.

Prior specifications in Bayesian methodology can be used to model desired graphical structures. It is common to assume *decomposable* graphs that are either complete or have a *decomposition* - there exists subsets $(A, B) \subset V$ such that 1. $A \cup B = V$, 2. $A \cap B$ is complete and 3. $A \cap B$ separates A from B. A non-decomposable graph is a graph with no decomposition.

For decomposable graphs, hyper-Markov priors were proposed by Dawid and Lauritzen, (1993) that essentially reproduce the factorization of likelihood at the a priori level. This is computationally attractive as it allows for inferences to be performed locally on a subset of vertices or a *clique*. However, the specification of hyper-Markov laws involves many hyperparameters and the constraints can be restrictive for modelling large graphs.

For complete decomposable graphs $\Omega > 0$, we can work with so-called *global priors* like Wishart that have positive definite realizations. Global priors are easily incorporated and are computationally attractive. For non-complete decomposable graphs Ω is constrained and prior elicitation is not straightforward. To this end, inference for non-decomposable graphs is typically complicated and has to rely heavily on numerical integration methods. Moreover, non-decomposable graphs are more important in theory than in practical applications.

We will work with the assumption that the graph G is *complete* with an edge between every pair of vertices. Then the precision matrix L is unconstrained and can be specified in a straightforward manner by a conjugate Wishart prior distribution. The Wishart distribution, $W_n(A, m)$ has $n \times n$ positive definite realizations when the degrees of freedom parameter $m > n$ and the average matriculate value is mA.

The rest of the work is organized as follows. In Section B, we review existing spectral techniques for graph cutting and the connection with kernel k-means clustering. In Section C, we describe the modelling of the precision matrix and the prior specification of parameters associated with the vertices by nonparametric Dirichlet process priors for clustering. The posterior inference of all the parameters is detailed in Section D. Therein, we also derive a loss function that has similarities with the normalized ratio loss function in graph theory. The empirical minimizaton of associated Bayes risk via MCMC is addressed in Section E. Finally in Section F, we discuss the simulations and the results from the analysis of various datasets.

B. Graph Cutting

Let us define an undirected graph $G = (V, E)$ where $V = 1, \ldots, n$ is a set of vertices and E is a set of undirected edges. Two vertices are *connected* by an edge if there is an association. Every element in E is assigned a non-zero weight as a measure of association.

Define the *adjacency matrix* A with entries $a_{i,j} > 0$, if there is an edge $(i, j) \in E$ and the *Laplacian* L that has entries,

$$l_{i,j} = \begin{cases} -a_{i,j} & (i,j) \in E, \\ w_i & i = j, \\ 0 & otherwise, \end{cases}$$

where the w_i is the degree of vertex i equal to $\sum_j a_{i,j}$ and $W = \text{diag}(w_1, \ldots, w_n)$. Note that Laplacian L, which can be written as $L = W - A$, has the following properties:

1. L is a symmetric positive semi-definite matrix. Thus all eigenvalues of L are real and non-negative, and L has a full set of n real and orthogonal eigenvectors.

2. Let $\mathbf{1}_n = (1, \ldots, 1)'$, then $L\mathbf{1}_n = 0$. Thus 0 is an eigenvalue of L with $\mathbf{1}_n$ as the eigenvector.

3. If the graph G has k connected components then L has k eigenvalues that are zero.

4. For any vector x, $x'Lx = \sum_{i,j \in E} a_{i,j}(x_i - x_j)^2$.

The last property follows from property 2 and is useful for defining loss functions for graph cutting. Note that Property 4 holds even if some weights $a_{i,j}$ are negative.

The matrices A or L are important because their spectral decomposition tells us a lot about the graph G. For example, the second-to-smallest eigenvalue of L, λ_2 is

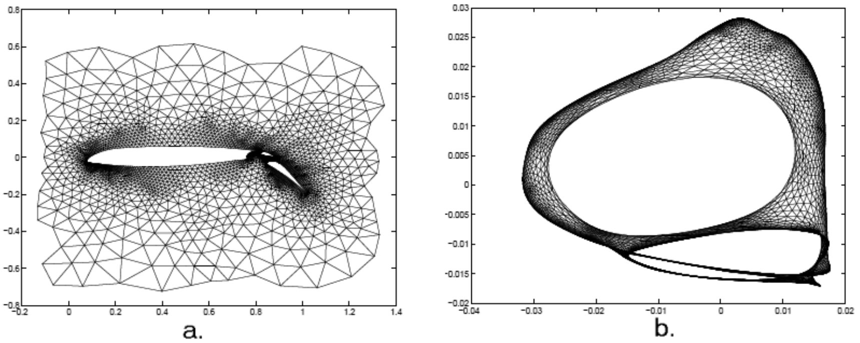

Fig. 26. Spectral decomposition of graphs. a. The popular airfoil graph, b. The graph drawn using the first two eigenvectors of the Laplacian of the graph as the x and y coordinates.

zero iff the graph G is disconnected (Fielder, 1975). The popular airfoil graph and the graph constructed from the first two eigenvectors of the Laplacian is shown in Figure 26. Roughly speaking, the dense areas in the new graph represent vertices that are highly connected with a significant interactions. These properties are used to derive graph cutting algorithms briefed below.

Graph cutting is a well-studied area in graph theory (Chung, 1997) and is commonly addressed as a constrained optimization problem where one tries to partition a graph into roughly equally sized pieces while minimizing the number of edges between those pieces. For this a simple cost function is given by the cut-ratio ϕ that is defined as follows. If S, S^c is a partition of V or the vertices, $\mathrm{vol}(S, S^c) = \sum_{i \in S, j \notin S^c} a_{i,j}$ is the sum of all edges between the two sets, and $\mathrm{vol}(S) = \sum_{i \in S} w_i$ then the cut-ratio $\phi(S) = \mathrm{vol}(S, S^c) / \min\{\mathrm{vol}(S), \mathrm{vol}(S^c)\}$. The cut of minimum ratio is given by the set S minimizing $\phi(S)$ and its quality is called the *conductance* of the graph,

$$\phi(G) = \min_{S \subset V} \phi(S). \tag{3.1}$$

An alternate way is to approximate the cut-ratio by minimizing

$$\phi_x = x'Lx/x'Wx \tag{3.2}$$

where L is the Laplacian and W is the diagonal degree matrix. Applying Property 4 of the Laplacian, it is clear that the ratio represents a trade-off between the sum of squares of the lengths of the edges in the numerator and a measure of vertex distance from the origin weighted by the degree of each vertex in the denominator. The weighted quadratic term in the denominator penalizes unbalanced partitions. It is interesting to note the resemblence with the loss function in Fisher's linear discriminant analysis.

The ratio ϕ_x is minimized by the solution to the generalized eigenvalue problem $Lx = \lambda Wx$, which are the eigenvectors of $W^{-1/2}LW^{-1/2}$. The smallest eigenvalue λ_1 is zero with the eigenvector $\mathbf{1}_n$. We are interested in the second smallest eigenvalue $\lambda_2 = \min_{x \perp 1_n} \phi_x$ as it provides a good approximation of the conductance of the graph (3.1). In particular, $\lambda_2 = \min_{x \perp 1_n} \phi_x$ can be bound (Lovascz, 1996) by Cheeger's Theorem,

$$\frac{\phi(G)^2}{8} \leq \lambda_2 \leq \phi(G).$$

One can also work with a *normalized Laplacian* $\mathcal{L} = W^{-1/2}LW^{-1/2}$ and the normalized ratio cut minimizes

$$\phi_x = x'\mathcal{L}x/x'x. \tag{3.3}$$

At this point there are two possible strategies to graph cutting. In a bi-partition or binary classification problem, ϕ_x is minimized by x constrained to lie in the space of binary indicator vectors depicting all possible class memberships of the n vertices. Then the solution vector provides the best binary classification of the vertices in the graph with respect to the above defined criterion (Chung, 1997). This procedure

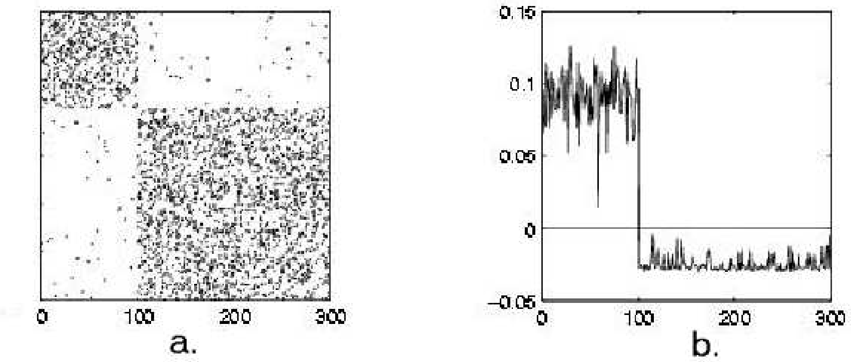

Fig. 27. Binary cutting of a graph. a. An adjacency matrix showing two dense clusters of vertices with sparse connections, b. The second eigenvector of the adjacency matrix clearly defines the boundary between the two classes.

can be extended to a multi-classification problem by seeking solutions of multi-class indicator vectors. To see why this is an effective way to cut a graph, we show an example adjacency matrix from a graph and its second eigenvector in Figure 27. The eigenvector clearly defines the boundary between the two dense clusters with sparse connections.

Alternatively, each row of the eigenvector(s) corresponding to λ_2 (and $\lambda_3, \lambda_4, \ldots$) are linked with the vertices and then clustered using a suitable algorithm (Dhillon, 2001).

1. The k-means connection

In a recent paper (Dhillon *et al.*, 2004) in maching learning, it was shown that normalized ratio-based graph cutting (3.3) is equivalent to a kernel k -means clustering. For particular choices of the kernel and by rewriting the cost function of the kernel k-means algorithm as a trace maximization problem it is possible to derive this equivalence.

We provide a shorter version of the proof given in Dhillon *et al.*, (2004). Consider a $p \times n$ matrix $\mathbf{Y} = (\mathbf{Y}_1, \ldots, \mathbf{Y}_n)$ of $p \times 1$ response vectors. Each data point or column of $\mathbf{Y}$ can be linked with a vertex in a graph. Assume a linear functional $\varphi : \mathbf{Y} \to \mathbf{Y}K$ for the kernel k-means algorithm that lineary maps each $1 \times n$ row of $\mathbf{Y}$. The corresponding kernel is given by $\varphi(\mathbf{Y})'\varphi(\mathbf{Y}) = K'\mathbf{Y}'\mathbf{Y}K$. Define the cluster configuration $\mathcal{C}_n$ with d_n set entries corresponding to the number clusters such that $\mathcal{C}_n(i)$ points to all indices in the ith cluster. Then the kernel k-means algorithm minimizes the loss function:

$$L(\mathcal{C}_n) = \sum_{i=1}^{s_n} \sum_{j \in \mathcal{C}_n(i)} \|\varphi(\mathbf{Y}_j) - \mu_i\|^2, \tag{3.4}$$

where $\mu_i \equiv \frac{1}{n_i} \sum_{j \in \mathcal{C}_n(i)} \varphi(\mathbf{Y}_i)$ is the center or location of the ith cluster which has $n_i = |\mathcal{C}_n(i)|$ entries. The equivalence with the normalized ratio cut (3.3) is proved next.

Theorem 2 *Let $\mathbf{Y}$ be a $p \times n$ matrix of observations. The objective loss function (3.4) for the kernel k-means algorithm with a kernel $K'\mathbf{Y}'\mathbf{Y}K$ is the equivalent to the normalized ratio cut loss function (3.3).*

Proof: Assume a configuration $\mathcal{C}_n$ having s_n clusters and define $\bar{\mathbf{Y}}_i = (\mathbf{Y}_j : j \in \mathcal{C}_n(i))$ as the $p \times n_i$ matrix of all n_i observables in the ith cluster.

The loss function (3.4) can be rewritten as

$$\begin{aligned} L(\mathcal{C}_n) &= \sum_{i=1}^{d_n} \sum_{j \in \mathcal{C}_n(i)} \|\varphi(\mathbf{Y}_j) - \varphi(\bar{\mathbf{Y}}_i)\frac{1}{n_i}\mathbf{1}_{n_i}\|^2 \\ &= \sum_{i=1}^{d_n} \mathrm{tr}\{(\varphi(\bar{\mathbf{Y}}_i) - \varphi(\bar{\mathbf{Y}}_i)\mathbf{J}_i)'(\varphi(\bar{\mathbf{Y}}_i) - \varphi(\bar{\mathbf{Y}}_i)\mathbf{J}_i)\} \\ &= \mathrm{tr}\{(\varphi(\mathbf{Y}) - \varphi(\mathbf{Y})\mathbf{J})'(\varphi(\mathbf{Y}) - \varphi(\mathbf{Y})\mathbf{J})\} \\ &= \mathrm{tr}\{\mathbf{Y}KK'\mathbf{Y}'\} - \mathrm{tr}\{\mathbf{J}K'\mathbf{Y}'\mathbf{Y}K\mathbf{J}\} \end{aligned}$$

where $\mathbf{J}_k = \frac{1}{n_k}\mathbf{1}_{n_k}\mathbf{1}_{n_k}'$, $\mathbf{J} = \mathrm{diag}(\mathbf{J}_1, \ldots, \mathbf{J}_{d_n})$ and the last step in the derivation follows from the idempotency of $\mathbf{J}$ and the invariance of trace to the order of multiplication, i.e. $\mathrm{tr}(AB) = \mathrm{tr}(BA)$.

In this form, the best clustering configuration is given by

$$\mathcal{C}_n = \mathrm{argmin}_{\mathcal{C}_n} L(\mathcal{C}_n) = \mathrm{argmax}_{\mathcal{C}_n} \mathrm{tr}\{\mathbf{J}K'\mathbf{Y}'\mathbf{Y}K\mathbf{J}\}. \tag{3.5}$$

This is equivalent to the normalized ratio cut (3.3) objective when the kernel $K'\mathbf{Y}'\mathbf{Y}K$ is set to $W^{-1/2}LW^{-1/2}$ mentioned in Section 1.1. ■

The additive loss function of the k-means algorithm is in fact the sum of the within-cluster variation over all the clusters in the kernel-transformed space. In the next section, we will develop a nonparametric Bayes model for clustering based on the Dirichlet Process (Ferguson, 1973). For a given clustering sampled from this model, we will show that a loss function similar to (3.5) arises naturally, when the likelihood is a multivariate normal distribution. The nonparametric modelling allows us to consider the marginal Bayes risk over all possible number of clustering combinations of different cluster sizes.

C. Graphical Partitioning Model

Let $\mathbf{Y} = (\mathbf{Y}_1, ..., \mathbf{Y}_n)$ be a set of $p \times 1$ observations vectors that follows a probability distribution P. We associate each vertex of the graph G with an observation in $\mathbf{Y}$ and assume that P is *Markov* over G. In particular, $P \in \mathcal{P}$ where $\mathcal{P}$ denotes the family of multivariate centered Gaussian distributions $N(0, \mathbf{I}_p, \Omega^{-1})$ with respect to G and Ω is the precision matrix. Due to the Markov property, a missing edge between any two vertices i and j is equivalent to setting $\omega_{i,j} = 0$ that in turn implies conditional independence, $\mathbf{Y}_i \perp \mathbf{Y}_j | \mathbf{Y}_{-(i,j)}$.

We want to determine the graphs that best explain the model uncertainty given a set of observations $\mathbf{Y}$. The likelihood of the Gaussian graphical model is then written as

$$\begin{aligned} \mathbf{Y}|G &\sim N(0, \mathbf{I}_p, \Omega^{-1}) \\ &= \frac{|\Omega|^{p/2}}{(2\pi)^{np/2}} \exp\left\{-\frac{1}{2}\mathrm{tr}\left(\mathbf{Y}'\mathbf{Y}\Omega\right)\right\} \end{aligned} \tag{3.6}$$

Gaussian graphical models are similar to the covariance selection models (Wong *et al.*, 2003) where the objective is to try and identify the zero elements in the precision matrix. Zeroes in the precision matrix imply a non-complete graph for which Ω is not necessary positive definite. For the problem of graph cutting, however, that is approached as a clustering problem it suffices to use assume a complete graph for which Ω is easily specified by a prior distribution on the space of all $n \times n$ positive definite matrices.

The precision matrix can be associated with a complete undirected graph by writing down the implied conditional distributions from the joint Gaussian distribution using standard distribution theory. We have

$$\mathbf{Y}_i|\mathbf{Y}_{-i} \sim N\left(-\omega_{i,i}^{-1}\sum_{j\neq i}\omega_{i,j}\mathbf{Y}_j, \omega_{i,i}^{-1}\right). \tag{3.7}$$

Switching the sign the off-diagonal terms in equation (3.7) can be identified as the edge weights $a_{i,j}$ of an undirected graph.

In general, the set of conditional distributions $\{\mathbf{Y}_i|\mathbf{Y}_{-i}\}, \forall i$ does not guarantee a joint distribution except in the special case of autoregressive models. The conditional autoregressive or CAR model (Besag, 1974) fits a high-dimensional distribution in sets of univariate conditional distributions depicting patterns of association in a set of variables. In other words, there is a consistency between the local and the global properties of the graph that is useful for practical interpretation. However, this is accomplished at the cost of a positive semi-definite Ω and a degenerate joint distribution, which poses significant problems for flexible hierarchical Bayesian modelling.

A well-defined multivariate normal model, such as (3.14) can always be augmented (with an extra missing variable) into a CAR model. More specifically, we can work with a linear combination of $\mathbf{Y}$ that has a precision matrix $\Omega^* = \mathbf{U}\Omega\mathbf{U}'$ where $\mathbf{U} = [\mathbf{I}_n| - \mathbf{1}_n]$. If $\mathbf{U}^+$ satisfies $\mathbf{U}^+\mathbf{U} = \mathbf{I}_n$, then the linear combination is given by $\mathbf{Y}^* = \mathbf{Y}\mathbf{U}^{+\prime}$ and has a distribution proportional to

$$\exp\left(-\frac{1}{2\sigma^2}\mathrm{tr}(\mathbf{Y}^{*\prime}\mathbf{Y}^*\Omega^*)\right). \tag{3.8}$$

Note that as $n \to \infty$, the linear combination $\mathbf{Y}^* = \mathbf{Y}\mathbf{U}^{+\prime} \to (\mathbf{Y}, \mathbf{Y}_{n+1})$ at the rate of $O(n^{-1})$. Thus for large values of n, we can use this distribution to graphically model an augmented collection of responses $(\mathbf{Y}, \mathbf{Y}_{n+1})$. As before the joint distribution can be factored into conditional regressions of the form (3.7) with the contribution of an extra response $\mathbf{Y}_{n+1}$ or an extra vertex in the graph which has little physical interpretation.

D. Prior Elicitation

We define a $m \times n$ matrix of random variables $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_n)$ following the distribution, $N(\boldsymbol{\theta}, \sigma^2\mathbf{I}_m, \mathbf{I}_n)$, where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$ is a matrix of cluster locations. For example, if $\beta_1, \beta_3, \beta_4$ belong to the same cluster then $\theta_1 = \theta_3 = \theta_4$. When $m \geq n$ this implies a noncentral Wishart distribution for $\Omega = \boldsymbol{\beta}'\boldsymbol{\beta}$,

$$\begin{aligned}\Omega &\sim W_n'(\sigma^2\mathbf{I}_n, \Psi, m), m \geq n && (3.9)\\ &= \frac{|\Omega|^{(m-n-1)/2}}{(2\sigma^2)^{mp/2}\Gamma(m/2)}\mathrm{etr}\left\{-\frac{1}{2\sigma^2}\Omega - \frac{1}{2}\Psi\right\} F_0^1\left(\frac{\nu}{2}, \frac{1}{4\sigma^2}\Psi\Omega\right) && (3.10)\end{aligned}$$

where $\mathrm{etr}(\cdot) = \exp(\mathrm{tr}(\cdot))$, F_0^1 is the Bessel function, $\Psi = \sigma^{-2}\boldsymbol{\theta}'\boldsymbol{\theta}$ is the non-centrality parameter and m is the degrees of freedom. We will show later that this specification allows us to approach graph cutting as a clustering problem.

In the special case, when the cluster locations $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$ are known and when m is large, we can approimate implied non-central Wishart distribution for Ω by its central counterpart. More precisely, the distribution (3.10) can be approximated upto the order $O(m^{-2})$ by the central Wishart distribution (Gupta and Nagar, 1999),

$$W_n\left(\sigma^2\mathbf{I}_n + \frac{1}{m}\boldsymbol{\theta}'\boldsymbol{\theta}, m\right), \tag{3.11}$$

which is conjugate to the normal likelihood and can be used for fitting Ω directly.

When unknown, the location parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$ are specified by Dirichlet process priors for clustering. The Dirichlet Process (DP) is a non-parametric two-parameter conjugate family in the sense that there is a positive probability that a sample distribution will approximate arbitrarily well any distribution that is dominated by the base distribution H_ϕ. DPs are also a.s. discrete and comprise a certain partitioning of the parameter space. These properties allow us to model clustering configurations of a set of variables by DP priors without fixing the number of clusters beforehand. The reader is referred to Section C of Chapter I for more details about Dirichlet process priors and their applications in clustering.

In a sequence of draws $\theta_1, \theta_2, \ldots$ from the Polya urn representation of the Dirichlet process (Blackwell and MacQueen, 1973), the nth sample is either distinct with a small probability $\alpha/(\alpha + n - 1)$ or is tied to previous sample with positive probability to form a cluster. Let $\boldsymbol{\theta}_{-n} = \{\theta_1, \ldots, \theta_n\} - \{\theta_n\}$ and d_{n-1}= number of preexisting clusters of tied samples in θ_{-n} at the nth draw, then we have

$$f(\theta_n|\boldsymbol{\theta}_{-n}, \alpha, \phi) = \frac{\alpha}{\alpha + n - 1}H_\phi + \sum_{j=1}^{d_{n-1}} \frac{n_i}{\alpha + n - 1}\delta_{\bar{\theta}_j} \tag{3.12}$$

where H_ϕ is the base prior, and the jth cluster has n_j tied samples that are commonly expressed by $\bar{\theta}_j$ subject to $\sum_{j=1}^{d_{n-1}} n_j = n-1$. After n sequential draws from the Polya urn, there are several ties in the sampled values and we denote the set of distinct samples (or in this case, the cluster centers) by $\{\bar{\theta}_1, \ldots, \bar{\theta}_{d_n}\}$, where d_n is essentially the number of clusters.

Assuming an inverse gamma prior distribution is assumed for the variance parameter, $\sigma^2 \sim IG(u, v)$. The developments above can be summarized in the Bayesian hierarchical model,

$$\begin{aligned}
\mathbf{Y}|\Omega &\sim N(0, \mathbf{I}_p, \Omega^{-1}), && (3.13)\\
\Omega|\boldsymbol{\theta}, \sigma^2 &\sim W(\sigma^2\mathbf{I}_n + \boldsymbol{\theta}'\boldsymbol{\theta}/m, m), && (3.14)\\
&\approx W'(\sigma^2\mathbf{I}_n, \sigma^{-2}\boldsymbol{\theta}'\boldsymbol{\theta}, m) && (3.15)\\
\Omega &= \boldsymbol{\beta}'\boldsymbol{\beta}\\
\cdots &\cdots \cdots
\end{aligned}$$

$$\begin{aligned}
\boldsymbol{\beta}|\boldsymbol{\theta}, \sigma^2 &\sim N(\boldsymbol{\theta}, \sigma^2\mathbf{I}_m, \mathbf{I}_n) && (3.16)\\
\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_n)|\sigma^2, g &\sim F \sim D(\alpha, N(0, g\sigma^2\mathbf{I}_m)), && (3.17)\\
\sigma^2 &\sim IG(u, v), && (3.18)\\
g &\sim IG(r, s), && (3.19)\\
\alpha &\sim Gamma(d_0, \eta_0), && (3.20)
\end{aligned}$$

where $\boldsymbol{\phi} = g$.

E. Posterior Inference

The posterior conditional distributions of all the parameters are in exact form due to the conjugate prior specifications. The posterior inference of the precision parameter α remains the same as Section D.3 in Chapter I and is not discussed here.

1. Conditional distributions for the precision matrix

When the cluster locations are known, the posterior distribution of Ω is calculated by combining the approximate Wishart prior distribution $W(\sigma^2\mathbf{I}_n + \boldsymbol{\theta}'\boldsymbol{\theta}/m, m)$ with the likelihood (3.14). This gives

$$\Omega|\mathbf{Y}, \boldsymbol{\theta}, \sigma^2 \sim W\left(\mathrm{tr}(\mathbf{Y}'\mathbf{Y}) + \sigma^2\mathbf{I}_n + \boldsymbol{\theta}'\boldsymbol{\theta}/m, m+p\right). \tag{3.21}$$

However, when the cluster locations are unknown, we sample first sample $\boldsymbol{\beta}$ from its posterior normal distribution and then calculate Ω as $\boldsymbol{\beta}'\boldsymbol{\beta}$. This helps us to avoid problems of uniqueness in calculating $\boldsymbol{\beta}$ from Ω. The posterior distribution for $\boldsymbol{\beta}$ with respect to the multivariate normal prior distribution $N(\boldsymbol{\theta}, \sigma^2\mathbf{I}_m, \mathbf{I}_n)$. The posterior distribution is given by

$$\boldsymbol{\beta}|\mathbf{Y}, \boldsymbol{\theta}, \sigma^2 \sim N\left(m(m\hat{\Sigma} + \sigma^{-2}\mathbf{I}_n)^{-1}\hat{\Sigma}\boldsymbol{\theta}, (m\hat{\Sigma} + \sigma^{-2}\mathbf{I}_n)^{-1}, \mathbf{I}_m\right) \tag{3.22}$$

with $\hat{\Sigma} = \mathbf{Y}'\mathbf{Y}/m$. Note that the posterior distributional form is exact and does not involve any approximations.

2. Conditional distributions for clustering

To update the unknown cluster locations θ_i, we can picture (3.17) as the likelihood function for $\boldsymbol{\theta}$ and calculate the posterior distribution with respect to the conjugate

Polya urn prior (3.12). We get a similar looking posterior,

$$\theta_i|\sigma^2, \boldsymbol{\theta}_{-i}, \boldsymbol{\beta}, g \propto q_{i,0} H^*_{\phi,i} + \sum_{j \neq i}^{d_i - 1} q_{i,j} \delta_{\bar{\theta}_j} \tag{3.23}$$

where

$$H^*_{\phi,i} = N(\theta^*_i, g^*_i \sigma^2 \mathbf{I}_m) \text{ with} \tag{3.24}$$

$$g^*_i = (g^{-1} + \sigma^{-2})^{-1} \text{ and } \theta^*_i = g^*_i \beta_i. \tag{3.25}$$

The weights $q_{i,\cdot}$ follow from the likelihood and marginals in the Normal-Inverse Gamma family (Escobar and West, 1995) and determine the posterior inclination towards new distinct samples. We have $q_{i,j} \propto n_j \phi(\beta - \bar{\theta}_j, \sigma^2 \mathbf{I}_m)$ as the conditional distribution of θ_i given the jth cluster parameters and $q_{i,0} \propto \alpha \phi\{\beta_i, \sigma^2(1+g)\mathbf{I}_m\}$ is the marginal distribution.

3. Posterior sampling of the scaling parameter

The scaling parameters g determines the posterior shrinkage or smoothing of the location parameters θ_i over all the vertices of the graph. Small values of g are associated with higher smoothing and lower number of clusters in the model.

To obtain the posterior distribution of the scaling parameter g, we use the conditional independence of the cluster locations. The inverse Gamma prior for the scaling parameters, $g \sim IG(r, s)$, when combined with joint prior distribution of the distinct cluster locations gives us

$$g|d_n, \sigma^2, \{\bar{\theta}_j\}_{j=1}^{d_n} \sim IG(r^*, s^*), \tag{3.26}$$

with $r^* = d_n m + r$ and $s^* = \sum_{j=1}^{d_n} ||\bar{\theta}_j||^2 + s$.

4. Posterior sampling of the dispersion parameter

The posterior distribution of σ^2 given the number of clusters is calculated by exploiting the conditional independence of the distinct cluster locations (Korwar and Hollander, 1973), that is $\bar{\theta}_j|\mathcal{C}_n \sim_{iid} N(0, g\sigma^2\mathbf{I}_m)$. Then, the posterior distribution is

$$\sigma^2|\boldsymbol{\beta}, g, \mathcal{C}_n \sim \text{IG}\left\{u + \sum_{i=1}^{d_n}\left(\sum_{j\in\mathcal{C}_n(i)} \beta_j'\beta_j - \mu_i^{*\prime}(G_i^*)^{-1}\mu_i^*\right), v + mn\right\}$$
$$\text{where } \mu_i^* = G_i^* \sum_{j\in\mathcal{C}_n(i)} \beta_j \text{ and } G_i^* = (g^{-1} + n_i)^{-1}\mathbf{I}_m.$$

It is interesting to note the resemblence of

$$E(\sigma^2|\boldsymbol{\beta}, g, \mathcal{C}_n) = \frac{u + \sum_{i=1}^{d_n}\left(\sum_{j\in\mathcal{C}_n(i)} \beta_j'\beta_j - \mu_i^{*\prime}(G_i^*)^{-1}\mu_i^*\right)}{v + mn - 2}$$

with the k-means cost function (3.5). This suggests that the posterior mean of σ^2 can be used as a loss function for graph partitioning. In fact, given the number of clusters and the ordering of vertices, the minimization of the posterior mean of σ^2 in the present model has some similarities with the minimization of the normalized ratio cut loss function (3.3) of an undirected graph that has an adjacency matrix, $A = -\omega_{i,j}I(i \neq j)$ and weight matrix, $W = \text{diag}(\omega_{i+}^{-1})$.

To see this, first note that by conditional independence, we have $\bar{\theta}_j|\mathcal{C}_n \sim_{iid} N(0, g\sigma^2\mathbf{I}_m)$. Define a $p \times d_n$ matrix $\bar{\boldsymbol{\theta}}$ that contains only the distinct elements in $\boldsymbol{\theta}$ and a $d_n \times n$ indicator matrix $\mathbf{T}_n$ satisfying $\bar{\boldsymbol{\theta}}\mathbf{T}_n = \boldsymbol{\theta}$ so that $\mathbf{T}_n$ is essentially a function of the present clustering configuration $\mathcal{C}_n$. Then the likelihood (3.17) can be rewritten as

$$\begin{aligned}\boldsymbol{\beta}|\bar{\boldsymbol{\theta}}, \sigma^2 &\sim N(\bar{\boldsymbol{\theta}}\mathbf{T}_n, \sigma^2\mathbf{I}_m, \mathbf{I}_n), \text{ or}\\ \boldsymbol{\beta}|\sigma^2 &\sim N(0, \sigma^2\mathbf{I}_m, (\mathbf{I}_n + g\mathbf{T}_n'\mathbf{T}_n)),\end{aligned}$$

where the last step follows by averaging out $\bar{\boldsymbol{\theta}}$. Combining this marginal distribution with the inverse gamma prior (3.19) for σ^2 gives the posterior distribution

$$\sigma^2|\boldsymbol{\beta}, g, \mathcal{C}_n \sim IG(u + \mathrm{tr}(\Omega(\mathbf{I}_n + g\mathbf{T}_n'\mathbf{T}_n)^{-1}), v + mn), \tag{3.27}$$

where in the last step Ω is used instead of $\boldsymbol{\beta}'\boldsymbol{\beta}$.

F. Loss Functions for Graph Partitioning

Define the matrix $\mathbf{x} = (\mathbf{I}_n + g\mathbf{T}_n'\mathbf{T}_n)^{-1/2}$ that is a function of present clustering configuration $\mathcal{C}_n$. Identifying Ω as the kernel, it is clear that the posterior mean of σ^2,

$$E(\sigma^2|\boldsymbol{\beta}, g, \mathcal{C}_n) \propto \mathrm{tr}\,(\mathbf{x}'\Omega\mathbf{x}) \tag{3.28}$$

has a quadratic form similar to the loss functions (3.5) or (3.3). The only problem is that now the kernel $\Omega > 0$ and does not satisfy $\Omega\mathbf{1} \neq 0$.

Like before we define an adjacency matrix A such that $a_{i,j} = -\omega_{i,j}I(i \neq j)$, then for any real symmetric Ω, we have the identity

$$\mathrm{tr}\,(\mathbf{x}'\Omega\mathbf{x}) = \sum_i \omega_{i+}||x_i||^2 + \sum_{i<j} a_{i,j}||x_i - x_j||^2, \tag{3.29}$$

with $\omega_{i+} = \sum_j \omega_{i,j}$. The quadratic form is minimized by simultaneously minimizing $\sum_{i<j} a_{i,j}||x_i - x_j||^2$ and $\sum_i \omega_{i+}||x_i||^2$. Equivalently, this represents the minimization of the normalized ratio cut of a graph that has A as the adjacency matrix and $W = \mathrm{diag}(\omega_{i+}^{-1})$ as the weight matrix which has the inverse of the total variation associated with any node as its diagonal elements. Regarding the inverse variation as a measure of information, the minimization of the loss function (3.29) supports equal sized partitions with respect to total information with each subgraph.

There is another important difference from the graphs used in spectral graph theory. There the edge weights $a_{i,j}$ are always positive and this is an important assumption underlying graph cutting techniques based on the minimization of the cut-ratio. The idea is that smaller cut-ratios are achieved when the vertices sharing strong edges are assigned a common class label and vertices sharing weak edges are assigned different class labels. In our model, the edge weights can be negative and smaller cut-ratios may not pertain to any kind of graph cutting. However with minor modifications the minimization of (3.28) can still lead to meaningful graph cuts as explained below.

1. Interaction graph cuts

In a Bayesian approach, we generate MCMC samples of all the model parameters that can be combined later to define new loss functions. We want to find sub-graphs that contain highly interacting vertices irrespective of the sign of the correlation between those points. A typical application is in gene pathway analysis where the objective is to classifiy a set of genes as highly interacting even if they are negatively correlated.

We define a new loss function by assigning a negative sign to all off-diagonal terms of Ω in (3.28). The associated posterior risk can be empirically minimized to estimate the optimal graph cut.

2. Estimation of the graph cut

We define the optimal graph cut to minimize the posterior expectation of the quadratic loss function described above with respect to the unknown paramters in the model. The posterior risk can be evaluated by simple Monte Carlo averaging of the loss function with respect to g and Ω. Alternatively, we can first estimate Ω and calculate the conditional posterior risk averaging with respect to g.

Ideally, the minimization of the posterior risk would entail the exploration of the space of all possible clustering configurations. However, even for moderately large values of n this can be a computationally challenging task and we resort to a split-merge risk minimization approach like Quintana and Iglesias (2003) which evaluates the risk by sequentially splitting and merging different partitions in the direction that minimizes it.

Algorithm for Risk Minimization

Let $S \subset \{1, \ldots, n\}$. For any $i \in S$, let $\text{vol}(i, S) = \sum_{j \in S - \{i\}} a_{i,j}$ be the sum of all edge wts for all edges to vertex i.

1. Set $j = 1$, $\mathcal{C}_n^1(1) = \{1, \ldots, n\}$, $\mathcal{C}_n^1 = \{\mathcal{C}_n^1(1)\}$ & evaluate $\ell(\mathcal{C}_n^1)$.

2. Set $j = 2$. Find $k_1^* = \arg\min_{i \in \mathcal{C}_n^1(1)} \text{vol}(i, \mathcal{C}_n^1(1))$.

3. Create new partition, $\mathcal{C}_n^2 = \{\mathcal{C}_n^1(1) - \{k_1\}, \{k_1\}\}$.

4. If $\ell(\mathcal{C}_n^2) > \ell(\mathcal{C}_n^1)$ set $\mathcal{C}_n^* = \mathcal{C}_n^1$ and Stop. Else, continue to next step.

5. Set $j = j + 1$. Find $k_{j-1}^* = \arg\min_{i \in \mathcal{C}_n^{j-1}(1)} \text{vol}(i, \mathcal{C}_n^{j-1}(1))$.

6. Let $\mathcal{C}_n^{j,1} = \{\mathcal{C}_n^{j-1}(1) - \{k_{j-1}^*\}, \ldots, \mathcal{C}_n^{j-1}(|\mathcal{C}_n^{j-1} - 1|), \{k_{j-1}^*\}\}$, and for $i = 2, \ldots, |\mathcal{C}_n^{j-1}|$, create partitions $\mathcal{C}_n^{j,i} = \{\mathcal{C}_n^{j-1}(1) - \{k_{j-1}^*\}, \ldots, \mathcal{C}_n^{j-1}(i) \cup \{k_{j-1}^*\}, \ldots, \mathcal{C}_n^{j-1}(|\mathcal{C}_n^{j-1} - 1|)\}$.

7. Set $\mathcal{C}_n^j = \arg\min_{\mathcal{C}_n^{j,i}} \ell(\mathcal{C}_n^{j,i})$.

8. If $\ell(\mathcal{C}_n^j) > \ell(\mathcal{C}_n^{j-1})$ set $\mathcal{C}_n^* = \mathcal{C}_n^{j-1}$ and Stop. Else, goto Step 5.

Note that the optimal cut may not be unique and there might be more than one set of clusters that would minimize the posterior risk. Since the preceding algorithm is greedy, we run the algorithm several times by randomly permuting the vertex set. This allows us to uncover multiple solutions and compare between them using some model choice criterion.

G. Examples

The proposed methodology is applied to the problem of identifying significant subgraphs given a large graph of interactions for a collection of genes. We use two microarray datasets, namely a Melanoma dataset and a Gastrointestinal tumor dataset. More information about significant subnetworks is available for the first dataset and it is therefore used to check the performance of the proposed methodology. Following Chapter I, all the reported results were averaged over 100 simulations with 10,000 iterations per simulation and a burn-in period of 1000.

1. Melanoma dataset

This dataset consists of gene expression profiles for roughly 7000 genes from 31 melanoma samples. Kim *et al.*, (2002) used stochastic modelling to study the relationships between genes and found WNT5A gene to play a central role in a signal transduction pathway that might trigger an invasive phenotype in melanoma cells. First, 51 well predicted or good predictor genes were selected. Finally, 10 genes (shown in Table VII) that are known to play a significant role in the WNT5A driven pathway, were analyzed.

For our study we use all 51 genes. We get two optimal interaction graph cuts with four and five subgraphs (shown in the tables VIII and IX). In the first model, eight out of ten genes shown in Table VII are part of the same subgraph. This group of eight genes are split up into two subgraphs in the five subgraph model.

For visualization, we project the estimates of $\boldsymbol{\beta}$ by principal component analysis onto a lower dimensional space. Figures 28 and 29 plot the subgraphs along the first two principal components of estimated $\boldsymbol{\beta}$.

Table VII. Predictors for 10 genes likely to be a part of the WNT5A driven pathway.

Target	Predictor 1	Predictor 2	Predictor 3
Pirin	WNT5A	STC2	HADHB
WNT5A	pirin	S100P	RET-1
S100P	WNT5A	RET-1	Synuclein
RET-1	Pirin	WNT5A	S100P
MMP-3	S100P	RET-1	HADHB
PHO-C	MART-1	Synuclein	STC2
MART-1	Pirin	WNT5A	MMP-3
HADHB	Pirin	WNT5A	MMP-3
Synuclein	Pirin	S100P	MART-1
STC2	Pirin	WNT5A	PHO-C

Table VIII. Model with four subgraphs for melanoma data.

Subgraph	Genes
1	PHO-C, RET-1, ...
2	WNT5A, STC2, MMP3,Pirin, S100P, Synuclein, MART-1, HADHB,...
3	....
4	....

Table IX. Model with five subgraphs for melanoma data.

Subgraph	Genes
1	PHO-C, RET-1, ...
2	WNT5A, STC2, MMP3, ...
3	Pirin, S100P, Synuclein, MART-1, HADHB,...
4	....
5	....

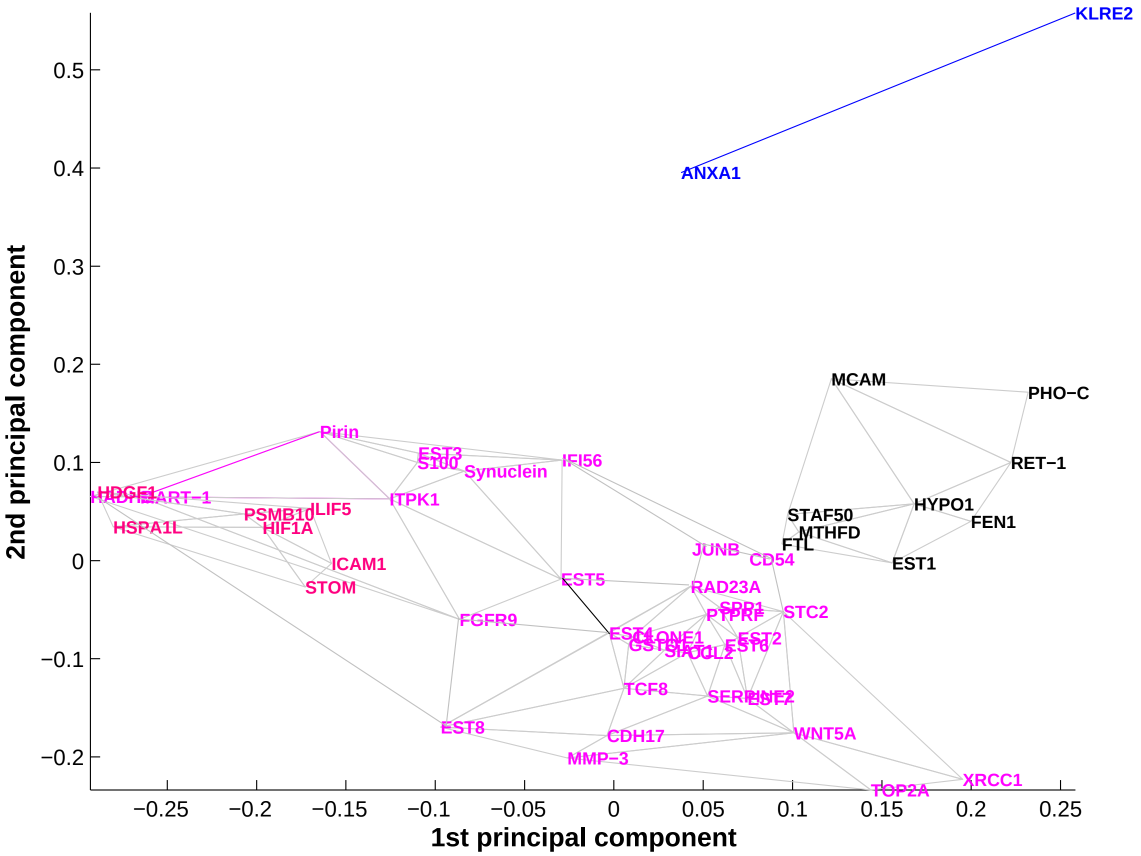

Fig. 28. Plot of four subgraphs for melanoma data along first two principal components of estimated beta.

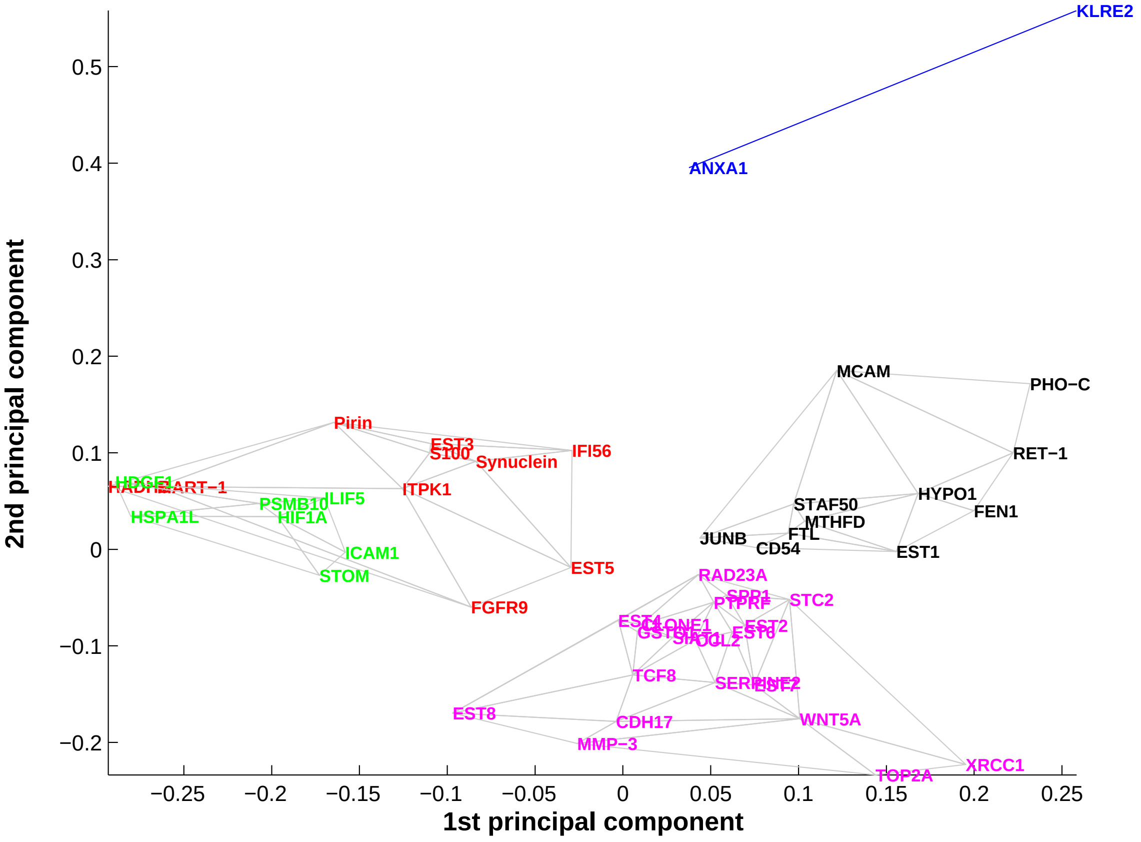

Fig. 29. Plot of five subgraphs for melanoma data along first two principal components of estimated beta.

2. Gastrointestinal cancer dataset

Nykter *et al.*, (2006) recently analyze a microarray data to compare the gene expressions in two types of tumors namely, gastrointestinal stromal tumors (GIST) and leiomyosarcomas (LMS) of the gastrointestinal tract. In the past, both have been classified as gastrointestinal sarcomas but new evidence shows that GIST has an unique phenotype characterized by the gain of function mutations in the cKIT gene compared to what is observed in LMS.

Microarray experiments were performed on 44 tissues for GIST and 32 tissues for LMS with the aim of finding differentially expressed genes between them. For this Nykter *et al.*, (2006) perform an unsupervised clustering the genes by visualizing them into multiple dimensions.

For our analysis, the dimension of the dataset was reduced to 84 genes with significant expression levels in both types of cancers. Following the estimation steps mentioned above, the estimated number of clusters for the cancers under both loss functions varies from 4 to 6.

To stress the relevance of a graph partitioning as compared to a simple clustering algorithm we plot the graphs data in two ways. Figures 30 and 31 plots the subgraphs generated by using the first two principal components of estimate $\boldsymbol{\beta}$ for GIST and LMS respectively. In all these figures, the colors indicate different clusters or subgraphs which were estimated using interaction loss function. It is evident that clustering interactions can be very different from clustering based on simple distance metrics.

The four estimated subgraphs for GIST and LMS are shown in tables X and XI. There are noticeable differences in the clustered genes between the two types of cancer and a comparison of the subgraphs can be used for differential analysis of genes.

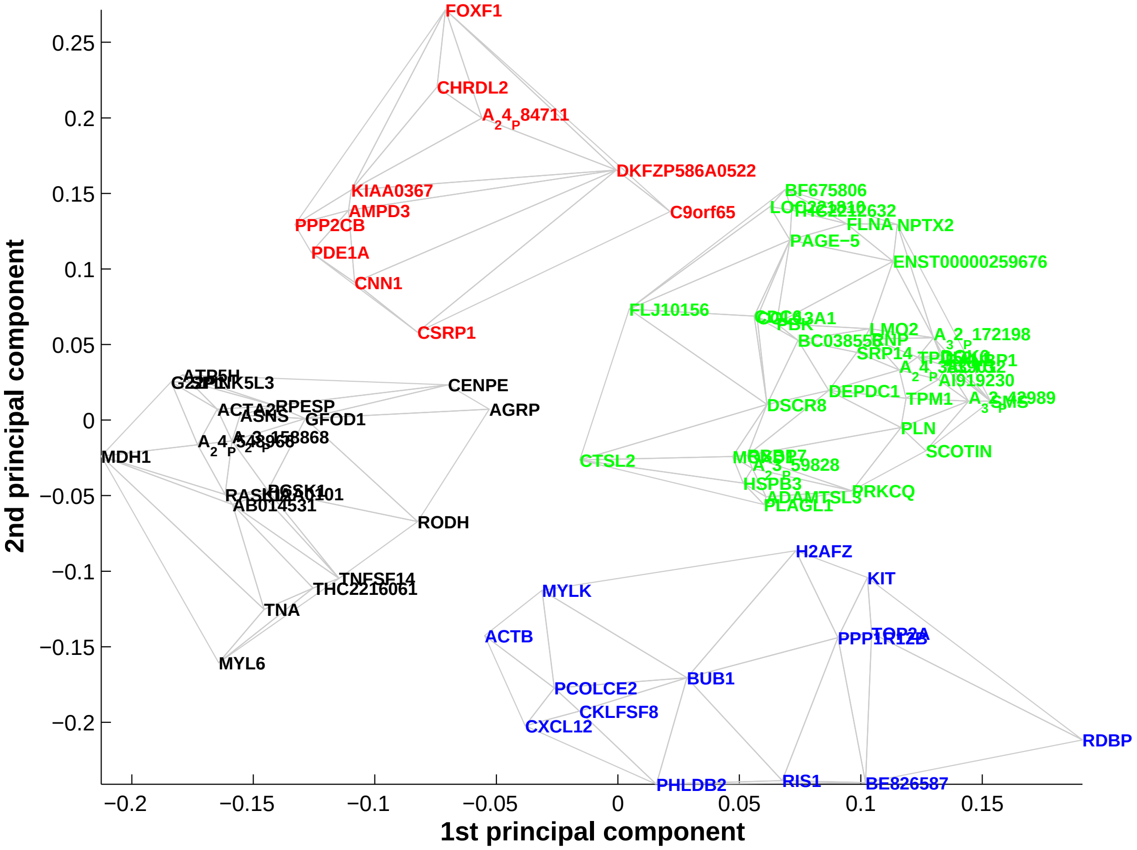

Fig. 30. Plot of four subgraphs for GIST data along first two principal components of estimated beta.

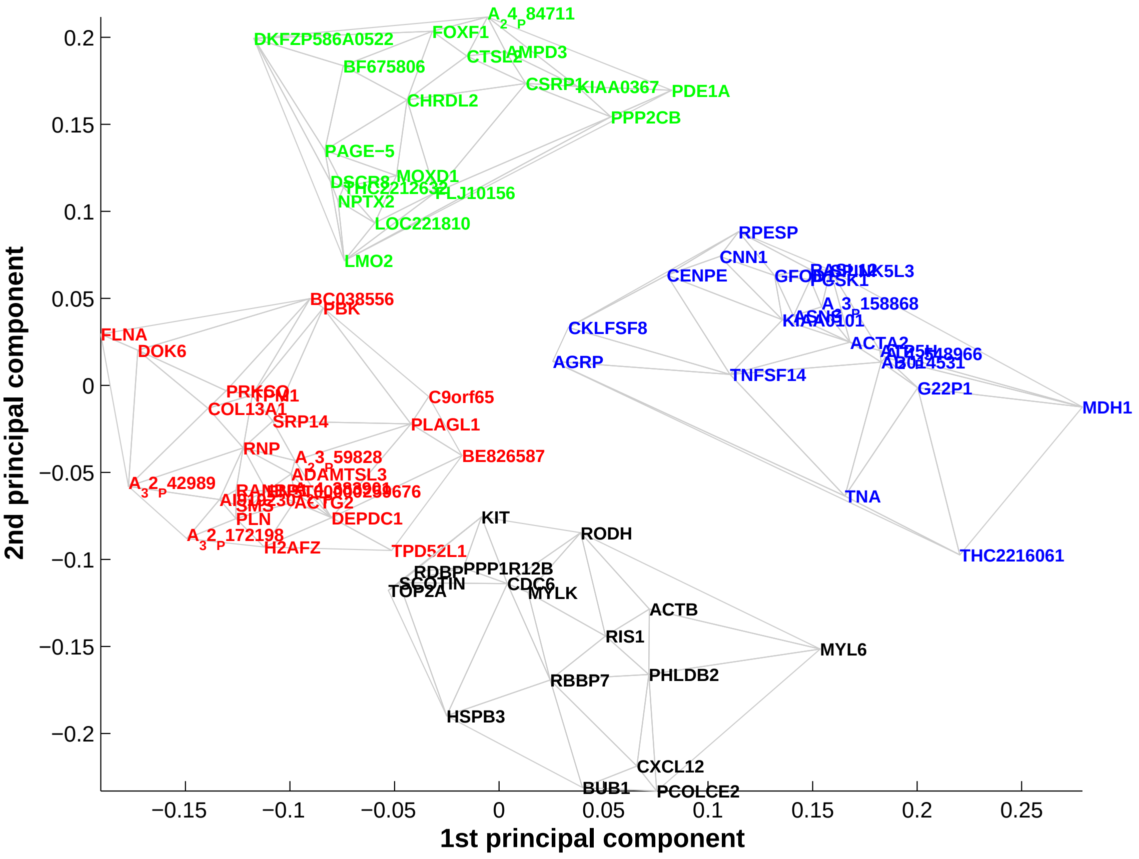

Fig. 31. Plot of four subgraphs for LMS data along first two principal components of estimated beta.

Table X. Four estimated clusters for GIST.

Clusters	Gene Names
1	AMPD3, A24P84711, CHRDL2, CNN1, CSRP1, DKFZP586A0522, FOXF1, KIAA0367, PDE1A, PPP2CB
2	ACTG2, AI919230, A23P59828, A24P383901, A32P172198, A32P42989, BC038556, BF675806, C9orf65, CDC6, COL13A1, DEPDC1, DOK6, DSCR8, ENST00000259676, FLJ10156, FLNA, LMO2, LOC221810, MOXD1, NPTX2, PAGE-5, PBK, PLN, PRKCQ, RANBP1, RBBP7, RNP, SCOTIN, SMS, SRP14, THC2212632, TPD52L1, TPM1
3	ACTB, ADAMTSL3, BE826587, BUB1, CKLFSF8, CXCL12, H2AFZ, HSPB3, **KIT**, MYLK, PCOLCE2, PHLDB2, PLAGL1, PPP1R12B, RDBP, RIS1, TOP2A
4	AB014531, ACTA2, AGRP, ASNS, ATP5H, A23P158868, A24P548966, CENPE, CTSL2, G22P1, GFOD1, KIAA0101, MDH1, MYL6, PCSK1, RASL12, RODH, RPESP, SPINK5L3, THC2216061, TNA, TNFSF14

Table XI. Four estimated clusters for LMS.

Clusters	Gene Names
1	AB014531, ACTA2, AGRP, ASNS, ATP5H, A23P158868, A24P548966, CENPE, CKLFSF8, CNN1, G22P1, GFOD1, KIAA0101, MDH1, PCSK1, PDE1A, PPP2CB, RASL12, RPESP, SPINK5L3, THC2216061, TNA, TNFSF14
2	AMPD3, A24P84711, BC038556, BF675806, CHRDL2, CSRP1, CTSL2, DKFZP586A0522, DSCR8, FLJ10156, FOXF1, KIAA0367, LMO2, LOC221810, MOXD1, NPTX2, PAGE-5, THC2212632
3	ACTG2, ADAMTSL3, AI919230, A23P59828, A24P383901, A32P172198, A32P42989, BE826587, C9orf65, COL13A1, DEPDC1, DOK6, ENST00000259676, FLNA, H2AFZ, PBK, PLAGL1, PLN, PRKCQ, RANBP1, RNP, SCOTIN, SMS, SRP14, TOP2A, TPM1
4	ACTB, BUB1, CDC6, CXCL12, HSPB3, **KIT**, MYL6, MYLK, PCOLCE2, PHLDB2, PPP1R12B, RBBP7, RDBP, RIS1, RODH, TPD52L1

REFERENCES

Abramovich, F., Sapatinas, T. and Silverman, B.W. (1998) Wavelet thresholding via a Bayesian approach. *J. R. Stat. Soc.* B, **60**, 725-749.

Antoniak, C.E. (1974) Mixtures of Dirichlet processes with applications to non-parametric problems. *Ann. Statist.*, **2**, 1152-1174.

Banerjee, S. (2004) On geodetic distance computations in spatial modelling. *Biometrics*, **61**, 617-625

Banfield, J.D. and Raftery, A.E. (1993) Model-based Gaussian and non-Gaussian clustering. *Biometrics*, **49**, 803-821.

Basu, S. and Chib, S. (2003) Marginal likelihood and Bayes factors for Dirichlet process mixture models. *J. Am. Stat. Assoc.*, **98**, 224-235.

Besag, J. (1974) Spatial interaction and the statistical analysis of lattice systems (with discussion). *J. Roy. Statist. Soc., Series B,* **36**, 192-236.

Bittner M., Meltzer P., Chen Y., Jiang Y., Seftor E. Hendrix M., Radmacher M., Simon R., Yakhini Z., Ben-Dor A., Sampas N., Dougherty E., Wang E., Marincola F., Gooden C., Lueders J., Glatfelter A., Pollock P., Carpten J., Gillanders E., Leja D., Dietrich K., Beaudry C., Berens M, Alberts D., Sondak V., Hayward N. and Trent J. (2000) Molecular classification of cutaneous malignant melanoma by gene expression profiling. *Nature*, **406**, 536-540.

Blackwell, D. and MacQueen, J.B. (1973) Ferguson distributions via polya urn schemes. *Ann. Statist.*, **1**, 353-355.

Bush, C.S. and MacEachern, S.N. (1996) A semi-parametric Bayesian model for randomized block designs. *Biometrika*, **83**, 221-227.

Chaloner, K. and Brant, R. (1988) A Bayesian approach to outlier detection and residual analysis. *Biometrika*, **75**, 651-660.

Chung, F.R.K. (1997) *Spectral Graph Theory.* CBMS Lecture Notes, Providence RI: AMS Publication.

Clyde, M. and George, E.I. (2000) Flexible empirical Bayes estimation for wavelets. *J. R. Stat. Soc.* B, **62**, 681-698.

Clyde, M., Parmigiani, G., and Vidakovic, B. (1998) Multiple shrinkage and subset selection in wavelets. *Biometrika*, **85**, 391-402.

Daubechies, I. (1992) *Ten Lectures in Wavelets*. Philadelphia: SIAM.

Daubechies, I. and Sweldens, W. (1996) Factoring wavelet transforms into lifting steps. *J. Fourier Anal. Appl.*, **4**, 245 - 267.

Dawid, A.P. and Lauritzen, S.L. (1993) Hyper Markov laws in the statistical analysis of decomposable graphical models. *The Annals of Statistics*, **21**, 1272-1317.

De Canditiis, D. and Vidakovic, B. (2004) Wavelet Bayesian block shrinkage via mixtures of normal inverse gamma priors. *J. Comput. Graph. Stat.*, **13**, 383-398.

Delouille, V. (2002) *Nonparametric Stochastic Regression Using Design-adapted wavelets.* Thesis, Louvain: Universit catholique de Louvain.

Dhillon, I.S. (2001) Co-clustering documents and words using bipartite spectral graph partitioning. *Proc. 7th ACM SIGKDD Int'l Conf on Knowledge Discovery and Data Mining (KDD)*, 269-274.

Dhillon, I.S., Guan, Y., and Kulis, B. (2004) Kernel k-means, spectral clustering and normalized cuts. *Proc. 10th ACM SIGKDD Int'l Conf on Knowledge Discovery and Data Mining (KDD)*, 551-556.

Donoho, D.L. and Johnstone, I.M. (1994) Ideal spatial adaptation by waveket shrinkage. *Biometrika*, **81**, 425-455.

Donoho, D.L. and Johnstone, I.M. (1995) Adapting to unknown smoothness via wavelet shrinkage. *J. Am. Stat. Assoc.*, **90**, 1200-1224.

Escobar, M.D. and West, M. (1995) Bayesian density estimation and inference using mixtures. *J. Am. Stat. Assoc.*, **90**, 577-588.

Ferguson, T.S. (1973) A Bayesian analysis of some nonparametric problems. *Ann. Statist.*, **1**, 209-230.

Fielder, M. (1975) Eigenvectors of acyclic matrices. *Czech. Math. J.*, **25**, 607-618.

Gelfand, A.E., Kim, H.K., Sirmans, C.F. and Banerjee, S. (2003) Spatial modelling with spatially varying coefficient processes. *J. Am. Stat. Assoc.*, **98**, 387-396.

Green, P.J. (1995) Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, **82**, 711-732.

Gupta, A.K. and Nagar, D.K. (1999) *Matrix Variate Distributions.* New York: Chapman and Hall/CRC Monographs and Surveys in Pure and Applied Mathematics **104**.

James, G., and Sugar, C. (2003) Clustering for sparsely sampled functional data. *J. Am. Stat. Assoc.*, **98**, 397-408.

Johnson, J.M., Castle, J., Garrett-Engele, P., Kan, Z., Loerch, P.M., Armour, C.D., Santos, R., Schadt, E.E., Stoughton, R. and Shoemaker, D.D. (2003) Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. *Science*, **302**, 2141-2144.

Kim, S., Li, H., Chen, Y., Cao, N., Dougherty, E.R., Bittner, M.L. and Suh, E.B. (2002) Can Markov chain models mimic biological regulation? *J. Biol. Syst.*, **10**, 337-357.

Klevecz, R.R. (2000) Dynamic architecture of the yeast cell cycle uncovered by wavelet decomposition of expression array data. *Functional and Integrative Genomics*, **1**, 186-192.

Korwar, R.M. and Hollander, M. (1973) Contributions to the theory of Dirichlet processes. *Ann. Probab.*, **1**, 705-711.

Kovac, A. and Silverman, B.W. (2000) Extending the scope of wavelet regression methods by coefficient-dependent thresholding. *J. Am. Stat. Assoc.*, **95**, 172-183.

Lauritzen, S.L. (1996) *Graphical Models*. Oxford: Oxford University Press.

Lovascz, L. (1996) Random walks on graphs: a survey. In *Combinatorics, Paul Erdös is Eighty* (eds T. Szonyi, D. Miklos, and V. T. Sos), Vol. 2, pp. 353-398, Budapest: Janos Bolyai Mathematical Society.

Medvedovic, M. and Sivaganesan, S. (2002) Bayesian infinite mixture model-based clustering of gene expression profiles. *Bioinformatics*, **18**, 1194-1206.

Nykter, M., Hunt, K.K., Pollock, R.E., El-Naggar, A.K., Taylor, E., Shmulevich, I., Yli-Harja, O. and Zhang, W. (2006) Unsupervised analysis uncovers changes in histopathologic diagnosis in supervised genomic studies. *Technology in Cancer Research and Treatment*, **5**, 177-182.

Pensky, M., and Vidakovic, B. (2001) On non-equally spaced wavelet regression. *Ann. Inst. Stat. Math.*, **53**, 681-690.

Quintana, F.A. and Iglesias, P.L. (2003) Bayesian clustering and product partition models. *J. R. Stat. Soc.* B, **65**, 557-574.

Reale, M.A., Hu, G., Zafar, A.I., Getzenberg, R.H., Levine, S.M. and Fearon, E.R. (1994) Expression and alternative splicing of the deleted in colorectal cancer (DCC) gene in normal and malignant tissues. *Cancer Research*, **54**, 4493-4501.

Ren, B., Robert, F., Wyrick, J.J., Aparicio, O., Jennings, E.J., Simon, I., Zeitlinger, J., Schreiber, J., Hannett, N., Kanin, E., Volkert, T.L., Wilson, C.J., Bell, S.P. and Young, R.A. (2000) Genome-wide location and function of DNA binding Proteins. *Science*, **290**, 2306-2309.

Rivoirard, J. (1994) *Introduction to Disjunctive Kriging and Non-Linear Geostatistics.* Oxford: Clarendon Press.

Scholkopf, B., Smola, A. and Muller, K.R. (1998) Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, **10**, 1299-1319.

Shoemaker, D.D., Schadt, E.E., Armour, C.D., He, Y.D., Garrett-Engele, P., McDonagh, P.D., Loerch, P.M., Leonardson, A., Lum, P.Y., Cavet, G., Wu, L.F., Altschuler, S.J., Edwards, S., King, J., Tsang, J.S., Schimmack, G., Schelter, J.M., Koch, J., Ziman, M., Marton, M.J., Li, B., Cundiff, P., Ward, T., Castle, J., Krolewski, M., Meyer, M.R., Mao, M., Burchard, J., Kidd, M.J., Dai, H., Phillips, J.W., Linsley, P.S., Stoughton, R., Scherer, S. and Boguski, M. S. (2001) Experimental annotation of the human genome using microarray technology. *Nature*, **409**, 922-927.

Spellman, P.T., Sherlock, G., Zhang, M.Q., Iyer, V.R., Anders, K., Eisen, M.B., Brown, P.O., Botstein, D., and Futcher, B., (1998) Comprehensive identification of cell cycle-regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization. *Mol. Biol. Cell.*, **9**, 327-397.

Sweldens, W. (1996) The lifting scheme: A custom-design construction of biorthogonal wavelets. *Appl. Comput. Harmon. Anal.*, **3**,186-200.

Vidakovic, B. (1998) Nonlinear wavelet shrinkage with Bayes rule and Bayes factors. *J. Am. Stat. Assoc.*, **93**, 173-179.

Wakefield, J., Zhou, C., and Self, S. (2003) Modelling gene expression over time: curve clustering with informative prior distributions. In *Bayesian Statistics, Proc. 7th Valencia Int'l Meeting* (eds J.M. Bernardo, M.J. Bayarri, J.O. Berger, A.P. Dawid, D. Heckerman, A.F.M. Smith, and M. West), Vol 7, pp. 721-732, Oxford: Oxford University Press.

Widmann, M. and Bretherton, C.S. (2000) Validation of mesoscale precipitation in the NCEP reanalysis using a new grid-cell data set for the northwestern United States. *J. Climate*, **13**, 1936–1950.

Wong, F., Carter, C.K. and Kohn, K. (2003) Efficient estimation of covariance selection models. *Biometrika*, **90**, 809-830.

Yeung, K., Fraley, C., Raftery, A., and Ruzzo, W. (2001) Model-based clustering and data transformations for gene expression data. *Bioinformatics*, **17**, 977-987.

APPENDIX A

MARGINAL CALCULATIONS FOR THE HETEROSCEDASTIC MODEL

Recall $\mathbf{Y}(n \times m)$ was used as the collection of n functional responses of length m. For convenience, let $\mathbf{Y}_{\mathcal{C}_n(i)}(n_i \times m)$ denote all the responses falling in the i^{th} cluster $\mathcal{C}_n(i)$. We can write the likelihood as $f(\mathbf{Y}|\{\bar{\beta}, \bar{\sigma}\}_{i=1}^{d_n}, \mathcal{C}_n) = \prod_{i=1}^{d_n} f(\mathbf{Y}_{\mathcal{C}_n(i)}|\bar{\boldsymbol{\beta}}_i, \bar{\sigma}_i^2, \mathcal{C}_n)$ and the marginal as $f(\mathbf{Y}|\boldsymbol{\gamma}, \mathbf{g}, \mathcal{C}_n) = \prod_{i=1}^{d_n} \int f(\mathbf{Y}_{\mathcal{C}_n(i)}|\bar{\boldsymbol{\beta}}_i, \bar{\sigma}_i^2, \mathcal{C}_n) f(\bar{\boldsymbol{\beta}}_i, \bar{\sigma}_i^2) d\bar{\boldsymbol{\beta}}_i d\bar{\sigma}_i^2$. Suppose the i^{th} cluster has n_i responses and $(\bar{\boldsymbol{\beta}}_i, \bar{\sigma}_i^2) \sim \text{NIG}(0, \mathbf{V}; u, v)$ a priori, the i^{th} (of the d_n) integral inside the product can be written as

$$\begin{aligned}
& \frac{(u/2)^{v/2}}{|\mathbf{V}|^{1/2}(2\pi)^{(n_i m+p)/2}\Gamma(v/2)} \int \frac{1}{\bar{\sigma}_i^{(n_i m+v+p+2)/2}} \exp\left\{-\frac{1}{2\bar{\sigma}_i^2} \sum_{i' \in \mathcal{C}(i)} (\mathbf{Y}_{i'} - \mathbf{X}\bar{\boldsymbol{\beta}}_i)^t (\mathbf{Y}_{i'} - \mathbf{X}\bar{\boldsymbol{\beta}}_i)\right\} \\
& \times \exp\left\{-\frac{\text{tr}(\bar{\boldsymbol{\beta}}_i^t \mathbf{V}^{-1} \bar{\boldsymbol{\beta}}_i) + u}{2\bar{\sigma}_i^2}\right\} d\bar{\boldsymbol{\beta}}_i d\bar{\sigma}_i^2 \\
= \; & \frac{|\mathbf{V}_i^*|^{1/2}(u/2)^{v/2}}{|\mathbf{V}|^{1/2}(2\pi)^{(n_i m)/2}\Gamma(v/2)} \int \frac{e^{-u^*/2\bar{\sigma}_i^2}}{\bar{\sigma}_i^{(n_i m+v+2)/2}} \\
& \times \int \frac{1}{(2\pi\bar{\sigma}_i)^{p/2}|\mathbf{V}_i^*|^{1/2}} \exp\left\{(\bar{\boldsymbol{\beta}}_i - \boldsymbol{\mu}_i^*)^t \mathbf{V}_i^{*-1} (\bar{\boldsymbol{\beta}}_i - \boldsymbol{\mu}_i^*)\right\} d\bar{\boldsymbol{\beta}}_i d\bar{\sigma}_i^2 \\
= \; & \frac{|\mathbf{V}_i^*|^{1/2}(u/2)^{v/2}}{|\mathbf{V}|^{1/2}(2\pi)^{(n_i m)/2}\Gamma(v/2)} \int \frac{e^{-u^*/2\bar{\sigma}_i^2}}{\bar{\sigma}_i^{(n_i m+v+2)/2}} d\sigma_i^2 \propto \frac{\Gamma((v+mn_i)/2)}{\pi^{(n_i m)/2}} \frac{|\mathbf{V}_i^*|^{1/2}}{|\mathbf{V}|^{1/2}} (u_i^*)^{-(v+mn_i)/2}
\end{aligned}$$

where $\mathbf{V}_i^* = (n_i I + \mathbf{V}^{-1})^{-1}$, $\boldsymbol{\mu}_i^* = \mathbf{V}_i^* \sum_{i' \in \mathcal{C}_n(i)} \mathbf{X}^t \mathbf{Y}_{i'}$ and $u_i^* = u + \sum_{i' \in \mathcal{C}_n(i)} \mathbf{Y}_{i'}^t \mathbf{Y}_{i'} - \boldsymbol{\mu}_i^{*t} \mathbf{V}_i^{*-1} \boldsymbol{\mu}_i^*$. Multiplying over all d_n clusters, we get

$$f(\mathbf{Y}|\boldsymbol{\gamma}, \mathbf{g}, \mathcal{C}_n) \propto \prod_{i=1}^{d_n} \frac{\Gamma((v+mn_i)/2)}{\pi^{n_i m/2}} \frac{|\mathbf{V}_i^*|^{1/2}}{|\mathbf{V}|^{1/2}} (u_i^*)^{-(v+mn_i)/2}$$

APPENDIX B

PROOF FOR BESOV PRIORS

This is an extension of the proof for Theorem 2 in Abramovich *et al.* (1998) for the special case of finite Besov scales $p, q < \infty$. This condition ensures that the complete metric parameter space is separable (for example, Blackwell & MacQueen, 1973). Also, we do not consider a third parameter ρ satisfying $g_j = c_1 j^\rho 2^{-aj}$.

In univariate notation, the prior on the wavelet coefficients is $f(\beta_{jk}|g_j, \sigma^2) = N(0, \sigma^2 g_j \gamma_{jk} I)$ and $f(g_j) \sim IG(r_j, s_j)$ implies $f(\beta_{jk}|\gamma_{jk}, \sigma^2) = t_{s_j}(0, r_j\gamma_{jk}\sigma^2)$. We will need the moments $\mathrm{E}||\boldsymbol{\beta}_j||_p^p = \sum_k \mathrm{E}\beta_{jk}^p$ and $\mathrm{E}(||\boldsymbol{\beta}_j||_p^{2p}) = \sum_k \mathrm{E}\beta_{jk}^{2p} + \sum_{k\neq k'} \mathrm{E}(\beta_{jk}\beta_{jk'})^p$, where $\boldsymbol{\beta}_j$ are the coefficients at the j^{th} resolution. If ν_p is the p^{th} moment of $N(0,1)$, then

$$\mathrm{E}\beta_{jk}^p = \mathrm{E}(\mathrm{E}(\beta_{jk}^p|g_j)) = c_2\sigma^p 2^{-bj}\nu_p \mathrm{E}(g_j^{p/2}).$$

Thus, we have $\mathrm{E}||\boldsymbol{\beta}_j||_p^p = c_2\sigma^p 2^{(1-b)j}\nu_p\mathrm{E}(g_j^{p/2})$ and $\mathrm{E}(||\boldsymbol{\beta}_j||_p^{2p}) = c_2\sigma^{2p}2^{(1-b)j}\nu_{2p}\mathrm{E}(g_j^p) + c_2^2\sigma^{2p}2^j(2^j-1)2^{-2bj}\nu_{2p}^2\mathrm{E}(g_j^p) \leq c_2\sigma^{2p}\nu_{2p}2^{(1-b)j}(1 + c_2 2^{(1-b)j}\nu_{2p})\mathrm{E}(g_j^p)$.

Given these moments, by Chebyshev Inequality, we have for some $\epsilon > 0$,

$$\Pr(|2^{-(1-b)j}||\boldsymbol{\beta}_j||_p^p - \sigma^p c_2\nu_p\mathrm{E}(g_j^{p/2})| > \epsilon) \leq \epsilon^2 2^{-2(1-b)j}\mathrm{E}(||\boldsymbol{\beta}_j||_p^{2p})$$

and to apply the Borel-Cantelli Lemma

$$\textstyle\sum_{j=0}^\infty \Pr(|2^{-(1-b)j}||\boldsymbol{\beta}_j||_p^p - \sigma^p c_2\nu_p\mathrm{E}(g_j^{p/2})| > \epsilon) \leq \epsilon^2 \sum_{j=0}^\infty 2^{-2(1-b)j}\mathrm{E}(||\boldsymbol{\beta}_j||_p^{2p}) < \infty$$

Thus $2^{-(1-b)j}||\boldsymbol{\beta}_j||_p^p \to c_2\sigma^p\nu_p\mathrm{E}(g_j^{p/2})$ almost surely.

Since this is true for $j = 0, 1, \ldots$, the Besov sequence norm is finite if

$$\textstyle\sum_{j=0}^\infty 2^{j(\ell+1/2-1/p)q}2^{(1-b)jq/p}\mathrm{E}(g_j^{p/2})^{q/p} < \infty$$

The infinite sum on the right hand side is finite, if $\mathrm{E}(g_j^{p/2})^{1/p} \propto 2^{-aj}$ for all $j = 0, 1, \ldots$, where a satisfies $(b-1)/p + (a-1)/2 \geq \ell - 1/p$.

To summarise, if for all j,

$$\mathrm{E}(g_j^{p/2})^{1/p} = \frac{{r_j}^{1/2}}{\{(s_j - 2)(s_j - 4) \ldots (s_j - p)\}^{1/p}} = c_1 2^{-aj}$$

where a satisfies $(b-1)/p + (a-1)/2 \geq \ell - 1/p$, then the Besov correspondence holds.

APPENDIX C

PROOF OF THEOREM 1

We consider the conditional posterior of $(\boldsymbol{\beta}_{n+1}|\mathbf{Y}, \boldsymbol{\beta}_{-(n+1)}, \sigma^2)$ in the oversmoothed model. The probability that $\boldsymbol{\beta}_{n+1}$ is not tied to any of the previous samples is

$$q_{n+1} = \frac{\sum_{i=1}^{n} \phi(\mathbf{Y}_{n+1}|\mathbf{X}\boldsymbol{\beta}_i, \sigma^2\mathbf{I}_m)}{\alpha\phi(\mathbf{Y}_{n+1}|0, \sigma^2(\mathbf{I}_m + \mathbf{X}\mathbf{V}\mathbf{X}')) + \sum_{i=1}^{n} \phi(\mathbf{Y}_{n+1}|\mathbf{X}\boldsymbol{\beta}_i, \sigma^2\mathbf{I}_m)} \leq \frac{1}{1 + \alpha R^*_{n+1}/n}$$

where

$$R^*_{n+1} = \frac{\phi(\mathbf{Y}_{n+1}|0, \sigma^2(\mathbf{I}_m + \mathbf{X}\mathbf{V}\mathbf{X}'))}{\phi(\mathbf{Y}_{n+1}|\mathbf{X}\bar{\boldsymbol{\beta}}, \sigma^2\mathbf{I}_m)}$$

and $\bar{\boldsymbol{\beta}} \in \{\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_n\}$ s.t. $||\mathbf{Y}_{n+1} - \mathbf{X}\bar{\boldsymbol{\beta}}||_2$ is minimum. Also, we can write

$$\begin{aligned} \log \mathrm{E}R^*_{n+1} &= \log \mathrm{E}_{\boldsymbol{\beta}} \frac{\phi(\mathbf{Y}_{n+1}|\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_m)}{\phi(\mathbf{Y}_{n+1}|\mathbf{X}\bar{\boldsymbol{\beta}}, \sigma^2\mathbf{I}_m)} \geq \mathrm{E}_{\boldsymbol{\beta}} \log \frac{\phi(\mathbf{Y}_{n+1}|\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_m)}{\phi(\mathbf{Y}_{n+1}|\mathbf{X}\bar{\boldsymbol{\beta}}, \sigma^2\mathbf{I}_m)} \\ &= \frac{1}{2\sigma^2}\left\{||\mathbf{Y}_{n+1} - \mathbf{X}\bar{\boldsymbol{\beta}}||_2^2 - ||\mathbf{Y}_{n+1}||_2^2 - \sigma^2 \mathrm{tr}(\mathbf{V})\right\} \equiv \log R_{n+1} \end{aligned}$$

Finally, writing $q_{n+1} \leq (1 + \alpha R_n/n)^{-1}$, we get

$$\mathrm{E}_{\mathbf{Y}_{n+1}} q_{n+1} \leq \mathrm{E}\left(\frac{1}{1 + \alpha R_n/n}\right) \approx \frac{1 + \alpha \mathrm{E}R_n/n}{\exp(2\mathrm{E}\log(1 + \alpha R_n/n))}.$$

Since R_n is small for large n, $\exp(2\mathrm{E}\log(1+\alpha R_n/n)) = \exp(2\alpha\mathrm{E}R_n/n) - O(n^{-2})$ and if $\mathbf{X}\boldsymbol{\beta}_{n+1}$ is the actual function underlying $\mathbf{Y}_{n+1}$, we have

$$\mathrm{E}_{\mathbf{Y}_{n+1}} q_{n+1} \leq \frac{1 + \alpha\mathrm{E}R_n/n}{\exp(2\alpha\mathrm{E}R_n/n)},\ \mathrm{E}R_n = e^{5||\bar{\boldsymbol{\beta}}||_2^2/8\sigma^2} e^{-\mathrm{tr}(\mathbf{V})/2} \exp\left\{-\frac{1}{2\sigma^2}\bar{\boldsymbol{\beta}}^t\boldsymbol{\beta}_{n+1}\right\} \quad \text{(C.1)}$$

since $\mathrm{E}e^{\mathbf{u}^t\mathbf{x}} = \exp(\mathbf{u}^t\boldsymbol{\mu} + \mathbf{u}^t\boldsymbol{\Sigma}\mathbf{u}/2)$ for $\mathbf{x} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. We assume that $||\bar{\boldsymbol{\beta}}||_2, ||\boldsymbol{\beta}_{n+1}||_2 < \infty$ almost surely by prior specification and let $\rho_{n+1} = \bar{\boldsymbol{\beta}}^t\boldsymbol{\beta}_{n+1}$ be the inner product of the actual functional and the closest available functional. For large n the sample space becomes dense and we expect ρ_{n+1} to increase. Then from (C.1)

$$\sum_{n=1}^{\infty} \mathrm{E}_{\mathbf{Y}_{n+1}} q_{n+1} \leq \sum_{n=1}^{\infty} \frac{1 + \frac{C_3}{n} e^{-\rho_{n+1}/2\sigma^2}}{e^{\frac{C_4}{n} e^{-\rho_{n+1}/2\sigma^2}}} < \infty, \, C_3, C_4 > 0$$

if $\rho_{n+1} \approx \sigma^2 \log(\log n)^{1+\delta}$ for some $\delta > 0$. By the Borell-Cantelli lemma, this means that the new sample is almost surely distinct if the inner-product (or the ℓ_2 distance) is less (greater) than $\sigma^2 O(\log(\log n)^{1+\delta})$.

VITA

Shubhankar Ray was born in Calcutta, India on July 18, 1979. He graduated from the Indian Institute of Technology (IIT) at Guwahati, India in May 2000 with a Bachelor of Technology (B.Tech.) in Electronics and Communications Engineering. In May 2002, he received a Master of Science degree in Electrical Engineering from Texas A&M University. He received his Ph.D. in Statistics from Texas A&M University in August of 2006 under the direction of Dr. Bani K. Mallick and Dr. Raymond J. Carroll.

The typist for this thesis was Shubhankar Ray.