Computational Statistics and Data Analysis 109 (2017) 93–104

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Adjusting background noise in cluster analyses of longitudinal data

Shengtong Han [a], Hongmei Zhang [a,*], Wilfried Karmaus [a], Graham Roberts [b], Hasan Arshad [c]

[a] School of Public Health, University of Memphis, Memphis, TN, United States
[b] Paediatric Allergy and Respiratory Medicine, University of Southampton, Southampton, UK
[c] Allergy and Clinical Immunology, Clinical and Experimental Sciences, University of Southampton, Southampton, UK

ARTICLE INFO

Article history:
Received 31 March 2016
Received in revised form 11 November 2016
Accepted 12 November 2016
Available online 27 November 2016

Keywords:
Dirichlet process
Clustering
Bayesian methods
Longitudinal data

ABSTRACT

Background noise in cluster analyses can potentially mask the true underlying patterns. To tease out patterns uniquely to certain populations, a Bayesian semi-parametric clustering method is presented. It infers and adjusts background noise. The method is built upon a mixture of the Dirichlet process and a point mass function. Simulations demonstrate the effectiveness of the proposed method. The method is then applied to analyze a longitudinal data set on allergic sensitization and asthma status.

1. Introduction

An underlying background pattern refers to a pattern shared by all subjects. In cluster analyses, existing methods do not explicitly infer background patterns and adjust for their effects when identifying unique patterns. This work is motivated by an epidemiological study examining allergic sensitization patterns longitudinally among subjects with different asthma statuses. We are interested in whether allergic sensitization to grass changes across different time points from pre-schoolers to young adult (ages 4–18 years). If it does, what is the longitudinal pattern for each group of subjects? Of particular interest is the unique patterns among subjects with diseases after adjusting for the background pattern, which is a temporal allergic sensitization trend shared in the general population. A pattern after excluding the background is referred to as a unique pattern. Unique patterns are important in many applications, e.g., unique allergic sensitization patterns for different group of subjects offer a potential to predict allergic disease risks. To identify patterns, cluster analyses aiming to detect the similarity between subjects are commonly implemented. In general, all clustering methods are either non-parametric, e.g., the K-means approach, or model-based (semi-) parametric approaches (Fraley and Raftery, 2002; Bigelow and Dunson, 2009; Scott, 2009; Nieto-Barajas and Contreras-Cristan, 2014; Efron et al., 2001; Kim et al., 2006; Dunson et al., 2008). In this article, we propose a model-based approach built upon Bayesian splines to infer longitudinal patterns of subjects in background and unique clusters.

Most model-based methods perform clustering based on the means of a set of variables. Maximum likelihood approaches have been used to detect the clusters and estimate the associations (Qin and Self, 2006). Bayesian methods have been

* Corresponding author.
E-mail address: hzhang6@memphis.edu (H. Zhang).

http://dx.doi.org/10.1016/j.csda.2016.11.009
0167-9473/

developed for the purpose of automatic determination of the number of clusters and to take advantage of prior knowledge on the potential associations. Our proposed method is formulated under this framework. To estimate the number of clusters automatically, sampling strategies such as the reversible jump Markov Chain Monte Carlo (MCMC) method (Green, 1995) and the birth–death process (Stephens, 2000) have been proposed. Dirichlet process (DP) (Ferguson, 1973), introduced as the prior distribution for the coefficients evaluating the strength of associations, has been used more often recently to estimate the number of clusters and infer the coefficients (Escobar and West, 1995; Caron et al., 2014; Zhang et al., 2012; Kim et al., 2009).

There is a gap in the literature of clustering methods. Longitudinal data are common nowadays and measures of clustering variables may be further influenced by one or more external variables. Some existing clustering methods can be applied to longitudinal data, but they do not infer temporal patterns or cannot explicitly describe temporal patterns in each cluster. For instances, McNicholas and Murphy (2010) proposed a parametric approach to cluster overall means (i.e., to infer overall mean μ_g for cluster g as noted in the article) instead of temporal patterns, where a modified Cholesky decomposition is implemented to ease the process of clustering. A non-parametric approach, implemented in an R package `kml` (Genolini and Falissard, 2011), uses the K-means method to perform the clustering such that subjects with similar profiles (i.e., means) over time are grouped together. More importantly, these existing methods do not possess the feature of differentiating unique patterns from background patterns. In this article, we develop an approach to cluster subjects that has the ability to simultaneously identify background and unique temporal patterns. The method identifies clusters based on temporal trends with background pattern adjusted. It has the potential to be easily modified to fit non-longitudinal data. The detailed model specification, including model assumptions, parameter priors and posteriors are presented in Section 2. Numeric studies including simulations and an application example are given in Section 3. Summary and discussions are included in Section 4.

2. Model specification

2.1. Model

Let Y_{it} denote a measure of response for subject i at time t with vector $\boldsymbol{Y_i} = \{Y_{i1}, \ldots, Y_{iT}\}$ being the ith observation over T time units and $\boldsymbol{Y}_{I\times T} = \{\boldsymbol{Y_1}, \ldots, \boldsymbol{Y_I}\}$ denote all the observations. Assume $\boldsymbol{Y_i}$ has a temporal pattern and is associated with time invariant covariates $\boldsymbol{X_i}$,

$$Y_{it} = \boldsymbol{X_i^T\beta_0} + f_1(t_i; \boldsymbol{\gamma_0}, \boldsymbol{b_0}) + \boldsymbol{X_i^T\beta_i} + f_2(t_i; \boldsymbol{\gamma_i}, \boldsymbol{b_i}) + s_i + \epsilon_{it}, \tag{1}$$

where $f_1(\cdot)$ is an unknown function describing the temporal pattern applicable to all subjects (background pattern) and $f_2(\cdot)$ is for temporal pattern specific to subject i (with background adjusted), s_i represents subject random effects, and ϵ_{it} is measurement error. Model (1) consists of two parts. The first part $\boldsymbol{X_i^T\beta_0} + f_1(t_i; \boldsymbol{\gamma_0}, \boldsymbol{b_0})$ describes the background pattern common to all subjects and $\boldsymbol{X_i^T\beta_i} + f_2(t_i; \boldsymbol{\gamma_i}, \boldsymbol{b_i})$ describes the trend specifically for subject i. Note that in the situation of $\boldsymbol{X}_i$ being categorical, constraints on $\boldsymbol{X}_i$ are needed to avoid design matrix being singular. Assuming $\epsilon_{it} \sim N(0, \tau)$ and $s_i \sim N(0, \sigma_s^2)$, the distribution of $\boldsymbol{Y_i}$ satisfies a multivariate normal distribution

$$\boldsymbol{Y_i}|\boldsymbol{\theta_0}, \boldsymbol{\theta_i} \sim N(\boldsymbol{M_i}, \Sigma), \tag{2}$$

with $\boldsymbol{M_i} = \boldsymbol{X_i\beta_0} + f_1(\boldsymbol{t_i}; \boldsymbol{\gamma_0}, \boldsymbol{b_0}) + \boldsymbol{X_i\beta_i} + f_2(\boldsymbol{t_i}; \boldsymbol{\gamma_i}, \boldsymbol{b_i})$ being a $T \times 1$ vector, $\boldsymbol{t_i} = (t_{i1}, t_{i2}, \ldots, t_{iT})$, Σ a $T \times T$ matrix with $\sigma_s^2 + \tau$ on the diagonal and off diagonal σ_s^2, $\boldsymbol{\theta}_0 = (\boldsymbol{\beta}_0, \boldsymbol{\gamma}_0, \boldsymbol{b}_0)^T$ denoting parameters in the background, and $\boldsymbol{\theta}_i = (\boldsymbol{\beta}_i, \boldsymbol{\gamma}_i, \boldsymbol{b}_i)^T$ being the collection of parameters unique to subject i. As seen in the construction of (1), $\boldsymbol{\theta}_i$ is added onto $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}_i = \boldsymbol{0}$ if subject i does not have a unique temporal trend. Note that expression (2) collapses to a univariate distribution when there is no repeated measures, i.e., $T = 1$.

We take Bayesian P-splines (Eilers and Marx, 1996; Baladandayuthapani et al., 2005) with order 2 for functions $f_1(\cdot)$ and $f_2(\cdot)$ to estimate the unknown common and subject specific temporal trends. Specifically, we define

$$\begin{aligned} f_1(t_i, \boldsymbol{\gamma}_0, \boldsymbol{b}_0) &= \gamma_{00} + \gamma_{01}t_i + \gamma_{02}t_i^2 + \sum_{j=1}^{N} b_{0j}(t_i - t_{ij}^*)_+^2, \\ f_2(t_i, \boldsymbol{\gamma}_i, \boldsymbol{b}_i) &= \gamma_{i0} + \gamma_{i1}t_i + \gamma_{i2}t_i^2 + \sum_{j=1}^{N} b_{ij}(t_i - t_{ij}^*)_+^2, \end{aligned} \tag{3}$$

where $t_i \in (t_{i1}, t_{i2}, \ldots, t_{iT})$, and $(x)_+^2 = x^2 I(x \geq 0)$ and N is the number of knots. The knots can be chosen as equally spaced points such that $N \ll T$ with $t_{i1}^* < t_{i2}^* < \cdots < t_{iN}^*$ from the T time points.

2.2. Strategy of clustering and specification of prior distributions

The clustering of subjects and identification of background pattern are fulfilled by a careful design of the prior distribution of $\boldsymbol{\theta}_i$. In the following, we focus on the discussion of prior distribution of $\boldsymbol{\theta}_i$ and list the prior distributions for $\boldsymbol{\theta}_0$ and other

parameters at the end of this section. If subject i's temporal pattern is as in the general population, we have $\boldsymbol{\theta}_i = \mathbf{0}$. If not, then each subject has his/her own temporal trend which, in public health or medical studies, is possibly related to a manifestation of a disease, e.g., asthma. To this end and be flexible on the prior distribution of $\boldsymbol{\theta}_i$, we take the prior distribution of $\boldsymbol{\theta}_i$ as a mixture of distribution G_1 and point mass $\delta(\boldsymbol{\theta}_i = 0)$,

$$\boldsymbol{\theta}_i | G_1, \omega \sim \omega G_1 + (1-\omega)\delta(\boldsymbol{\theta}_i = 0), \tag{4}$$

with G_1 generated from a Dirichlet Process (DP), $G_1 \sim DP(\alpha_1, G_{01})$, where G_{01} is the base distribution and assumed to be truncated multivariate normal $G_{01} = TMVN(\boldsymbol{\mu}_{01}, \boldsymbol{\Sigma}_{01}, a)$ with a being the distance from zero, and α_1 a precision parameter controlling the distance between G_1 and G_{01}. G_1 is included for describing unique temporal trends and clustering subjects with similar trends, while $\delta(\cdot)$ is to put a subject into a group bearing a background pattern only. Parameter ω denotes the probability that subject i has a unique temporal trend. Forcing $\omega = 1$ reduces to a model for standard clustering in the Bayesian framework and patterns in the background will not be adjusted. The parameter a is pre-specified indicating our belief that θ_i is non-zero. In our application, we set $a = 0.0001$.

The mixture structure in (4) was motivated by the mixture of two normal distributions utilized in Bayesian variable selection modeling (George and McCulloch, 1997), which was later adapted to incorporate clustering into variable selection with the involvement of DP (Dunson et al., 2008). In our application, however, this structure moved away from its original variable selection functionality and, instead, is suited to model a situation of differentiating unique patterns from background patterns in addition to clustering. Following the properties of DP, we have the following conditional prior distribution for $\boldsymbol{\theta}_i$ with $(\cdot)$ denoting all other parameters,

$$\boldsymbol{\theta}_i | \boldsymbol{\theta}^{(i)}, (\cdot) \sim \omega \frac{1}{I-1+\alpha_1} \sum_{j \neq i} \delta(\boldsymbol{\theta}_j) + \omega \frac{\alpha_1}{I-1+\alpha_1} TMVN(\boldsymbol{\mu}_{01}, \boldsymbol{\Sigma}_{01}, a) + (1-\omega)\delta(\boldsymbol{\theta}_i = 0). \tag{5}$$

DP provides a non-parametric prior in the space of distribution functions and gives rise to a more flexible class of distributions than would be obtained by parametric approaches. Furthermore, the inherent clustering property of samples drawn from a distribution with DP prior ensures the formation of clusters among $\boldsymbol{\theta}_i$. Details of DP can be found in Ferguson (1973), Escobar and West (1995), Antoniak (1974), and among others. Taking this prior for $\boldsymbol{\theta}_i$, it assumes that, among subjects not following the temporal trend in the general population (determined by $\boldsymbol{\theta}_0$), there exist clusters of subjects such that subjects in one cluster share a unique temporal pattern on average. The feature of this prior is beneficial to public health and medical researchers due to the potential connection between temporal patterns and phenotypic manifestation, e.g., different temporal patterns of a phenotype may be linked to different levels of disease manifestation, providing a potential for disease prediction. To clarify, throughout the article, a "temporal pattern" refers to a "temporal trend" shared by a group of subjects. The parameter α_1 in G_1 influences the degree of clustering of $\boldsymbol{\theta}_i$'s. There are various methods for selecting or estimating α_1 (Escobar and West, 1995; Murugiah and Sweeting, 2012; Ritter and Tanner, 1992; Dorazio, 2009). However, most methods are burdensome and suffer from insensitivity to underlying patterns. Instead, in this article, we choose α_1 by optimizing the deviance information criterion (DIC) (Gelman et al., 2003). Note that the recently raised concern on DIC is in terms of over-fitting, which will not influence the selection of α_1 (Spiegelhalter et al., 2014). In general, the larger the value of α_1, the more clusters will be generated. Thus, alternatively, if we do not expect a large number of patterns, then setting α_1 relatively small, e.g., $\alpha_1 = 0.1$, is acceptable in general.

We now discuss the hyper-prior distributions for parameters in (5), including $\boldsymbol{\mu}_{01}$, $\boldsymbol{\Sigma}_{01}$, and ω. Parameters $\boldsymbol{\mu}_{01}$ and $\boldsymbol{\Sigma}_{01}$ are in the base distribution G_{01} for $\boldsymbol{\theta}_i$. The prior distribution of $\boldsymbol{\mu}_{01}$ is assumed to follow a multivariate normal distribution with mean $\mathbf{0}$ and known large diagonal covariance matrix $\boldsymbol{\Sigma}_{\mu_{01}}$. Covariance matrix $\boldsymbol{\Sigma}_{01}$ is a diagonal matrix composed of variance parameters corresponding to $\boldsymbol{\beta}_i$, $\boldsymbol{\gamma}_i$, and $\boldsymbol{b}_i$ in $\boldsymbol{\theta}_i$ (Section 2.1). We take inverse gamma (IG) with small and known hyper-prior parameters as the prior distributions for each of its components. Specifically, for variances linked to $\boldsymbol{\beta}_i$, we have $\sigma_c^2 \sim IG(\alpha_{\sigma_c}, \beta_{\sigma_c})$, $c = 1, 2, \ldots, C$, assuming C covariates in total. For variance corresponding to $\boldsymbol{\gamma}_i$ coefficients in the P-spline, we have $\sigma_c^2 \sim IG(\alpha_{\sigma_c}, \beta_{\sigma_c})$, $c = C+1, \ldots, C+3$, and finally for $\boldsymbol{b}_i$, since they represent the random part of P-splines, we assume a constant variance across all b_{ij}'s, that is, $\sigma_b^2 \sim IG(\alpha_{\sigma_b}, \beta_{\sigma_b})$. For the weight parameter ω, we assume $\omega \sim \text{Beta}(2, 2)$, which is symmetric and bell-shaped.

Finally, we present the prior distributions for parameters in $\boldsymbol{\theta}_0$, τ, and σ_s^2. The prior distribution of $\boldsymbol{\theta}_0$ is assumed to be multivariate normal with mean $\mathbf{0}$ and covariance $\boldsymbol{\Sigma}_{\theta 0}$, a known diagonal matrix with large components. The variance component τ is assumed to follow an inverse gamma distribution, $\tau \sim IG(\alpha, \beta)$ with α and β known and chosen small. The same prior distribution is assigned to σ_s^2.

2.3. Posterior distribution computing

The joint posterior distribution of all parameters, $\mathcal{A} = \{\boldsymbol{\theta}_i, \boldsymbol{\mu}_{01}, \boldsymbol{\Sigma}_{01}, \omega, \boldsymbol{\theta}_0, \tau, \sigma_s^2\}$, is (up to a normalization constant),

$$P(\mathcal{A}|\boldsymbol{Y}) \propto p(\boldsymbol{Y}|\boldsymbol{\theta}_i, \boldsymbol{\theta}_0, \tau, \sigma_s^2) p(\boldsymbol{\theta}_i|G_1, \omega) p(G_1|G_{01}, \alpha_1) p(\boldsymbol{\mu}_{01}) p(\boldsymbol{\Sigma}_{01}) p(\omega) p(\boldsymbol{\theta}_0) p(\tau) p(\sigma_s^2), \tag{6}$$

where $G_{01} = TMVN(\boldsymbol{\mu}_{01}, \boldsymbol{\Sigma}_{01}, a)$.

Posterior inference of $\mathcal{A}$ is obtained by successively simulating values from their full conditional posterior distributions through a Gibbs sampling scheme. We briefly discuss these distributions below and the detail of each conditional posterior

distribution and its derivation are included in the Appendix. It can be derived that with $\omega = 1$ the conditional posterior distribution of $\boldsymbol{\theta}_i$ is a Dirichlet process. In this case, we will use the Algorithm 8 in Neal (2000) to sample $\boldsymbol{\theta}_i$. In the Appendix, we listed the conditional posterior distributions for $\boldsymbol{\theta}_i$ needed for Algorithm 8 to update unique values of each cluster. For the hyper-prior parameters, $\boldsymbol{\mu}_{01}$, $\boldsymbol{\Sigma}_{01}$, in the base distribution G_{01}, the conditional posterior distribution $\boldsymbol{\mu}_{01}$ is $N(\boldsymbol{\mu}, \boldsymbol{\Delta})$ with $\boldsymbol{\mu} = \boldsymbol{\Delta}\boldsymbol{\Sigma}_{01}{}^{-1}\sum_i \boldsymbol{\theta}_i$ and $\boldsymbol{\Delta} = (\boldsymbol{\Sigma}_{\mu_{01}}{}^{-1} + I\boldsymbol{\Sigma}_{01}{}^{-1})^{-1}$, and the conditional posterior distributions of the variance components in $\boldsymbol{\Sigma}_{01}$ are still inverse gamma with updated shape and scale parameters,

$$\sigma_c^2|(\cdot) \sim IG\left(\alpha_{\sigma_c} + \frac{I}{2}, \beta_{\sigma_c} + \frac{1}{2}\sum_i (\theta_{ic} - \mu_{01}(c))^2\right),$$

$$\sigma_b^2|(\cdot) \sim IG\left(N\left(\alpha_b + 1 + \frac{I}{2}\right) - 1, \beta_b N + \frac{1}{2}\sum_{c=C+3+1}^{C+3+N}\sum_i (\theta_{ic} - \mu_{01}(c))^2\right),$$

where $\mu_{01}(c)$ is the cth element of $\boldsymbol{\mu}_{01}$, $i = 1, 2, \ldots, C+3$; $c = C+3+1, \ldots, C+3+N$.

For parameters in $\boldsymbol{\theta}_0$, below is the conditional posterior distribution of $\boldsymbol{\beta}_0$.

$$\boldsymbol{\beta}_0|(\cdot) \sim N(\boldsymbol{\mu}, \Delta), \quad \text{with}$$

$$\boldsymbol{\mu} = \Delta\sum_{i=1}^{I} \boldsymbol{X}_i^T\boldsymbol{\Sigma}^{-1}\boldsymbol{R}_i(\boldsymbol{\beta}_0),$$

$$\Delta = \left(\boldsymbol{\Sigma}_{\beta_0}{}^{-1} + \sum_{i=1}^{I}\boldsymbol{X}_i^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X}_i\right)^{-1},$$

$$\boldsymbol{R}_i(\boldsymbol{\beta}_0) = \boldsymbol{Y}_i - f(\boldsymbol{t}_i; \boldsymbol{\gamma}_0, \boldsymbol{b}_0) - \boldsymbol{X}_i\boldsymbol{\beta}_i - f(\boldsymbol{t}_i; \boldsymbol{\gamma}_i, \boldsymbol{b}_i).$$

The conditional posterior distributions of $\boldsymbol{\gamma}_0$, $\boldsymbol{b}_0$ are in similar forms and included in the Appendix.

For the variance components τ and σ_s^2 in the distribution of ϵ and s_i, respectively, their conditional posteriors are not standard and will be sampled using the Metropolis–Hastings (M–H) algorithm in the Gibbs sampler. Their proposal functions are both log-normal distributions with mean centered at the current posterior sample and variance selected to achieve convergence efficiency. Similarly, for ω, we will use M–H algorithm as well and the proposal function will be uniform centered at the current posterior sample with range selected to achieve convergence efficiency.

The number of clusters is inferred based on the "least-squares clustering" method proposed by Dahl (2006) and applied in our recent work (Zhang et al., 2012). Basically, this method uses converged MCMC simulation samples to estimate a probability matrix showing the probability of each pair of subjects being clustered together, and then chooses one specific MCMC simulation sample such that the clustering pattern in that iteration is closest to the probability matrix in Euclidean distance. In terms of MCMC simulations, it is noticed that model (1) is actually a linear mixed model and the coefficients can be easily estimated using the R function *lmer* instead of going through conditional posterior sampling. To improve the efficiency of MCMC, we introduce the idea of empirical Bayes (Efron et al., 2001) into the Gibbs sampling process in that at each iteration of MCMC after we conclude the cluster assignment, we estimate the coefficient parameters via the R function *lmer*, which gives maximum likelihood estimators. All the posterior sampling and results summary are programmed into R and available to readers.

3. Numerical studies

3.1. Simulated experiments

Simulation scenarios: To evaluate the performance of the proposed clustering method, we consider various scenarios. Three different sample sizes are used, $I = 60, 100, 400$, of which 50% subjects are in the background. Another 50% subjects are equally assigned into 2 clusters. As for the unique temporal patterns added onto background (the pattern for $\boldsymbol{\theta}_i$), we consider two settings. In the first setting, both patterns are linear but with different slopes; for cluster 1, $f_2(\cdot) = -10 + 50t$, and cluster 2, $f_2(\cdot) = 30 - 70t$. In the second setting, we let one temporal trend be linear and the other be quadratic; cluster 1, $f_2(\cdot) = -7 + 23t$, and cluster 2, $f_2(\cdot) = 15 + 50t - 10t^2$. For the background or the common temporal pattern in the general population, we take a linear pattern, that is, $f_1(\cdot) = -3 + 9t$. The variances in random subject effects and random errors, σ_s^2 and τ, are taken as 0.25. One dimensional covariate $\boldsymbol{X_i}$ is included in the simulated data. In total, 100 Monte Carlo (MC) replicates are considered for each scenario. To evaluate the method, for each MC replicate, we calculate the sensitivity and specificity with respect to cluster and background identifications and summarize these two statistics across 100 MC replicates using the means and standard deviations.

Results: For the precision parameter α_1, as noted in Section 2.2, we determine it via grid search by minimizing DIC. Fast convergence of the MCMC chains is observed in all of our simulation scenarios. Fig. 1 shows the trace plot of the two scale parameters τ, σ_s^2 as an illustration. Other parameters showed similar patterns of fast convergence. Thus, for each MC

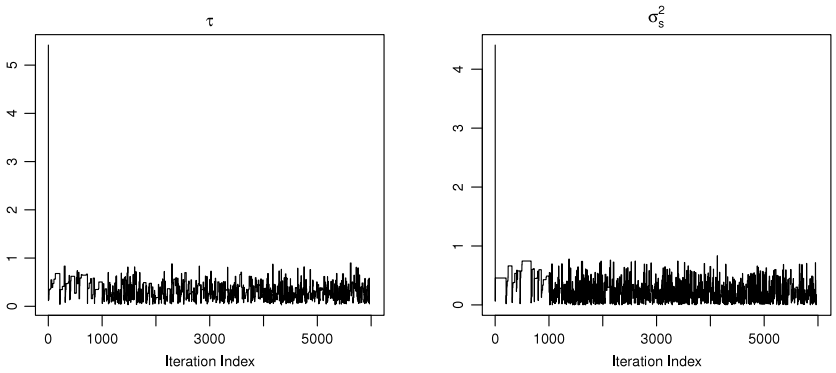

Fig. 1. Trace plots of two scale parameters $\tau = 0.25$, $\sigma_s^2 = 0.25$ with sample size $I = 400$ in setting 2.

Table 1
Summary of sensitivity and specificity across 100 MC replicates for varying sample sizes under setting 1. Sensitivity: probability of true cluster or background assignment. Specificity: probability of true exclusion from a cluster or the background. SD stands for standard deviation. 100* is the unbalanced case where 10 subjects are in cluster 1, 30 subjects are in cluster 2 and the rest 60 are in background. All other cases have equal number of subjects in clusters 1 and 2.

			Sample Size			
			60	100	100*	400
Background	Sensitivity	Mean	0.9997	1.0000	1.0000	1.0000
		SD	0.0032	0.0000	0.0000	0.0000
	Specificity	Mean	1.0000	1.0000	1.0000	1.0000
		SD	0.0000	0.0000	0.0000	0.0000
Cluster 1	Sensitivity	Mean	0.9728	1.0000	0.9846	1.0000
		SD	0.1351	0.0000	0.1083	0.0000
	Specificity	Mean	0.9903	0.9992	0.9876	0.9988
		SD	0.0406	0.0031	0.0090	0.0017
Cluster 2	Sensitivity	Mean	0.9745	0.9981	0.9604	0.9997
		SD	0.0944	0.0084	0.0370	0.0017
	Specificity	Mean	0.9951	1.0000	1.0000	0.9894
		SD	0.0267	0.0000	0.0000	0.0000

replicate, 1800 iterations were run for burn-in and additional 200 iterations were used to infer the parameters and clusters. The posterior inferences are obtained from posterior samples of one chain after burn-in iterations.

For both scenarios, precision parameter α_1 is chosen as 0.01 by minimizing DIC. Summary statistics of clustering quality including sensitivity and specificity are displayed in Tables 1 and 2. Overall, a pattern of improvement in mean sensitivity and specificity is observed for background, cluster 1, and cluster 2 as the sample size increases from 60 to 400 in both scenarios, the variation (indicated by standard deviations) of both statistics decreases. We also notice that for the same sample size, sensitivities and specificities in setting 1 are overall slightly better than those in setting 2. This is likely due to the more complicated temporal trend in setting 2. The slightly larger variations when the sample size of 100 compared to those with a sample size of 60 is likely due to the complexity of setting 2, sampling errors in the process of MC replicates generations, and relatively closeness between the two sample sizes. Higher mean sensitivities and specificities for samples of 100 indicate that most MC replicates result in high values in those two statistics, but some MC replicates produce relatively lower sensitivities and specificities, which causes the larger variations. This type of uncertainty is clearly improved when the sample size is increased from 100 to 400 as seen in Table 2.

We further assess the fitness of the estimated patterns to the true patterns. One hundred MC replicates each with sample size 400 under setting 2 (two clusters one with a linear pattern and the other a quadratic pattern) are used in this assessment. Fig. 2 displays the average of fitted curves for the background and each cluster along with 95% empirical confidence bands, based on 100 MC replicates. The variations of fitted curves are not clearly shown in Fig. 2 due to the wide range in $f_1(\cdot)$ and $f_2(\cdot)$. To demonstrate the variations, Fig. 3 is a magnifier plot of distances from the fitted curve and the confidence bands to the true curve, respectively, for the temporal pattern in cluster 1 in time interval $(1, 2)$. When sample size is smaller, the distances are larger but overall the true cluster patterns are well estimated even with a sample size of 60.

Comparisons with other methods: We compared the proposed method with two approaches. The first approach is a non-parametric clustering method applied to longitudinal data and is implemented in an R package `kml` (Genolini and Falissard, 2011). It uses the K-means method for clustering such that subjects with similar profiles over time are grouped together. We use data simulated under scenario 2 with sample sizes of 100 and 400 to compare the performance of the proposed method and the method in `kml`. The results are included in the parentheses of Table 3. Overall, regardless of the sample size, sensitivities and specificities from the proposed method are higher than those from the non-parametric approach.

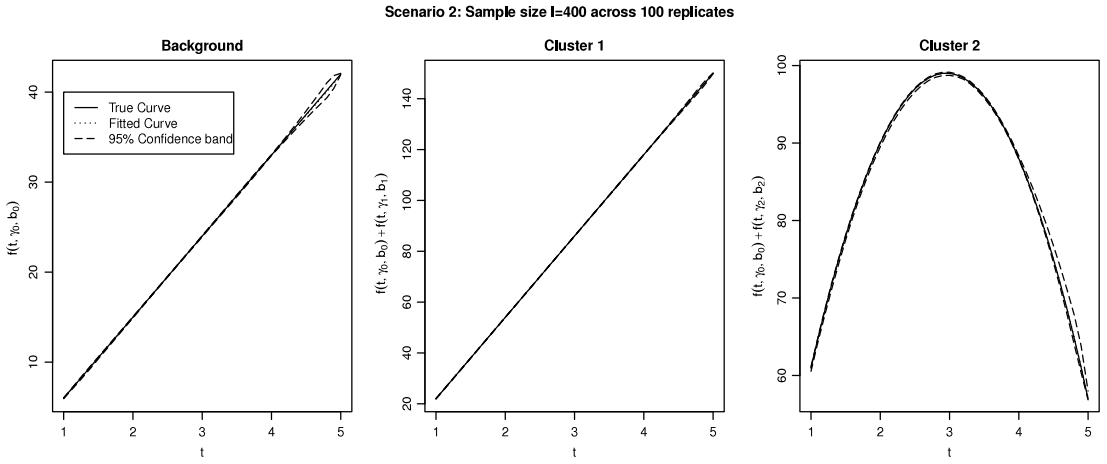

Fig. 2. True temporal curves, fitted curves and 95% confidence bands over time across 100 replicates when sample size $I = 400$ in setting 2.

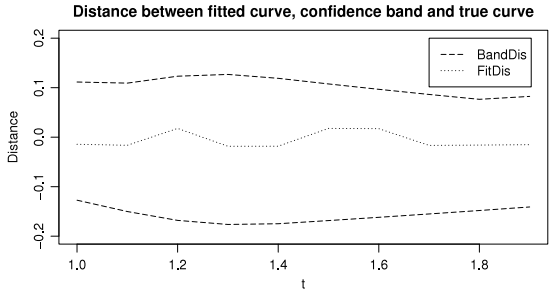

Fig. 3. Distance between fitted curves, confidence bands and the true curve (cluster 1 in time interval (1, 2) under setting 2 for $I = 400$). BanDis: distance between empirical confidence bands and the true curve; FitDis: distance between the fitted curve and the true curve.

Table 2
Summary of sensitivity and specificity across 100 MC replicates for varying sample sizes under setting 2. SD stands for standard deviation. 100* is the unbalanced case where 10 subjects are in cluster 1, 30 subjects are in cluster 2 and the rest 60 are in background. All other cases have equal number of subjects in clusters 1 and 2.

			Sample Size			
			60	100	100*	400
Background	Sensitivity	Mean	1.0000	1.0000	0.9994	1.0000
		SD	0.0000	0.0000	0.0039	0.0000
	Specificity	Mean	1.0000	1.0000	1.0000	1.0000
		SD	0.0000	0.0000	0.0000	0.0000
Cluster 1	Sensitivity	Mean	0.2737	0.7874	0.6084	1.0000
		SD	0.0400	0.3237	0.3781	0.0000
	Specificity	Mean	0.7737	0.9394	0.9816	0.9997
		SD	0.0400	0.1016	0.0179	0.0010
Cluster 2	Sensitivity	Mean	0.5304	0.8421	0.8359	0.9991
		SD	0.0278	0.2193	0.0819	0.0028
	Specificity	Mean	0.8484	0.9617	1.0000	1.0000
		SD	0.0279	0.0685	0.0000	0.0000

The method in `kml` only performs cluster analyses and does not consider patterns in the background, which might be the reason of inferior results. To further assess the advantage of accounting for background, we use the proposed method but without adjusting for background, that is, we only perform cluster analyses. The 100 MC replicates generated under scenario 2 with sample size of 400 are used in this comparison. The results are displayed in the last column of Table 3. Again, both clustering sensitivity and specificity suffer from the absence of background adjustment. Observations from these two comparisons demonstrate the need of adjusting for background patterns in order to improve the quality of unique clusters.

Further assessment: In all the above analyses, parameter a in the multivariate truncated normal distribution was set at 0.0001. Our further simulations with larger values of a, e.g., $a = 0.01$, give results (not shown) comparable to those in Tables 1 and 2, implying that the underlying unique clusters are potentially quite different from the background. We suggest the selection of a rely on users' belief in the strength of "signal" instead of imposing an arbitrary choice. If no background patterns are

Table 3
Comparison between the proposed method and the method in `kml` in terms of sensitivity and specificity across 100 MC replicates and the results without adjusting background under setting 2. SD stands for standard deviation. There are equal number of subjects in clusters 1 and 2, making one half number of subjects and the rest half are in background. Results from `kml` are included in parentheses.

	Sample	Size	Proposed method (vs. `kml`)		No adjustment
			100	400	400
Background	Sensitivity	Mean	1.0000 (0.4358)	1.0000 (0.9350)	0.9990
		SD	0.0000 (0.0783)	0.0000 (0.0711)	0.0028
	Specificity	Mean	1.0000 (0.9171)	1.0000 (0.9407)	0.9929
		SD	0.0000 (0.0796)	0.0000 (0.0377)	0.0081
Cluster 1	Sensitivity	Mean	0.7874 (0.7920)	1.0000 (0.9500)	0.7930
		SD	0.3237 (0.1818)	0.0000 (0.0848)	0.2481
	Specificity	Mean	0.9394 (0.9440)	0.9997 (0.9851)	0.9878
		SD	0.1016 (0.0466)	0.0010 (0.0237)	0.0103
Cluster 2	Sensitivity	Mean	0.8421 (0.5896)	0.9991 (0.8855)	0.7937
		SD	0.2193 (0.3548)	0.0028 (0.1538)	0.2380
	Specificity	Mean	0.9617 (0.9054)	1.0000 (0.9680)	0.9939
		SD	0.0685 (0.0702)	0.0000 (0.0369)	0.0054

Table 4
Summary statistics of misclassification rates of subjects with partial missing time points over 100 MC replicates in setting 2 with sample size of 400. SD stands for standard deviation.

		10% missingness	20% missingness
Background	Mean	0.0000	0.0050
	SD	0.0000	0.0500
Cluster 1	Mean	0.0095	0.0125
	SD	0.0221	0.0534
Cluster 2	Mean	0.0000	0.0000
	SD	0.0000	0.0000

expected, then setting $\omega = 1$ in the prior distribution of $\boldsymbol{\theta}_i$, expression (4), brings the method back to standard clustering in the Bayesian framework. Besides the assessment in *a*, we further performed a set of sensitivity analyses, including sensitivity with respect to unbalanced data, missingness, and signal-noise ratio.

It is known that many existing clustering methods prefer clusters with similar sizes (e.g., similar numbers of subjects across different clusters). To assess whether the proposed approach is sensitive to unbalanced clusters, for $I = 100$, we include 10 subjects in cluster 1, 30 subjects in cluster 2 and the remaining in background. Summary statistics of sensitivities and specificities across 100 MC replicates for settings 1 and 2 are in Tables 1 and 2 (the column indicated by 100*). Compared to the results from data with balanced clusters, the variations of sensitivities and specificities in this unbalanced situation are slightly larger, but still quite small compared to the mean statistics.

To test the performance of the method on data set with missing values, we choose q% subjects from clusters 1 and 2 and background, respectively. For these selected subjects, samples at time points 1 and 3 are deleted. Then these data are used to infer the patterns and perform cluster analyses. In this assessment, we use scenario 2 with sample size of 400, q takes two values, $q = 10, 20$, and for each q, 100 MC replicates are simulated. A misclassification rate for those subjects with missing values is recorded for each MC replicate, and the results are summarized in Table 4. Apparently, more subjects with missing values lead to higher misclassification rates. However, even when 20% subjects with missing values, most of these individuals are still correctly grouped together, which implies that the proposed method has the potential to handle missingness as long as the percentage of subjects with missing data is small or moderate.

To assess how the method performs for lower signal to noise ratios, we increased the "noise" (i.e., variance in the random error) with "signal" (i.e., the pattern in the background and each unique cluster) intact. Two directions are considered in this assessment. In the first direction, we increase the variance only in the background from $\tau = 0.25$ to 1, and in the second direction, we increase the variances both in the background and in the unique clusters from $\tau = 0.25$ to 1. Other settings are the same as in scenario 2. As before, 100 MC replicates each with a sample size of 400 are used in this assessment. The results are summarized in Table 5 indicated as "Case 1" and "Case 2", where high sensitivities and specificities are observed and are comparable to those in Table 2. Overall, these simulations demonstrate that the proposed method has the ability to handle data that are unbalanced, with missing values, or with low signal-noise ratios.

3.2. Real data applications

We applied the proposed method to the motivating example to identify background and unique temporal patterns with respect to wheal sizes in reaction to grass allergens. The information was collected in a birth cohort on the Isle of Wight in the United Kingdom. Details of the cohort and related information on data collection can be found elsewhere (Arshad and

Table 5
Sensitivity and specificity over 100 MC replicates in setting 2 with sample size of 400 in three additional cases. SD stands for standard deviation. Case 1: Increased variance in background only, i.e. $\tau = 1$; Case 2: Increased variance in both background and two unique clusters, i.e., $\tau = 1$.

			Case 1	Case 2
Background	Sensitivity	Mean	1.0000	1.0000
		SD	0.0000	0.0000
	Specificity	Mean	1.0000	1.0000
		SD	0.0000	0.0000
Cluster 1	Sensitivity	Mean	0.9833	0.9708
		SD	0.1054	0.0144
	Specificity	Mean	0.9994	0.9923
		SD	0.0015	0.0428
Cluster 2	Sensitivity	Mean	0.9821	0.9730
		SD	0.1053	0.1299
	Specificity	Mean	1.0000	1.0000
		SD	0.0000	0.0000

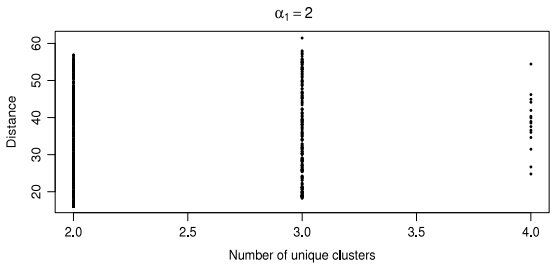

Fig. 4. Varying distance with different unique clusters (i.e., background excluded) when $\alpha_1 = 2$.

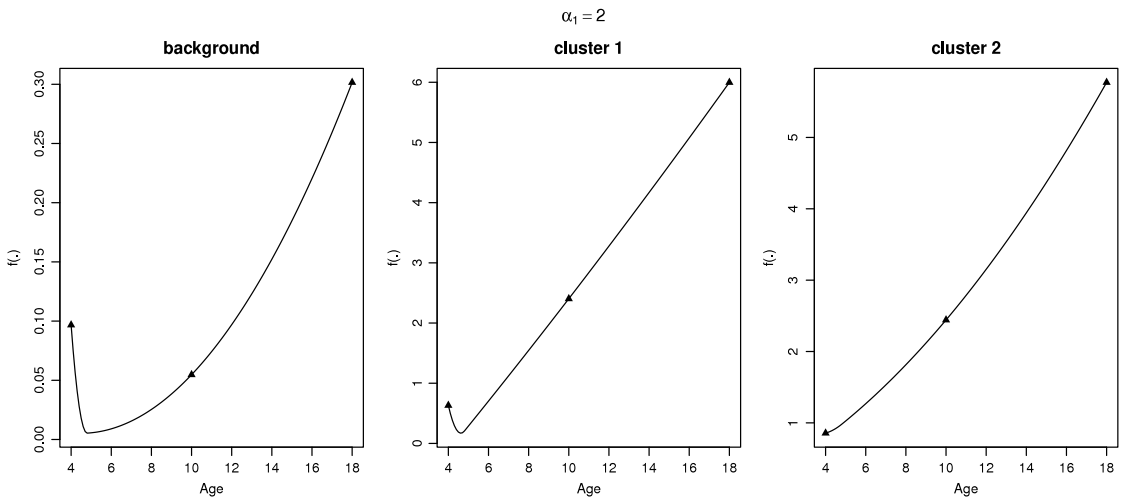

Fig. 5. Inferred temporal patterns in background with 493 subjects and 2 unique clusters with the cluster 1 of 1 subject and cluster 2 of 130 subjects.

Hide, 1992). In this analysis, wheal sizes of 624 subjects in reaction to grass allergens at three time points (ages 4, 10, and 18) are included in the analyses.

We ran one but long MCMC chain than in simulations, in total, 9000 iterations with the first 6000 as burn-in. The next 2000 are used to estimate the probability matrix and the last 1000 iterations to infer the clusters and parameters. As in simulations, the precision parameter is chosen via a grid search by minimizing DIC, and the inferences are made based on MCMC samples corresponding to the selected precision parameter.

In this application, the precision parameter is selected as 2 which achieves the minimum DIC. Based on the last 1000 iterations, distances for different clusters are plotted (Fig. 4). The minimum distances at different numbers of clusters are not far from each other, but overall minimum distance across all possible cases is reached at 2 clusters (besides background). Out of the 624 subjects, 493 subjects only have a background pattern, and the remaining 131 subjects are divided into two clusters with one cluster of size 1 (Fig. 5). Although in general wheal size increases after the age of 4 years, the wheal size of

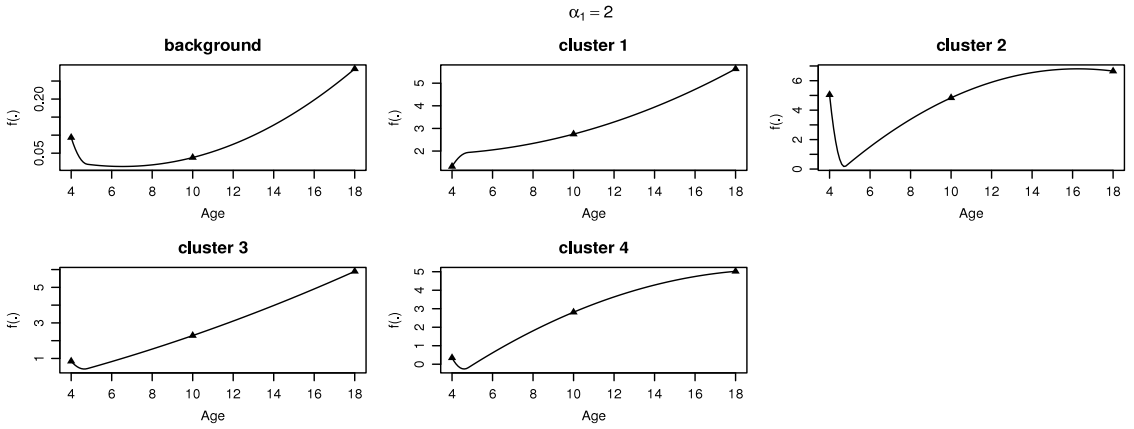

Fig. 6. Inferred temporal patterns in background with 493 subjects and 4 unique clusters. There is 1 subject in cluster 1, 3 subjects are in cluster 2, 103 subjects in cluster 3 and the remaining 24 subjects are in cluster 4.

subjects in the two clusters increases much faster than in the background (and recall that the unique patterns in clusters are added onto the background). Our closer observation at the clusters found that grouping of subjects into the background is consistent across all three clustering efforts (2, 3, and 4 clusters); regardless of the numbers of unique patterns, the background always includes the same 493 subjects. In terms of unique patterns, the clustering patterns in 3 clusters are quite similar to the patterns in 2 clusters and the additional cluster only has 2 subjects. In the case of 4 clusters, although the minimum Euclidean distance for 4 clusters is not the overall minimum across all clustering situations obtained, it is not far from the overall minimum. A discussion with our collaborators in epidemiology and clinical studies indicated that the patterns in the 4 clusters are of great interest. To further understand the clustering patterns clinically, we linked the 4 clusters (Fig. 6) to asthma status over time due to the close relationship between allergy and asthma (Soto-Ramirez et al., 2013). Among all the 493 subjects with background pattern only, about 76% are asthma free. For subjects in cluster 3, 50% suffered at least one time of asthma from ages 4 to 18, and about 20% of subjects had persistent asthma (i.e., asthma never disappeared). For subjects in cluster 4, 37% had at least asthma once from ages 4 to 18, and around 10% had persistent asthma. These findings are consistent with the wheal size temporal patterns, that is, faster increase of wheal size over time (cluster 3) indicating more difficulty in recovering from asthma (i.e., asthma persists) and slowing increasing (cluster 4) providing a hope of asthma curing in the future. For the two clusters with small cluster sizes (clusters 1 and 2), the wheal sizes are large on average with cluster 2 having a much larger mean wheal size at each time point (Fig. 6). However, most of these subjects (3 out of 4) are asthma free, which may deserve a further investigation. It is worth pointing out that since the patterns in these 4 clusters are not associated with the minimum Euclidean distance, the interpretation of the 4 clusters should be implemented with caution.

4. Summary and discussion

We proposed a semi-parametric method to cluster subjects based on their unique temporal patterns after adjusting for a background pattern. *P*-splines are used to describe the temporal pattern in the background and also for subjects with unique patterns. To differentiate between unique patterns and background patterns, we utilized a mixture of the Dirichlet process and a point mass function to achieve this. The proposed method can be easily simplified to fit non-longitudinal data.

Different simulation scenarios were applied to demonstrate and assess the method in terms of its sensitivity and specificity in cluster identification and pattern estimations. High sensitivity and specificity were observed across all simulations. Our comparison with an existing non-parametric clustering approach for longitudinal data also demonstrated the advantage of the proposed method and the need of taking background pattern into consideration. Besides this non-parametric method, a parametric approach for clustering longitudinal data is available. It was proposed by McNicholas and Murphy (2010). However, this approach may have a focus different from that in the proposed method. It clusters overall means (i.e., to infer overall mean μ_g for cluster g as noted in the article) from longitudinal data, where a modified Cholesky decomposition is implemented to ease the process of clustering. Their clustering approach cannot be directly applied to the clustering of temporal patterns, and thus was not included in the comparisons.

Furthermore, in real data applications, the estimated unique temporal patterns in wheal sizes were sensible and informative, which clearly linked allergic sensitization patterns to different diseases statuses over time. That is, a subject with faster increase in wheal size over time is more likely to result in persistent asthma than a subject with slower increase. This finding has the potential to assist clinicians to predict allergic diseases based on earlier allergic sensitization temporal patterns and gives a potential to prevent the disease from occurring at an earlier stage. In this work, background is defined as the largest cluster with most subjects. This definition is reasonable in the context of its application since more subjects

are expected to be in the general population or background. Other definition of background can also be used and should be guided by study questions.

In this study, the cluster analysis is designed for one variable and in the application, applied to one allergen, grass pollen. It is possible that multiple allergens may behave similarly. The predictability will be improved if we can combine evidence from different allergens. Thus, there is a need to cluster variables (e.g., allergens) in addition to clustering subjects. However, the variables may be correlated, and taking correlations into account potentially increases the complexity of the modeling, which warrants further investigations.

Acknowledgments

The research work is supported by National Institutes of Health research fund, R21 AI099367, Hongmei Zhang (PI), R01 AI121226, Hongmei Zhang (MPI), and R01 AI091905, Wilfried Karmaus (PI).

Appendix. Derivations of conditional posterior probabilities

- The conditional posterior distribution of $\boldsymbol{\theta_0}$.

$$P(\boldsymbol{\theta_0}|(\cdot)) \propto \exp\left\{-\frac{1}{2}\boldsymbol{\theta_0}^T{\boldsymbol{\Sigma}_{\boldsymbol{\theta 0}}}^{-1}\boldsymbol{\theta_0}\right\} \times \exp\left\{-\frac{1}{2}\sum_{i=1}^{I}(\boldsymbol{Y_i}-\boldsymbol{M_i})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{Y_i}-\boldsymbol{M_i})\right\},$$

where $(\cdot)$ denote data and all other variables, $\boldsymbol{\theta_0}$ has three independent components, $\boldsymbol{\beta_0}$, $\boldsymbol{\gamma_0 b_0}$. Since $\boldsymbol{\Sigma_{\theta 0}}$ is a diagonal matrix, let $\boldsymbol{\Sigma_{\theta 0}} = diag\{\boldsymbol{\Sigma}_{\beta_0}, \boldsymbol{\Sigma}_{\boldsymbol{\gamma_0}}, \boldsymbol{\Sigma}_{\boldsymbol{b_0}}\}$, $\boldsymbol{R_i}(\boldsymbol{\beta_0}) = \boldsymbol{Y}_i - f(\boldsymbol{t}_i; \boldsymbol{\gamma_0}, \boldsymbol{b}_0) - \boldsymbol{X}_i\boldsymbol{\beta_i} - f(\boldsymbol{t}_i; \boldsymbol{\gamma_i}, \boldsymbol{b}_i)$ with $\boldsymbol{\beta}_0$ excluded. In the following, $\boldsymbol{R}_i(\Phi)$ denotes an expression excluding parameter Φ. The conditional posterior of $\boldsymbol{\beta_0}$ is proportional to

$$\begin{aligned}
&\exp\left\{-\frac{1}{2}\boldsymbol{\beta_0}^T{\boldsymbol{\Sigma}_{\boldsymbol{\beta_0}}}^{-1}\boldsymbol{\beta_0}\right\} \times \exp\left\{-\frac{1}{2}\sum_{i=1}^{I}(\boldsymbol{R_i}(\boldsymbol{\beta_0}) - \boldsymbol{X}_i\boldsymbol{\beta_0})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{R_i}(\boldsymbol{\beta_0}) - \boldsymbol{X}_i\boldsymbol{\beta_0})\right\} \\
&= \exp\left\{-\frac{1}{2}\boldsymbol{\beta_0}^T\left(\boldsymbol{\Sigma}_{\boldsymbol{\beta_0}}{}^{-1} + \sum_{i=1}^{I}\boldsymbol{X}_i^T\boldsymbol{\Sigma}\right)^{-1}\boldsymbol{X}_i\boldsymbol{\beta_0} + \left[\frac{1}{2}\sum_{i=1}^{I}\boldsymbol{R_i}\boldsymbol{\beta_0}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X}_i\right]\boldsymbol{\beta_0}\right. \\
&\quad \left. + \boldsymbol{\beta_0}^T\left[\frac{1}{2}\sum_{i=1}^{I}\boldsymbol{X}_i^T\boldsymbol{\Sigma}^{-1}\boldsymbol{R_i}(\boldsymbol{\beta_0})\right]\right\} + c.
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\boldsymbol{\beta_0}|(\cdot) \sim MN(\boldsymbol{\mu}, \boldsymbol{\Delta}), \quad \text{with} \\
&\boldsymbol{\mu} = \boldsymbol{\Delta}\sum_{i=1}^{I}\boldsymbol{X}_i^T\boldsymbol{\Sigma}^{-1}\boldsymbol{R_i}(\boldsymbol{\beta_0}), \\
&\boldsymbol{\Delta} = \left(\boldsymbol{\Sigma}_{\boldsymbol{\beta_0}}{}^{-1} + \sum_{i=1}^{I}\boldsymbol{X}_i^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X}_i\right)^{-1}.
\end{aligned} \tag{A.1}$$

Similarly, the posterior probability of $\boldsymbol{\gamma_0}$ is proportional to

$$\exp\left\{-\frac{1}{2}\boldsymbol{\gamma_0}^T{\boldsymbol{\Sigma}_{\boldsymbol{\gamma_0}}}^{-1}\boldsymbol{\gamma_0} - \frac{1}{2}\sum_{i=1}^{I}(\boldsymbol{R_i}(\boldsymbol{\gamma_0}) - (\boldsymbol{I_{T\times T}} \otimes \boldsymbol{\gamma_0^T})\boldsymbol{T_i})^T\boldsymbol{\Sigma}^{-1} \times (\boldsymbol{R_i}(\boldsymbol{\gamma_0}) - (\boldsymbol{I_{T\times T}} \otimes \boldsymbol{\gamma_0^T})\boldsymbol{T_i})\right\}$$

where $\otimes$ is an outer product, $\boldsymbol{R_i}(\boldsymbol{\gamma_0}) = \boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\beta_0} - (\boldsymbol{b_0}^T \triangledown \boldsymbol{T}_{\boldsymbol{1}*}^{(\boldsymbol{i})}, \ldots, \boldsymbol{b_0}^T \triangledown \boldsymbol{T}_{\boldsymbol{T}*}^{(\boldsymbol{i})})^T - \boldsymbol{X}_i\boldsymbol{\beta_i} - f(\boldsymbol{t}_i; \boldsymbol{\gamma_i}, \boldsymbol{b}_i)$, $\boldsymbol{T_i} = (\boldsymbol{T_{i1}}, \ldots, \boldsymbol{T_{iT}})^T$; $\triangledown \boldsymbol{T}_{\boldsymbol{l}*}^{(\boldsymbol{i})} = ((t_{il} - t_{i1}^*)_+^2, (t_{il} - t_{i2}^*)_+^2, \ldots, (t_{il} - t_{iN}^*)_+^2)$, $l = 1, 2, \ldots, T$. This leads to

$$\begin{aligned}
&\gamma_{0j}|(\cdot) \sim N\left(\frac{\Pi}{2}\left(\Delta + \frac{1}{\sigma_{\gamma_{0i}}^2}\right), \frac{1}{\Delta + \frac{1}{\sigma_{\gamma_{0i}}^2}}\right), \quad \text{with} \\
&\Delta = \sum_{i=1}^{I}\sum_{t=1}^{T}\sum_{k=1}^{T}\Sigma_{kt}^{-1}t_{ik}^{2j}, \\
&\Pi = \sum_{i=1}^{I}\sum_{t=1}^{T}\sum_{k=1}^{T}(R_i(\gamma_{0i})_k + R_i(\gamma_{0i})_t)\Sigma_{kt}^{-1}t_{ik}^{j}, \\
&j = 0, 1, 2.
\end{aligned}$$

Similarly,

$$b_{0j}|(\cdot) \sim N\left(\frac{\Pi}{2}\left(\Delta + \frac{1}{\sigma^2_{b_{0j}}}\right), \frac{1}{\Delta + \frac{1}{\sigma^2_{b_{0j}}}}\right), \quad \text{with}$$
$$\Delta = \sum_{i=1}^{I}\sum_{t=1}^{T}\sum_{k=1}^{T} \Sigma^{-1}_{kt}(t_{ik} - t_{ij})^4_+,$$
$$\Pi = \sum_{i=1}^{I}\sum_{t=1}^{T}\sum_{k=1}^{T} (R_i(b_{0j})_k + R_i(b_{0j})_t)\,\Sigma^{-1}_{kt}(t_{ik} - t_{ij})^2_+,$$
$$j = 1, 2, \ldots, N,$$

where $R(\Phi)_k$ is the kth element of $\boldsymbol{R}(\Phi)$, and similar definition applied to Σ^{-1}_{kt}.

- Following a similar way as in the derivation of conditional posterior distributions of $\boldsymbol{\theta}_0$, conditional posterior distributions for $\boldsymbol{\theta_i}$ can be obtained. In the following, we list the conditional posterior distributions used to update the unique values for $\boldsymbol{\theta}_i$'s in the DP conditional on $\omega = 1$. As for $\boldsymbol{\beta_i}$,

$$\boldsymbol{\beta_i}|(\cdot) \sim \frac{1}{I - 1 + \alpha_1}\sum_{j\neq i}\delta(\boldsymbol{\beta_j}) + \frac{\alpha_1}{I - 1 + \alpha_1} TMVN(\boldsymbol{\mu}, \boldsymbol{\Delta}, a), \quad \text{with}$$
$$\boldsymbol{\mu} = \boldsymbol{\Delta}[{\boldsymbol{\Sigma}_{\boldsymbol{\beta_i}}}^{-1}\boldsymbol{\mu}_{\boldsymbol{\beta_i}} + \sum_{i=1}^{I}\boldsymbol{X}_i^T\boldsymbol{\Sigma}^{-1}\boldsymbol{R_i}(\boldsymbol{\beta_i})],$$
$$\boldsymbol{\Delta} = \left(\boldsymbol{\Sigma}^{-1}_{\beta_i} + \sum_{i=1}^{I}\boldsymbol{X}_i^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X}_i\right)^{-1},$$

where $\boldsymbol{R_i}(\boldsymbol{\beta_i}) = \boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\beta_0} - f(\boldsymbol{t}_i; \boldsymbol{\gamma_0}, \boldsymbol{b}_0) - f(\boldsymbol{t}_i; \boldsymbol{\gamma_i}, \boldsymbol{b}_i)$.
For $\boldsymbol{\gamma_i}$,

$$\gamma_{ih}|(\cdot) \sim \frac{1}{I - 1 + \alpha_1}\sum_{j\neq i}\delta(\gamma_{jh}) + \frac{\alpha_1}{I - 1 + \alpha_1} TMVN(\mu, \Delta, a), \quad \text{with}$$
$$\mu = \Delta\left(\frac{\Pi}{2} + \frac{\mu_{\gamma_{ih}}}{\sigma^2_{\gamma_{ih}}}\right),$$
$$\Delta = \frac{1}{\Phi + \frac{1}{\sigma^2_{\gamma_{ih}}}},$$
$$\Phi = \sum_{j:\ \text{subject } j \text{ has } \boldsymbol{\theta_i}}\sum_{t=1}^{T}\sum_{k=1}^{T}\Sigma^{-1}_{kt}t^{2h}_{ik},$$
$$\Pi = \sum_{i=1}^{I}\sum_{t=1}^{T}\sum_{k=1}^{T}(R_i(\gamma_{ih})_k + R_i(\gamma_{ih})_t)\,\Sigma^{-1}_{kt}t^h_{ik},$$
$$h = 0, 1, 2.$$

Finally, for $\boldsymbol{b_i}$,

$$b_{ih}|(\cdot) \sim \frac{1}{I - 1 + \alpha_1}\sum_{j\neq i}\delta(b_{jh}) + \frac{\alpha_1}{I - 1 + \alpha_1} TMVN(\mu, \Delta, a), \quad \text{with}$$
$$\mu = \Delta\left(\frac{\Pi}{2} + \frac{\mu_{b_{ih}}}{\sigma^2_{b_{ih}}}\right),$$
$$\Delta = \frac{1}{\Phi + \frac{1}{\sigma^2_{b_{ih}}}},$$
$$\Phi = \sum_{j:\ \text{subject } j \text{ has } \boldsymbol{\theta_i}}\sum_{t=1}^{T}\sum_{k=1}^{T}\Sigma^{-1}_{kt}(t_{ik} - t_{ih})^4_+,$$

$$\Pi = \sum_{i=1}^{I}\sum_{t=1}^{T}\sum_{k=1}^{T}(R_i(b_{ih})_k + R_i(b_{ih})_t)\,\Sigma_{kt}^{-1}(t_{ik} - t_{ih})_+^2,$$
$$h = 1, 2, \ldots, N.$$

References

Antoniak, C.E., 1974. Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. Ann. Statist. 2, 1152–1174.
Arshad, S.H., Hide, D.W., 1992. Effect of environmental factors on the development of allergic disorders in infancy. J. Allergy Clin. Immunol. 90, 235–241.
Baladandayuthapani, V., Mallick, B.K., Carroll, R.J., 2005. Spatially adaptive Bayesian penalized regression splines (p-splines). J. Comput. Graph. Statist. 14, 378–394.
Bigelow, J.L., Dunson, D.B., 2009. Bayesian semiparametric joint models for functional predictors. J. Amer. Statist. Assoc. 104, 26–36.
Caron, F., Teh, Y.W., Murphy, T.B., 2014. Bayesian nonparametric Plackett-Luce models for the analysis of preferences for college degree programmes. Ann. Appl. Stat. 8, 1145–1181.
Dahl, D.B., 2006. Model-based clustering for expression data via a Dirichlet process mixture model. In: Do, Kim-Anh, Müller, Peter, Vannucci, Marina (Eds.), Bayesian Inference for Gene Expression and Proteomics. Cambridge University Press.
Dorazio, R.M., 2009. On selecting a prior for the precision parameter of Dirichlet process mixture models. J. Statist. Plann. Inference 139, 3384–3390.
Dunson, D.B., Herring, A.H., Engel, S.M., 2008. Bayesian selection and clustering of polymorphisms in functionally related genes. J. Amer. Statist. Assoc. 103, 534–546.
Efron, B., Tibshirani, R., Storey, J.D., Tusher, V., 2001. Empirical bayes analysis of a microarray experiment. J. Amer. Statist. Assoc. 96, 1151–1160.
Eilers, P.H.C., Marx, B.D., 1996. Flexible smoothing with b-splines and penalties. Statist. Sci. 11, 89–121.
Escobar, M.D., West, M., 1995. Bayesian density estimation and inference using mixtures. J. Amer. Statist. Assoc. 90, 577–588.
Ferguson, T.S., 1973. A Bayesian analysis of some nonparametric problems. Ann. Statist. 1, 209–230.
Fraley, C., Raftery, A.E., 2002. Model-based clustering, discriminant analysis, and density estimation. J. Amer. Statist. Assoc. 97, 611–631.
Gelman, A., Carlin, J.B., Stern, H.S., Rubin, D.B., 2003. Bayesian Data Analysis, second ed.. Chapman and Hall/CRC.
Genolini, C., Falissard, B., 2011. Kml: A package to cluster longitudinal data. Comput. Methods Programs Biomed. 104, e112–e121.
George, E.I., McCulloch, R.E., 1997. Appraches for Bayesian variable selection. Statist. Sinica 7, 339–373.
Green, P.J., 1995. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika 82 (4), 711–732.
Kim, S., Dahl, D.B., Vannucci, M., 2009. Spiked Dirichlet process prior for Bayesian multiple hypothesis testing in random effects models. Bayesian Anal. 4, 707–732.
Kim, S.Y., Lee, J.W., Bae, J.S., 2006. Effect of data normalization on fuzzy clustering of DNA microarray data. BMC Bioinformatics 7, 134.
McNicholas, P.D., Murphy, T.B., 2010. Model-based clustering of longitudinal data. Canad. J. Statist. 38, 153–168.
Murugiah, S., Sweeting, T., 2012. Selecting the precision parameter prior in Dirichlet process mixture models. J. Statist. Plann. Inference 142, 1947–1959.
Neal, R.M., 2000. Markov chain sampling methods for dirichlet process mixture models. J. Comput. Graph. Statist. 9, 249–265.
Nieto-Barajas, L.E., Contreras-Cristan, A., 2014. A Bayesian nonparametric approach for time series clustering. Bayesian Anal. 9, 147–170.
Qin, L., Self, S., 2006. The clustering of regression models method with applications in gene expression data. Biometrics 62, 526–533.
Ritter, C., Tanner, M.A., 1992. Facilitating the gibbs sampler: The gibbs stopper and the griddy-gibbs sampler. J. Amer. Statist. Assoc. 87, 861–868.
Scott, J.G., 2009. Nonparametric Bayesian multiple testing for longitudinal perforance stratification. Ann. Appl. Stat. 3, 1655–1674.
Soto-Ramirez, N., Arshad, S.H., Holloway, J.W., Zhang, H., Schauberger, E., Ewart, S., Patil, V., Karmaus, W., 2013. The interaction of genetic variants and DNA methylation of the interleukin-4 receptor gene increase the risk of asthma at age 18 years. Clin. Epigenet. 5, 1–8.
Spiegelhalter, D.J., Best, N.G., Carlin, B.P., van der Linde, A., 2014. The deviance information criterion: 12 years on. J. R. Stat. Soc. Ser. B Stat. Methodol. 76, 485–493.
Stephens, M., 2000. Bayesian analysis of mixture models with an unknown number of components -an alternative to reversible jump methods. Ann. Statist. 28, 40–74.
Zhang, H., Ghosh, K., Ghosh, P., 2012. Sampling designs via a multivariate hypergeometric-Dirichlet process model for a multi-species assemblage with unknown heterogeneity. Comput. Statist. Data Anal. 56, 2562–2573.