Joint clustering with correlated variables

Hongmei Zhang *
School of Public Health, The University of Memphis, Memphis, TN
Yubo Zou
Blue Cross Blue Shield of South Carolina, Columbia, SC
Will Terry, Wilfried Karmaus
School of Public Health, The University of Memphis, Memphis, TN
and
Hasan Arshad
University of Southampton Faculty of Medicine, Southampton, UK

November 21, 2017

Abstract

Traditional clustering methods focus on grouping subjects or (dependent) variables assuming independence between the variables. Clusters formed through these approaches can potentially lack homogeneity. This article proposes a joint clustering method by which both variables and subjects are clustered. In each joint cluster (in general composed of a subset of variables and a subset of subjects), there exists a unique association between dependent variables and covariates of interest. To this end, a Bayesian method is designed, in which a semi-parametric model is used to evaluate any unknown relationships between possibly correlated variables and covariates of interest, and a Dirichlet process is utilized to cluster subjects. Compared to existing clustering techniques, the major novelty of the method exists in its ability to improve the homogeneity of clusters, along with the ability to take the correlations between variables into account. Via simulations, we examine the performance and efficiency of the proposed method. Applying the method to cluster allergens and subjects based on the association of wheal size in reaction to allergens with age, we found that a certain pattern of allergic sensitization to a set of allergens has a potential to reduce the occurrence of asthma.

Keywords: Bayesian methods, Dirichlet process, Semi-parametric modeling

*Corresponding author. This work is supported by the National Institute of Allergy and Infectious Diseases (grant numbers R21AI099367 [PI: H Zhang], R01AI121226 [MPI: H Zhang and J Holloway]). The authors acknowledge the Computational Research and Cyber Infrastructure Support Initiative at the University of South Carolina, and the High Performance Computing at the University of Memphis for providing the computing resources that contributed to the simulation results of this paper.

1 Introduction

The work presented in the article was motivated by an epidemiological effort to study patterns of wheal sizes in reaction to different allergens at different ages and how those patterns are associated with asthma risk. There exist various allergens in our daily lives such as pollens or peanuts. Being sensitized to an allergen increases the probability of asthma incidence. Some people are allergic to certain allergens but never develop asthma, while others experience asthma incidence at certain ages. In addition, asthma remissions are observed. It is postulated that asthma incidence and asthma remission are linked to particular allergic sensitization patterns at different ages. Specifically, there is a desire to sort out whether there exist groups of subjects such that in each group their reaction to certain allergens was different from their reaction to other allergens, and also different from subjects in other groups. A cluster analysis is commonly taken as an attempt to achieve such a goal. This type of analyses has become increasingly popular in areas of epidemiological research and genetic or epigenetic studies to identify patterns of clinical phenotypes of different health outcomes, or genetic/epigenetic patterns potentially associated with an outcome of interest.

Classical methods for cluster analysis generally focus on clustering subjects or variables (e.g., allergens) but not both. Non-parametric approaches such as the k-means and hierarchical methods are commonly utilized to cluster subjects. These methods detect the homogeneity in subjects, but not in variables. To achieve both, before conducting cluster analyses, some studies perform principal component analyses (PCA) or factor analyses (FA) on the variables to identify inherent homogeneity. The advantage of PCA and FA is their ability to address the interdependence between phenotypes. However, they cannot explain any external variable effects, such as time effects, on the formation of patterns. This is crucial when there is a need to explore patterns in variables, e.g., changes of allergic sensitization with ages (natural history).

The concept of biclustering has been more recognized recently. It was dated back to the 1970's (Hartigan, 1972), and later implemented to explore gene expression/microarray data (Cheng and Church, 2000). The biclustering scheme simultaneously clusters two-dimensional gene expression data and tries to optimize a pre-specified objective function.

There are two main classes of biclustering algorithms: systematic search algorithms and stochastic search algorithms, while each class of algorithm has several different ways to be implemented (Freitas et al., 2013). Various biclustering tools and methods are available: bicluster analysis in R (Kaiser and Leisch, 2008), BiVisu (Cheng et al., 2007), GEMS (Wu and Kasif, 2005), BicOverlapper (Santamaría et al., 2008), e-CCC-Biclustering (Madeira and Oliveira, 2009), parametric Bayesian BiClustering model (BBC) (Gu and Liu, 2008), as well as non-parametric Bayesian methods (Meeds and Roweis, 2007; Lee et al., 2013). The existing biclustering concept in these works considers the coherence of rows and columns in the data. Since the technique is not model-based, it is restricted to profiles in the variables and external variables do not have any contribution to the evaluation of similarity between different variables. Furthermore, some biclustering methods perform cluster analyses on the rows and columns separately, and do not simultaneously consider the interrelationship between the rows and columns. Most importantly, in our application (i.e., sensitization to allergens), sensitization measures are dependent, e.g., a person allergic to cat dander is likely to be allergic to dog dander as well. However, existing methods overlook the correlations between the clustering variables, which can potentially cause mis-clustering.

In this article, we propose a probabilistic clustering method, denoted as joint clustering, which takes into account the correlations between variables (e.g., sensitization measures) and the interrelationship between variables and subjects. The clusters are formed by consistent associations between a variable (or a "dependent variable") and covariates of interests among a subset of subjects for a set of variables. Each joint cluster is composed of a certain numbers of variables and a subset of subjects. To evaluate possibly non-linear associations between variables and covariates, a semi-parametric model via penalized splines (Eilers and Marx, 1996) is used. To cluster variables, an indicator variable is introduced for cluster assignment. To cluster subjects, a Dirichlet process mixture model is applied. The proposed joint clustering method has the ability to produce homogeneous clusters composed of a certain number of subjects sharing common features in the relationship between some variables and covariates.

The remainder of the article is organized as follows. Section 2 introduces the model of joint clustering under the Bayesian framework and settings for the priors. The full

conditional posteriors, detailed procedure and approach of joint clustering are also described in this section. We demonstrate and evaluate the performance of the proposed method in Section 3 through simulations. The proposed approach is then applied to analyze allergic sensitization data. We cluster subjects and allergens based on associations of wheal sizes in reaction to allergens with age. This is discussed in Section 4. We summarize our methods and findings and discuss limitations in Section 5.

2 The Method

We consider the following joint (two-dimensional) clusters which are illustrated in Figure 1. To ease the presentation, we dissect the unified clustering process into two parts. In part 1, variables are clustered; and in part 2, subjects within each variable cluster are further clustered to form refined clusters, where the correlations between the variables are taken into account. We then combine parts 1 and 2 and lay out the joint clustering scheme.

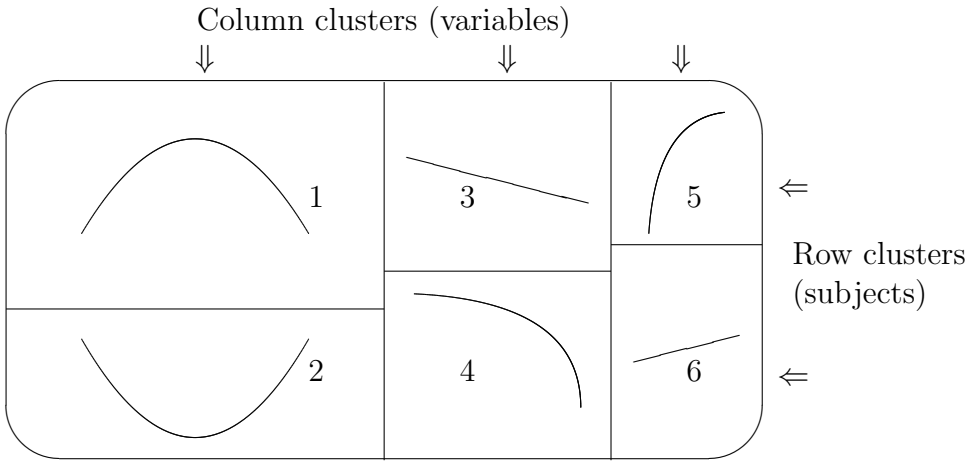


Figure 1: Illustration of joint clusters. In total 6 clusters and the numbers are joint cluster indices.

2.1 Clustering the Variables

We cluster variables based on agreement in relationships between variables and covariates of interest. Assume in total n subjects and K variables are under consideration for clustering.

For subject $i, i = 1, \cdots, n$, let $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{iK})$ denote the measures of K variables. Let M denote the number of clusters formed by the variables ($M \leq K$), and D be an $M \times K$ 0-1 matrix such that the k^{th} column contains $M-1$ zeros and one element with the number 1 indicating which cluster the k^{th} variable ($k = 1, 2, \ldots, K$) belongs to. The value of M will be determined via grid search, which is further discussed in Section 2.4. For a given M, elements in the m^{th} row of D, $m = 1, \cdots, M$, inform which variables are in cluster m. We formulate the variable clustering procedure into the following:

$$\boldsymbol{y}_{i,m}|D_{m.} = \boldsymbol{Q}(\boldsymbol{x}_i, \boldsymbol{\beta}_{i,m}) + \boldsymbol{\varepsilon}_{i,m}^T, \tag{1}$$

where $\boldsymbol{y}_{i,m} = (y_{i,(1)}, \ldots, y_{i,(k_m)})'$ is a vector of variables in variable cluster m, $\boldsymbol{x}_i$ is a vector of covariates potentially associated with $\boldsymbol{y}_{i,m}$, $\boldsymbol{\beta}_{i,m}$ describes the association of $\boldsymbol{y}_{i,m}$ with $\boldsymbol{x}_i$ in cluster variable m for subject i, and $\boldsymbol{\varepsilon}_{i,m}^T$ follows a multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix Σ_m. The covariance Σ_m informs the strength of correlations between the variables in cluster m. Based on the property of homogeneity in a cluster (assuming variables are properly transformed when necessary), the variances of the variables in one cluster are assumed to be the same, i.e., the diagonal elements are the same. For formulation simplicity, we also assume that variables in different variable clusters are independent. In the simulations, we demonstrate the robustness of this assumption. Function $\boldsymbol{Q}(\cdot)$ is a vector function which describes the relationship between variables and covariates of interest $\boldsymbol{x}_i$. We use semi-parametric models to model this relationship. Specifically, penalized splines (P-Splines) are applied (Eilers and Marx, 1996) due to its use of low rank bases. In order to achieve the smoothness, second order P-Splines is used. For one covariate x_i,

$$Q(x, \boldsymbol{\beta}_{i,m}) = a_{im,1}x_i + a_{im,2}x_i^2 + \sum_{l=1}^{g} b_{im,l}(x_i - z_l)_+^2,$$

where g is the number of knots, $\beta_m = (a_{im,1}, a_{im,2}, b_{im,1}, \ldots, b_{im,g})'$ is of length $(g+2)$ representing coefficients for the P-Splines for the m^{th} cluster, z_l's are the spline knots, and

$$(x_i - z_l)_+ = \begin{cases} 0, & \text{if } x_i \leqslant z_l, \\ x_i - z_l, & \text{if } x_i > z_l. \end{cases}$$

We write $X_i = (x_i, x_i^2, (x_i - z_1)^2, \ldots, (x_i - z_g)^2)'$, which gives $Q(x_i, \boldsymbol{\beta}_{i,m}) = X_i'\boldsymbol{\beta}_{i,m}$. Since the dependent variables are multivariate, we write $\boldsymbol{X}_i = X_i \otimes \mathbf{1}_{k_m}$. Where $\otimes$ is the Kronecker

product; $\mathbf{1}_{k_m}$ is a row vector of dimension k_m composed of 1's. Knots can be chosen to be evenly spaced between the range of x_i (Ruppert et al., 2003). As for the number of knots (g), it has been demonstrated that a large number of knots is not necessary (Baladandayuthapani et al., 2005; Ruppert et al., 2003). In our study, we use 10 knots in simulations as well as in the real data application.

A Bayesian approach is applied to infer the variable clusters. The following lists the prior distributions of $D_{\cdot k}, \Sigma_m$, and hyper-parameters, where $D_{\cdot k}$ denotes the k^{th} column of D representing which cluster the k^{th} variable belongs to.

$$\begin{aligned} D_{\cdot k}|\boldsymbol{\pi} &\sim \text{Multinomial}(1, \boldsymbol{\pi}), \\ \boldsymbol{\pi}|\zeta &\sim \text{Dirichlet}(\zeta \mathbf{1}_M), \\ \zeta &\sim p(\zeta) = \frac{1}{2} \text{ if } 0 < \zeta \leq 1, \text{ and } \frac{1}{2}\zeta^{-2} \text{ if } \zeta > 1, \\ \Sigma_m|(S,\nu) &\sim \text{InvWishart}(S,\nu), \end{aligned} \tag{2}$$

where $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_M)$ gives the probabilities that the k^{th} variable is in each of the M clusters. We choose a Dirichlet distribution for the prior of $\boldsymbol{\pi}$. The distribution of hyper-prior ζ in $\boldsymbol{\pi}$ is chosen following the suggestion of Good (1965). The parameter Σ_m is the variance-covariance matrix of $\boldsymbol{\epsilon}_{i,m}$ and an inverse Wishart distribution is selected as its prior distribution. The choice of hyper-prior parameters S and ν controls the variation in Σ_m for cluster m. To avoid non-necessary difficulty in differentiating different variable clusters, we take $S = \frac{1}{2}\boldsymbol{I}$ with I standing for identity matrix and select ν such that the prior mean of Σ_m is moderate on the diagonal. The prior distribution of $\boldsymbol{\beta}_{i,m}$ is discussed in the next section due to its involvement in subject clustering.

2.2 Clustering the Subjects

The subjects within each variable cluster are further grouped such that each group reflects a different relationship between variables and covariates of interest. We can follow what is done in the clustering of variables, i.e., introducing an indicator matrix and determining the number of subject clusters within each variable cluster via grid search. Doing so will significantly increase computing burden. To ease the computing complexity, we propose to use the Dirichlet process to fulfill the goal of clustering subjects. The Dirichlet process has

the ability to detect clusters without the need of defining a particular parameter for the number of clusters as done under the multinomial setting. Specifically, we assume that the prior distribution of $\boldsymbol{\beta}_{i,m}$ is generated from a Dirichlet process,

$$\begin{aligned}
\boldsymbol{\beta}_{i,m}|G &\sim G, \\
G &\sim \mathrm{DP}(G_0, \lambda), \\
G_0|\sigma_0^2 &\sim \boldsymbol{N}(\mathbf{0}, \sigma_0^2 I), \\
\sigma_0^2|(a,c) &\sim \mathrm{InvGamma}(a,c),
\end{aligned}$$

where $\boldsymbol{\beta}_{i,m}|G$ are independent given G, and $\mathrm{DP}(G_0, \lambda)$ represents the Dirichlet process with a measure having concentration λ and proportional to the base distribution $G_0 \sim \boldsymbol{N}(\mathbf{0}, \Sigma_0)$ with $\Sigma_0 = \sigma_0^2 I$. The prior of $\boldsymbol{\beta}_{i,m}$ conditional on $\boldsymbol{\beta}_{-i,m}$, the coefficients with $\boldsymbol{\beta}_{i,m}$ excluded, is a mixture distribution

$$\boldsymbol{\beta}_{i,m}\Big|\boldsymbol{\beta}_{-i,m} \sim \frac{1}{n-1+\lambda}\sum_{j\neq i}\delta_{\boldsymbol{\beta}_{i,m}}(\boldsymbol{\beta}_{j,m}) + \frac{\lambda}{n-1+\lambda}G_0,$$

where $j = 1, \cdots, n, j \neq i$, $\delta_{\boldsymbol{\beta}_{i,m}}(\boldsymbol{\beta}_{j,m})$ is a point mass concentrated at a single point where $\beta_{i,m} = \beta_{j,m}$ (i.e., $\delta_{\boldsymbol{\beta}_{i,m}}(\boldsymbol{\beta}_{j,m}) = 1$ if $\beta_{i,m} = \beta_{j,m}, j \neq i$), and λ is the concentration parameter. As for now, we assume λ is known and discuss its selection in Section 2.4.

For the hyper-prior parameters a and c, they are assumed to be known and selected to achieve vague priors. In particular, we set $a = c = 0.5$, assuming unit-information prior based on the suggestion by Kass and Wasserman (1995).

2.3 Joint Clustering

When clustering variables and subjects jointly, we combine Sections 2.1 and 2.2 to meet this need,

$$
\begin{aligned}
\boldsymbol{y}_{i,m}|(\boldsymbol{\beta}_{i,m}, D, \Sigma_m) &\sim \boldsymbol{N}\big(\boldsymbol{X}_i'\boldsymbol{\beta}_{i,m}, \Sigma_m\big), \\
\boldsymbol{\beta}_{i,m}|G &\sim G, \quad G \sim \mathrm{DP}(G_0, \lambda), \\
G_0|\sigma_0^2 &\sim \boldsymbol{N}(\mathbf{0}, \sigma_0^2 I), \\
\sigma_0^2|(a,c) &\sim \mathrm{InvGamma}(a,c), \\
D_{.k}|\boldsymbol{\pi} &\sim \mathrm{Multinomial}(1, \boldsymbol{\pi}), \\
\boldsymbol{\pi}|\zeta &\sim \mathrm{Dirichlet}(\zeta \mathbf{1}_M), \\
\zeta &\sim p(\zeta) = \frac{1}{2} \text{ if } 0 < \zeta \leq 1, \text{ and } \frac{1}{2}\zeta^{-2} \text{ if } \zeta > 1, \\
\Sigma_m|(S,\nu) &\sim \mathrm{InvWishart}(S,\nu),
\end{aligned}
$$

with settings for hyper-parameters S, ν, a, and c defined in Sections 2.1 and 2.2.

2.4 Posteriors Computing

Markov chain Monte Carlo (MCMC) simulations, specifically the Gibbs sampler with Metropolis-Hastings steps, are implemented to generate observations from full conditional posterior distributions, which are then used to infer the parameters of interest. In the following, we list the conditional posterior distributions.

For parameters related to clustering the dependent variables, we have the following conditional posterior distributions,

$$
\begin{aligned}
\boldsymbol{\pi}\Big|(D,\zeta) &\sim \mathrm{Dirichlet}(D_{.k} + \boldsymbol{\zeta}), \\
&\quad \text{where } \boldsymbol{\zeta} \text{ is a vector with all components equal to } \zeta. \\
D_{.k}\Big|\boldsymbol{y}_{i,m}, \boldsymbol{\beta}_{i,m}, \Sigma_m, \boldsymbol{\pi} &\sim \mathrm{Multinomial}(1, \boldsymbol{\pi_0}), \\
\boldsymbol{\pi}_{0m} &\propto p(\boldsymbol{y}_{i,m}|(\boldsymbol{\beta}_{i,m}, D_{.k}, \Sigma_m))p(D_{.k}|\boldsymbol{\pi}), m = 1, \cdots, M, \\
&\quad i = 1, \cdots, n,
\end{aligned} \tag{3}
$$

where sampling of $D_{.k}$ and $\boldsymbol{\pi}_{0m}$ depends on the coefficients in variable cluster m for subject i, $\boldsymbol{\beta}_{i,m}$. The conditional posterior distribution of ζ is not in a standard form and to sample ζ, we apply the Metropolis-Hastings algorithm and take the log-normal distribution as the proposal distribution.

The conditional posterior distributions in the procedure of further clustering subjects within each variable cluster include the conditional posterior distribution of $\boldsymbol{\beta}_{i,m}$. Assuming the data are exchangeable (Neal, 2000), we have:

$$\boldsymbol{\beta}_{i,m}\Big|(\boldsymbol{\beta}_{-i,m}, \boldsymbol{y}_{i,m}) \sim \sum_{j\neq i} q_{i,j}\delta(\boldsymbol{\beta}_{j,m}) + r_i H_i,$$

where

$$\begin{aligned}
q_{i,j} &= b\frac{1}{n-1+\lambda}(2\pi)^{-\frac{k_m}{2}}|\Sigma_m|^{-\frac{1}{2}} \\
&\quad \exp\Big[-\frac{1}{2}(\boldsymbol{y}_{i,m}-\boldsymbol{X}_i\boldsymbol{\beta}_{j,m})'\Sigma_m^{-1}(\boldsymbol{y}_{i,m}-\boldsymbol{X}_i\boldsymbol{\beta}_{j,m})\Big], \\
r_i &= b\frac{\lambda}{n-1+\lambda}(2\pi)^{-\frac{k_m}{2}}|\Sigma_m|^{-\frac{1}{2}}|\Sigma_0|^{-\frac{1}{2}}|\Sigma_{\beta_{i,m}}|^{\frac{1}{2}} \\
&\quad \exp\Big[-\frac{1}{2}\boldsymbol{y}_{i,m}'\Sigma_m^{-1}\boldsymbol{y}_{i,m}+\frac{1}{2}\boldsymbol{y}_{i,m}'\Sigma_m^{-1}\boldsymbol{X}_i\Sigma_{\beta_{i,m}}\boldsymbol{X}_i'\Sigma_m^{-1}\boldsymbol{y}_{i,m}\Big], \\
\Sigma_{\beta_{i,m}} &= (\boldsymbol{X}_i'\Sigma_m^{-1}\boldsymbol{X}_i+\Sigma_0^{-1})^{-1}, \\
H_i &\sim \boldsymbol{N}\Big(\Sigma_{\beta_{i,m}}(\boldsymbol{X}_i'\Sigma_m^{-1}\boldsymbol{y}_{i,m}), \Sigma_{\beta_{i,m}}\Big),
\end{aligned}$$

where b is a normalizing constant. To sample $\boldsymbol{\beta}_{i,m}$, we implement Algorithm 2 summarized in Neal (2000), initially proposed by Bush and MacEachern (1996). Related conditional posteriors implemented in this Algorithm are discussed in Appendix A. Other conditional posterior distributions involved in the clustering procedure are for Σ_m and Σ_0:

$$\begin{aligned}
\Sigma_m\Big|(Y,\boldsymbol{\beta}_{i,m}) &\sim \text{InvWishart}\Big(\sum_{i=1}^{n}(\boldsymbol{y}_{i,m}-\boldsymbol{X}_i'\boldsymbol{\beta}_{i,m})(\boldsymbol{y}_{i,m}-\boldsymbol{X}_i'\boldsymbol{\beta}_{i,m})'+S, n+\nu\Big), \\
\Sigma_0[j]\Big|(Y,\boldsymbol{\beta}_{i,m}) &\sim \text{InvGamma}\Big(a_2+\frac{n(2+g)}{2}, c_2+\sum_{i=1}^{n}\boldsymbol{\beta}_{i,m}[j]\Big),
\end{aligned}$$

where $\Sigma_0[j]$ denotes the j^{th} diagonal element in Σ_0, and $\boldsymbol{\beta}_{i,m}[j]$ is the j^{th} component of $\boldsymbol{\beta}_{i,m}$. A Gibbs sampler will then be used to sample from the aforementioned conditional posterior distributions. Details of the sampling procedure are outlined in Appendix B.

Now we turn to the selection of the concentration parameter λ in the Dirichlet process. This parameter controls the distribution of $\boldsymbol{\beta}_{i,m}$ over the number and sizes of clusters. If λ is large, the prior assigns distributions that are close to the base distribution. If we have prior knowledge on the number of clusters, we can pre-specify λ based on such knowledge.

For instance, we can set λ small assuming that the number of subject clusters is substantially smaller than the sample size. This assumption is realistic and does not lose generality in many real applications. However, it is noteworthy that simply pre-specifying the concentration parameter can potentially increase misclassifications. Antoniak (1974) also noted that a caution is needed when choosing too small values for λ. A prior distribution for λ, e.g., a gamma distribution, is suggested in some earlier studies, but sensitivity of the posterior inferences of λ to its prior choice has been discussed in various applications (Dorazio et al., 2008; McAuliffe et al., 2006). Doss (2008) indicates that parameter λ is typically the most difficult to estimate or defend as a fixed value. Different approaches have been proposed to infer or estimate λ. For instance, both McAuliffe et al. (2006) and Dorazio et al. (2008) adopted a numerical approach based on the work of Liu (1996) to estimate λ. However, our simulations (results not shown) indicate that this approach has the risk of under estimating λ. Kyung et al. (2010) also noted that the choice of λ based on Dorazio et al. (2008) may be far from the truth due to the possibility of flat likelihood of λ. Other approaches determining the concentration parameter have also been proposed, e.g., Doss (2008, 2012) and Kyung et al. (2010). Knowing the pros and cons of all these developed methods, to avoid a potential bias and to reduce complexity in the process of inferring λ, in this article, we decide to choose λ by maximizing the joint posterior likelihood.

To determine the number of variable clusters, M, we optimize the deviance information criterion (DIC) (Spiegelhalter et al., 2002). That is, we run the clustering process for a set of different M values and choose the results showing the smallest DIC. The final number of joint clusters is decided by identifying an iteration with "least-squares distance", a procedure adapted from Dahl (2006). Details are given in Appendix C.

3 Simulation Study

This section, via simulations, demonstrates the proposed method, assesses its sensitivity with respect to large variations in data and dependence between variable clusters, and compares the method with existing approaches. The program for the proposed method is written in R and available to readers of interest.

3.1 Settings

We generate 100 Monte Carlo (MC) replicates, with each of sample size 400 and having 10 dependent variables, and one covariate x_i generated from a uniform distribution between 1 and 6. The 10 variables are grouped into 3 clusters and within each variable cluster, the subjects are further clustered. Following is the setting of the clusters and the associations defined for each cluster:

- Cluster 1, $E(y_{ij}) = 6 + 5\sin(0.2\pi(x_i - 1))$ for $i = 1, \ldots, 250$ and $j = 1, \ldots, 5$
- Cluster 2, $E(y_{ij}) = -5 - 5\cos(0.2\pi(x_i - 3.5))$ for $i = 251, \ldots, 400$ and $j = 1, \ldots, 5$
- Cluster 3, $E(y_{ij}) = 10 - 0.8x_i$ for $i = 1, \ldots, 200$ and $j = 6, 7, 8$
- Cluster 4, $E(y_{ij}) = -5 - 3\exp\big(0.4(x_i - 1)\big)$ for $i = 201, \ldots, 400$ and $j = 6, 7, 8$
- Cluster 5, $E(y_{ij}) = 15 + 4\log\big(0.4(x_i - 0.8)\big)$ for $i = 1, \ldots, 180$ and $j = 9, 10$
- Cluster 6, $E(y_{ij}) = -2 + 0.1x_i$ for $i = 181, \ldots, 400$ and $j = 9, 10$

In total, we have 6 joint clusters with each cluster having a specific association between $\boldsymbol{y}$ and x for a subset of subjects. The distribution of random errors is assumed to be multivariate normal with mean $\mathbf{0}$ and the following variance-covariance matrices for the three variable clusters

$$\Sigma_1 = \begin{bmatrix} 1 & -0.25 & 0 & 0 & 0 \\ -0.25 & 1 & -0.5 & 0 & 0 \\ 0 & -0.5 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0.3 \\ 0 & 0 & 0 & 0.3 & 1 \end{bmatrix}, \Sigma_2 = \begin{bmatrix} 1 & 0.1 & -0.4 \\ 0.1 & 1 & -0.1 \\ -0.4 & -0.1 & 1 \end{bmatrix},$$

$$\Sigma_3 = \begin{bmatrix} 0.6 & -0.25 \\ -0.25 & 0.6 \end{bmatrix}.$$

The patterns of the 6 clusters are displayed in Figure 1. We denote this simulation design as Scenario 1. To assess the quality of clustering, we record the number of joint clusters, accurate rate calculated based on pairwise agreement of clustering (such that pairs

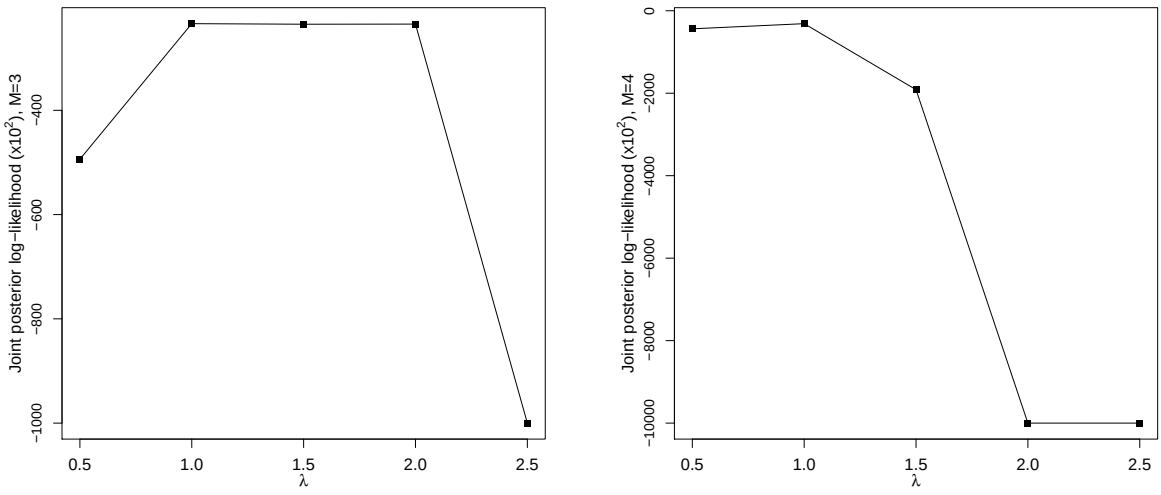
$[i, j]$ ([subject, variable]) and $[i', j']$ are in one cluster), sensitivity (Se=TP/(TP+FN)), and specificity (Sp=TN/(TN+FP)) with respect to a specific cluster, where "TP" denotes true positives (correct cluster identification), "FN" false negatives, "TN" true negatives, and "FP" false positives.

3.2 Results

We discuss findings from the fully Bayesian sampling scheme discussed in Section 2.4 such that $D_{.k}^{(t)}$ is sampled from the distribution in (3). In total, $g = 10$ evenly spaced knots are taken in the P-Splines. For the selection of concentration parameter λ of a given M, we use grid search by maximizing the posterior likelihood. Smaller values of λ are preferred as they are in expectation corresponding to parsimonious clusters (smaller numbers of clusters). This grid search is applied to a randomly selected data set from the 100 MC replicates, and 2,000 MCMC iterations are run with 1,000 MCMC iterations from one chain after 1,000 burn-in iterations used to estimate the parameters and calculate the joint posterior likelihood for that randomly selected data set. A value of λ that maximizes the posterior likelihood is then taken for all the remaining MC replicates. After λ determined for each M, for each MC replicate, supported by potentially fast convergence of MCMC chains, we run two MCMC chains with 1,000 iterations each chain for each MC replicate with 300 iterations as burn-in, the next 300 for the determination of the average clustering matrix, and the last 400 iterations for inferences.

Figure 2 demonstrates patterns of the posterior likelihoods with respect to different values of λ at given values of M. For instance, taking concentration parameter $\lambda = 1.0$ at $M = 3$ maximizes the likelihood. Our additional simulations (results not shown) demonstrate that taking λ in the neighbors of maximization point for a given M gives similar posterior inferences on clusters and parameters.

After λ is determined for each M, the final value of M for each MC replicate is determined by minimizing DICs. The pattern of DICs for each M across the 100 MC replicates are given in Figure 3. Out of 100 MC replicates, 73 replicates are optimized at $M = 3$ variable clusters. For the numbers of joint clusters among these 73 MC replicates, the median is 6 and a 95% empirical interval is [6, 8]. The sensitivity and specificity of the joint clus-



a) b)

Figure 2: Joint posterior log-likelihood v.s. λ (simulation scenario 1). a) $M = 3$, b) $M = 4$.

tering process with respect to the true clustering evaluated based on pairwise agreements are listed in Table 1 (in the columns named "Scenario 1"). Overall, high sensitivities and specificities are observed, indicating the effectiveness of the proposed method.

To illustrate the fitting performance of the proposed joint clustering method, we choose results from one data set. The fitted curves for all clusters are shown in Figure 4 along with the true curves. The fitted curves are all close to the true curves except slight deviation at two ends of the curves. This is further reflected by the 95% empirical posterior interval bands.

3.3 Further Assessment of the Method

In the above analysis, we demonstrated the robustness of the method via sensitivity and specificity with respect to different cluster patterns. In this section, via simulations, we evaluate the impact of large variation in data on the quality of clustering, assess the sensitivity on the independence assumption between variable clusters, and compare the proposed method with existing competing methods.

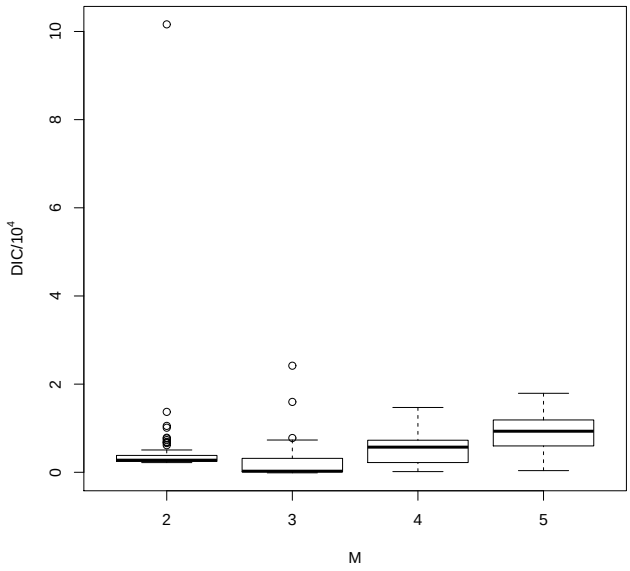


Figure 3: Box plots of DICs v.s. different numbers of variable clusters (M) (simulation scenario 1).

Impact of large variation. To assess this impact, we increase the values of variance components in the covariance matrices for the three variable clusters from $\{1, 1, 0.6\}$ to $\{5, 6, 6\}$, respectively. Other settings are kept the same as before. We denote this simulation scenario as Scenario 2. Table 1 summarizes the sensitivities and specificities across 100 MC replicates (the column indicated by "Scenario 2"). The sensitivities and specificities are not severely impacted by the substantially increased variations in the data. The lowest average sensitivity occurs for joint cluster 1 but is still higher than 0.90, indicating high sensitivity in most MC replicates. However, the standard deviations are larger than those under Scenario 1, implying increased uncertainty potentially due to larger variations in the data.

Sensitivity on the independence assumption between variable clusters. In the previous simulations, all the MC replicates are simulated such that variables between different variable clusters are independent, which follows the assumption of the proposed method. To demonstrate whether the method is sensitive to this assumption, we generated 100 MC

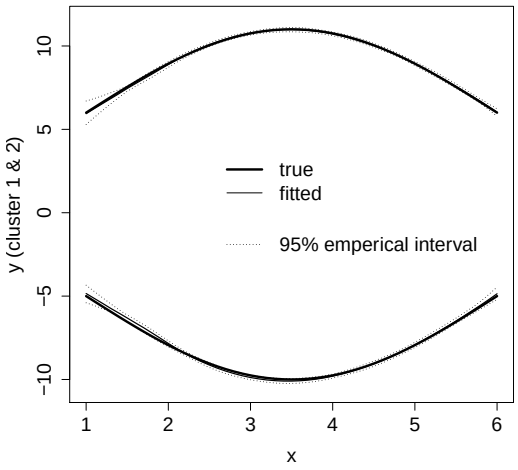
Table 1: The average sensitivity and specificity for the pre-specified 6 joint clusters with $D_{.k}^{(t)}$ sampled from (3). SD: Standard Deviations.

	Scenario 1		Scenario 2		Scenario 3	
	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity
Cluster	Mean (SD)					
1	0.980 (0.141)	0.997 (0.027)	0.959 (0.197)	0.994 (0.039)	0.990 (0.100)	1.000 (0.000)
2	0.994 (0.047)	0.992 (0.049)	0.992 (0.047)	0.981 (0.076)	0.999 (0.011)	0.996 (0.037)
3	0.998 (0.023)	1.000 (0.000)	0.980 (0.141)	0.999 (0.000)	1.000 (0.000)	1.000 (0.000)
4	0.990 (0.100)	1.000 (0.000)	0.980 (0.141)	0.999 (0.000)	1.000 (0.000)	1.000 (0.000)
5	0.999 (0.001)	0.999 (0.000)	0.990 (0.100)	0.999 (0.000)	0.990 (0.100)	0.999 (0.000)
6	0.994 (0.053)	1.000 (0.000)	0.996 (0.011)	0.999 (0.010)	0.994 (0.044)	0.999 (0.010)

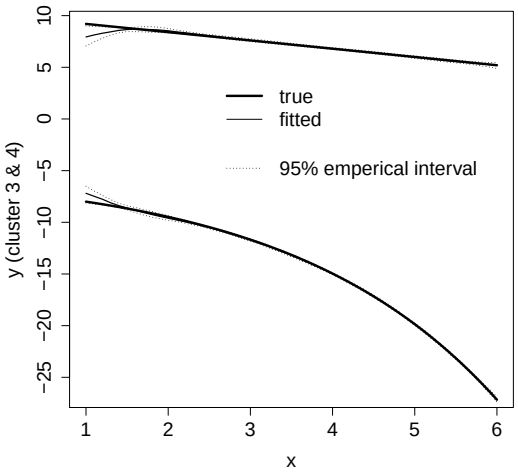
replicates such that all the variables are correlated and the correlation between every two consecutive variables is defined as $0.6^{|i-j|}, i, j = 1, \cdots, 10$. The variance of each variable is set at 1. The sensitivity and specificity statistics across the 100 MC replicates are summarized in Table 1 (the column indicated by "Scenario 3"). The results all show high average sensitivities and specificities and are comparable to those in Table 1 under Scenario 1. The slightly increased sensitivity and specificity for some clusters is likely due to the increased homogeneity in the variables, which benefits the quality of clustering.

Comparing with existing methods. Currently, there is no method available that can jointly cluster variables and subjects with effect from external variables incorporated. Existing biclustering methods do not take the contribution of external variables into account. To fulfill the goal of comparison and demonstrate the effectiveness of the proposed method even in the situation of no external covarites, we generate 100 MC replicates with each of sample size 400 and having 10 dependent variables. The 10 variables are grouped into 3 clusters and within each variable cluster, the subjects are further clustered. Following is the setting of the clusters and the associations defined for each cluster:

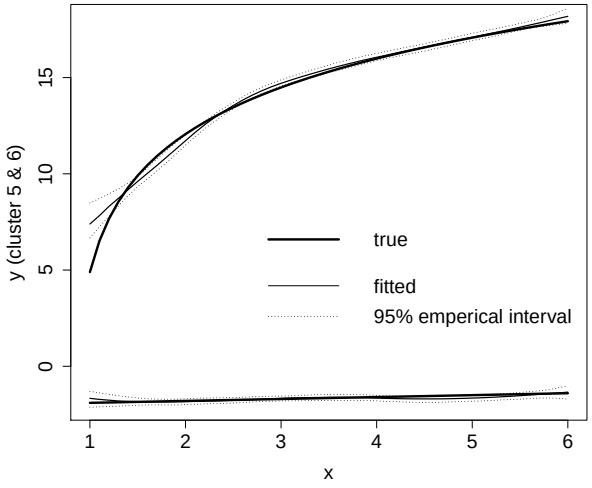
- Cluster 1, $E(y_{ij}) = 6$ for $i = 1, \ldots, 250$ and $j = 1, \ldots, 5$
- Cluster 2, $E(y_{ij}) = -5$ for $i = 251, \ldots, 400$ and $j = 1, \ldots, 5$
- Cluster 3, $E(y_{ij}) = 10$ for $i = 1, \ldots, 200$ and $j = 6, 7, 8$



(a) First variable cluster



(b) Second variable cluster



(c) Third variable cluster

Figure 4: The fitted curves vs true curves for all clusters.

- Cluster 4, $E(y_{ij}) = -8$ for $i = 201, \ldots, 400$ and $j = 6, 7, 8$
- Cluster 5, $E(y_{ij}) = 15$ for $i = 1, \ldots, 180$ and $j = 9, 10$
- Cluster 6, $E(y_{ij}) = -2$ for $i = 181, \ldots, 400$ and $j = 9, 10$

In total, we have 6 joint clusters. The distribution of random errors is assumed to be multivariate normal with mean 0 and variance-covariance matrix composed of three diagonal blocks for each of the three variable clusters, $\Sigma_1 = I_5, \Sigma_2 = I_3, \Sigma_3 = 0.6I_2$, where the subscript of I stands for the dimension of the identity matrix.

For the competing methods, we consider two commonly used bicluster approaches. One approach is proposed by Cheng and Church (2000) (BCCC), which identifies biclusters formed by genes and conditions by minimizing mean squared residues calculated based on mean expression levels, and the other is developed by Prelić et al. (2006) (BCBimax) utilizing a fast divide and conquer approach via a binary inclusion-maximal biclustering algorithm. These and other existing biclustering methods allow data points to be in more than one biclusters, while our proposed joint clustering method is mutually exclusive for the cluster assignment. We apply the proposed method and the two competing methods (BCCC and BCBimax) to the 100 MC replicates and summarize the findings by use of sensitivities and specificities. The results are shown in Table 2. Since this set of simulated data do not consider associations of the variables with external variables, as expected, sensitivities and specificities from the proposed methods are all high with small variations across different MC replicates; in Table 2, only the means of sensitivities and specificities from the proposed method are listed. For BCCC and BCBimax, besides means, we included more detailed information on the distribution of sensitivities and specificities, i.e., mean, median, and 95% empirical intervals. The BCCC method, in general, gives reasonable sensitivity and specificity, but is inferior to the proposed method. The BCBimax method, on the other hand, provides sensitivity comparable to the proposed method, but have a high risk to sacrifice specificity substantially. This is potentially due to the ambiguity in the clustering process of these biclustering methods, which may cause difficulty in the interpretation of findings.

4 Real Data Analysis

We apply the proposed method to measures of wheal sizes in reaction to 11 allergens on 972 children aged 4, 10, or 18 years, to identify joint clusters of allergens and subjects based on associations of wheal sizes with age. The data are extracted from a longitudinal

Table 2: Comparison of the average sensitivity for the pre-specified 6 joint clusters of the proposed method, BCCC and BCBimax. EI: empirical interval

	Sensitivity						
	Proposed	BCCC			BCBimax		
Cluster	Mean	Mean	Median	95% EI	Mean	Median	95% EI
1	0.999	0.713	0.712	(0.702,0.72)	0.990	1	(1,1)
2	0.999	0.997	1	(0.957,1)	0.990	1	(1,1)
3	0.998	0.891	0.890	(0.877,0.900)	0.990	1	(1,1)
4	0.999	0.748	0.750	(0.717,0.755)	0.990	1	(1,1)
5	0.999	0.990	0.989	(0.975,1)	0.950	1	(0.5,1)
6	0.997	0.680	0.682	(0.652,0.686)	0.765	0.8	(0.447,0.857)
	Specificity						
1	0.999	0.674	0.675	(0.655,0.681)	0.288	0.291	(0.291,0.291)
2	0.999	0.770	0.769	(0.766,0.779)	0.244	0.246	(0.246,0.246)
3	0.998	0.632	0.631	(0.617,0.639)	0.234	0.235	(0.235,0.235)
4	0.999	0.692	0.691	(0.689,0.705)	0.234	0.235	(0.235,0.235)
5	0.999	0.607	0.607	(0.594,0.614)	0.117	0.118	(0.082,0.153)
6	0.998	0.664	0.663	(0.661,0.678)	0.096	0.099	(0.067,0.125)

study cohort aiming to investigate the history of asthma. Details of the cohort is discussed elsewhere (Hide et al., 1996). Without loss of generality, we standardized the data before analyzing to avoid potential bias in clustering caused by heterogeneous scale.

We follow the same procedure as in simulations to choose λ and M. The only difference is that for the selected M, we ran two longer MCMC chains with each chain of 15,000 iterations. The results presented in the article are based on one chain, of which 6,000 iterations are used for burn-in, 4,500 iterations to calculate the average clustering matrix, and posterior inferences are drawn from the remaining 4,500 iterations.

Figure 5 indicates that the smallest distance occurs at 6 joint clusters, including 3 allergen clusters ($M = 3$) and subjects are further grouped within each allergen cluster. Specifically, 2 subject clusters are in allergen cluster {*Alternaria*, *Cladosporium*, Cod, Peanut, Egg, Milk, and Soya}, 2 subject clusters for allergen Grass, and 2 subject clusters in allergen cluster {Cats, Dogs, House dust mite} (Figure 6). The clustering of allergens is as expected; food allergens and fungi with food as their major sources are clustered together (*Alternaria*, *Cladosporium*, Cod, Peanut, Egg, Milk, and Soya), indoor allergens are clustered together (Cats, Dogs, House dust mite), and allergen grass represents outdoor allergens. In these 6 joint clusters, wheal sizes in clusters 2, 4, and 6 are all small and do not show a clear pattern over time. In the remaining joint clusters, the wheal sizes are generally larger, but temporal patterns vary between clusters (Figure 6). For allergens *Alternaria*, *Cladosporium*, Cod, Peanut, Egg, Milk, and Soya, wheal sizes first increases with age and then becomes roughly stable over time (cluster 1), for allergen Grass, a convex pattern is observed with a pattern of slow increase at a later age in adolescence; for allergens Cats, Dogs, and House dust mite, the wheal sizes at earlier ages decreases slightly and then shows a faster increase afterwards.

Sizes of the wheals reflect a potential severity of allergic sensitization (atopy) and atopy is linked to asthma. We further examined the percentages of subjects in each joint cluster who ever had asthma and linked the percentages to the identified joint cluster patterns. The prevalence of asthma ever in each joint cluster is recorded in Table 3. Although this is not a longitudinal analysis, the findings indicate that clusters showing increased wheal sizes with respect to food and food-related allergens at an earlier age (joint cluster 1) or a larger

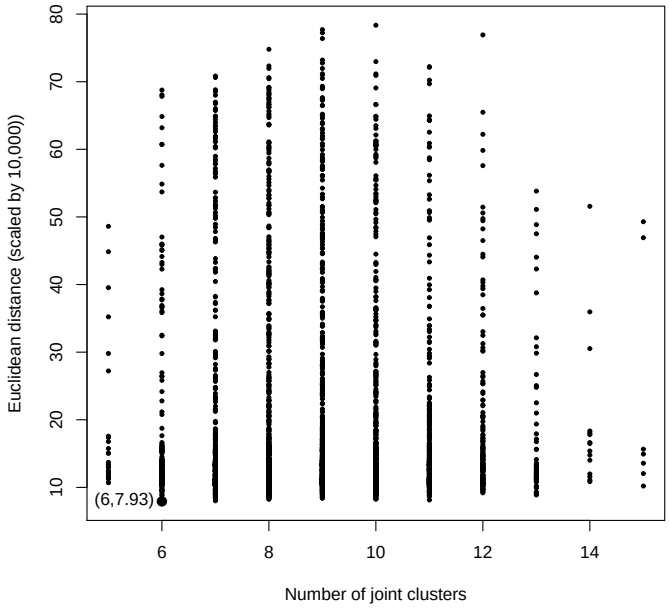
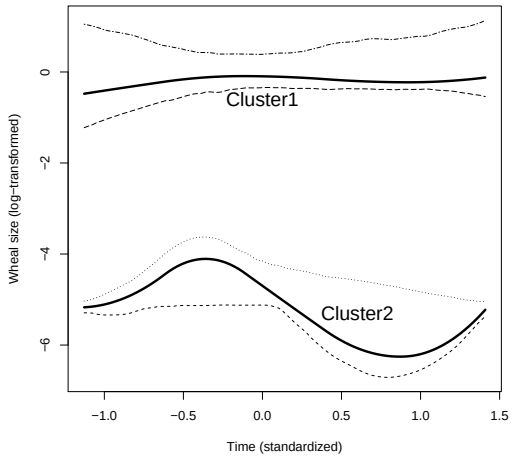


Figure 5: The Euclidean distances between cluster assignment matrix and the average clustering matrix at each of the 4,500 iterations. In the figure, (6, 7.93) refers to the minimum distance being 7.93 (scaled by 10,000) with 6 clusters. This is the smallest distance across all possible clusters.

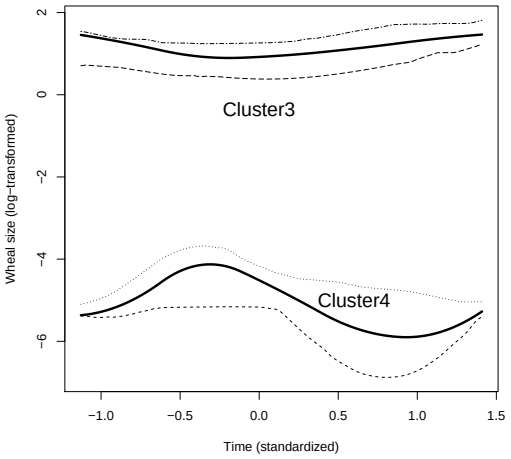
increase at a later age against indoor allergens during adolescence (joint cluster 5) are likely to have more subjects with asthma. However, the percentage is decreased by more than 10% if wheal sizes against outdoor allergens decrease at an earlier age before adolescence and are roughly stable at a later age (joint cluster 3). We postulate that promoting atopy remission during the transition period of adolescence has a potential to reduce acquisition of asthma and promote remission of asthma.

5 Conclusion and Discussion

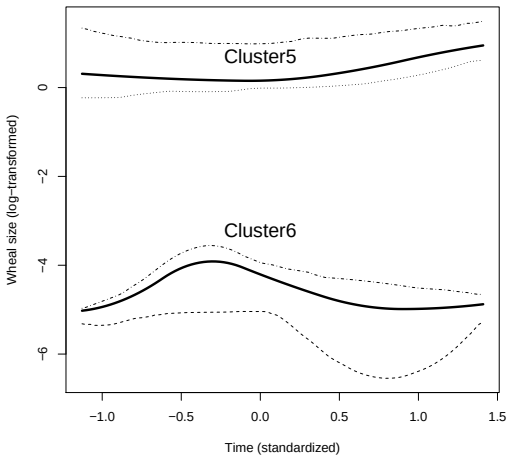
We proposed a joint clustering method in a Bayesian framework to probabilistically identify clusters composed of variables (such as measures of sensitization to allergens) and subjects.



(a) First allergen cluster (*Alternaria*, *Cladosporium*, Cod, Peanut, Egg, Milk, Soya)



(b) Second allergen cluster (Grass)



(c) Third allergen cluster (Cats, Dogs, and House dust mite)

Figure 6: The temporal patterns of wheal sizes of the 6 joint clusters. Solid lines are the fitted curves. The dotted lines provide empirical 95% confidence bands. The time in the X-axis represents standardized ages.

Table 3: Sizes of each identified cluster and proportions of asthma at age 18 years in each cluster. Allergen cluster A: *Alternaria*, *Cladosporium*, Cod, Peanut, Egg, Milk, Soya. Allergen cluster B: Grass. Allergen cluster C: Cats, Dogs, House dust mite.

Allergen clusters		
A	B	C
Joint cluster index; size; % of asthma		
1; 13; 61.5	3; 161; 49.7	5; 66; 59.1
2; 959; 27.0	4; 811; 23.1	6; 903; 25.0

In each cluster, there exists a unique association between a subset of variables and covariates of interest, and such an association is described by a semi-parametric model. Penalized splines are implemented to estimate the associations due to its low rank bases and ability to capture linear and non-linear effects. The joint clustering strategy clusters variables and subjects simultaneously, and takes into account potential dependence among variables.

Overall, the proposed methods can effectively identify the joint clusters with high sensitivity and specificity. We also demonstrated that the method is robust against large variations in the data and has the ability to handle dependence between and among variables. A comparison with two commonly used biclustering methods supports the effectiveness of the proposed method. We applied the method to sensitization measures and identified subsets of allergens as well as patterns of wheal size over time that may play an important role in the occurrence of asthma, which provides a significant insight into the understanding of allergic diseases and their relation to atopy.

The proposed methods are ready to be applied to other types of statistical models with multiple response variables which have different associations with a covariate or covariates of interest, for instance, logistic regressions or log-linear models. In addition, it can be directly applied to analyze other types of data. One example will be to, in genetic or epigenetic studies, jointly cluster genes and subjects based on an association of DNA methylation or gene expression levels with measures of environmental exposures; the joint cluster patterns can then be linked to health outcomes, e.g., allergic diseases or cancer.

The methods have some limitations that warrant a discussion. The joint clustering strat-

egy identifies subject clusters within each dependent variable cluster instead of completely allowing dependent variable cluster sizes vary for different subject clusters. Our ongoing work is exploring this more flexible clustering scheme, which will have the potential to further improve the homogeneity in the clusters. In addition, the use of semi-parametric regressions via P-Splines instead of regular linear regressions can substantially increase the number of parameters to be inferred, especially in the situation of a large number of covariates. In situations like this, features represented by different variables may need to be detected first before applying the method.

Appendix

A. Algorithm for Sampling $\boldsymbol{\beta}_{i,m}$

To sample $\boldsymbol{\beta}_{i,m}$, we implement Algorithm 2 summarized in Neal (2000), initially proposed by Bush and MacEachern (1996), in which all subjects are assigned to some clusters and the Gibbs sampling process becomes more efficient by drawing only those $\boldsymbol{\beta}_{c,m}$ that are currently associated with some subjects. The conditional posterior of c_i is

$$\begin{cases} \text{if } c = c_j \text{ for some } j \neq i : P(c_i = c \Big| c_{-i}, Y, \boldsymbol{\beta}) \\ \quad = b\dfrac{n_{-i,c}}{n-1+\lambda}(2\pi)^{-\frac{k_m}{2}}|\Sigma_m|^{-\frac{1}{2}} \exp\Big[-\tfrac{1}{2}(\boldsymbol{y}_{i,m}-\boldsymbol{X}_i\boldsymbol{\beta}_{j,m})'\Sigma_m^{-1}(\boldsymbol{y}_{i,m}-\boldsymbol{X}_i\boldsymbol{\beta}_{j,m})\Big], \\ P(c_i \neq c_j \text{ for all } j \neq i \Big| c_{-i}, Y, \boldsymbol{\beta}) \\ \quad = b\dfrac{\lambda}{n-1+\lambda}(2\pi)^{-\frac{k_m}{2}}|\Sigma_m|^{-\frac{1}{2}}|\Sigma_0|^{-\frac{1}{2}}|\Sigma_{\beta_{i,m}}|^{\frac{1}{2}} \\ \qquad \exp\Big[-\tfrac{1}{2}\boldsymbol{y}_{i,m}'\Sigma_m^{-1}\boldsymbol{y}_{i,m}+\tfrac{1}{2}\boldsymbol{y}_{i,m}'\Sigma_m^{-1}\boldsymbol{X}_i\Sigma_{\beta_{i,m}}\boldsymbol{X}_i'\Sigma_m^{-1}\boldsymbol{y}_{i,m}\Big], \end{cases} \tag{4}$$

where c_{-i} denotes all c_j for $j \neq i$; $n_{-i,c}$ is the number of c_j for $j \neq i$ that are equal to c; $\boldsymbol{\beta}$ represents the set of $\boldsymbol{\beta}_{c,m}$ currently associated with at least one observation.

The posterior for $\boldsymbol{\beta}_{c,m}$ is

$$\begin{aligned} \boldsymbol{\beta}_{c,m}\Big|(\boldsymbol{y}_{i,m}, \Sigma_m, D, c) \;\sim\; & N\Big((\sum_{i=1}^{n}\boldsymbol{X}_i'\Sigma_m^{-1}\boldsymbol{X}_i+\Sigma_0^{-1})^{-1}\sum_{c_i=c}\boldsymbol{X}_i'\Sigma_m^{-1}\boldsymbol{y}_{i,m}, \\ & (\sum_{c_i=c}\boldsymbol{X}_i'\Sigma_m^{-1}\boldsymbol{X}_i+\Sigma_0^{-1})^{-1}\Big). \end{aligned}$$

B. Sampling Procedure and Determination on the Number of Joint Clusters

In the following, we summarize the process of joint clustering. At iteration (t),

1. For the k^{th} variable, draw $\boldsymbol{\pi}$ from the Dirichlet distribution, $\boldsymbol{\pi}\Big|(D_{.k}^{(t-1)}, \zeta^{(t-1)})$, and $\zeta^{(t)}$ from its conditional posterior distribution via the Metropolis-Hastings algorithm.

2. Draw $D_{.k}^{(t)}$ from the distribution of $D_{.k}\Big|(\boldsymbol{y}_{i,m}, \boldsymbol{\beta}_{i,m}^{(t-1)}, \Sigma_m^{(t-1)}, \boldsymbol{\pi}^{(t)})$ given in (3).

For each variable cluster, we apply the Dirichlet Process to cluster subjects. For subject i, $i = 1 \dots n$.

3. Draw $c_i^{(t)}$ by distribution given by (4), where the state of c is $\{c_1^{(t)}, \dots, c_{i-1}^{(t)}, c_{i+1}^{(t-1)}, \dots, c_n^{(t-1)}\}$

4. Draw Σ_m from $\Sigma_m\Big|(Y, \boldsymbol{\beta}_{i,m}^{(t-1)})$, where $\boldsymbol{\beta}_{i,m}^{(t-1)} = \boldsymbol{\beta}_{i,m}^{(t'-1)}$ with $\boldsymbol{\beta}_{i,m}^{(t'-1)}$ being the coefficients at the latest iteration $t'-1$ such that variables form the same cluster m; if cluster m is unique up to iteration t, that is, $\boldsymbol{\beta}_{i,m}^{(t-1)}$ does not exist. In this case, to initiate the sampling of cluster m, we assume no subject clusters in variable cluster m, i.e., $\beta_{i,m} = \boldsymbol{\beta}_m = (a_{m,1}, a_{m,2}, b_{m,1}, \dots, b_{m,g})$ for all i. We set $\boldsymbol{\beta}_{i,m}^{(t-1)} = \boldsymbol{\beta}_m^{(t-1)}$ with $\boldsymbol{\beta}_m^{(t-1)}$ sampled from

$$\begin{aligned}\boldsymbol{\beta}_m\Big|(Y, \sigma^2, \sigma_m^2, \Sigma_m, D) \;\sim\; & \boldsymbol{N}\Big((\sum_{i=1}^{n} \boldsymbol{X}_i'\Sigma_m^{-1}\boldsymbol{X}_i + V^{-1})^{-1} \sum_{i=1}^{n} \boldsymbol{X}_i'\Sigma_m^{-1}\boldsymbol{y}_{i,m}, \\ & (\sum_{i=1}^{n} \boldsymbol{X}_i'\Sigma_m^{-1}\boldsymbol{X}_i + V^{-1})^{-1}\Big),\end{aligned}$$

and $\sigma_m^2\Big|\boldsymbol{\beta}_m \sim \text{InvGamma}(a_1 + \frac{g}{2}, c_1 + \frac{1}{2}\sum_{l=1}^{g} b_{m,l}^2)$, where $V = V(\sigma^2, \sigma_m^2)$ is a diagonal matrix with entries σ^2 and σ_m^2. This is concluded by assuming prior distributions $\boldsymbol{\beta}_m|(\sigma^2, \sigma_m^2, D) \sim \boldsymbol{N}\big(\boldsymbol{0}, V(\sigma^2, \sigma_m^2)\big)$ and $\sigma_m^2|(a_1, c_1) \sim \text{InvGamma}(a_1, c_1)$ with $a_1 = c_1 = 0.5$, where σ^2 is the variance of $a_{m,1}$ and $a_{m,2}$ in $\boldsymbol{\beta}_m$ and assumed to be known and large ($\sigma^2 = 100$), and σ_m^2 is the variance of b_l's in $\boldsymbol{\beta}_m$.

Draw $\boldsymbol{\beta}_{c,m}^{(t)}$ currently associated with at least one subject ($\boldsymbol{\beta}_{i,m}^{(t)} = \boldsymbol{\beta}_{c,m}^{(t)}$) for all $c_i = c$.

5. Draw a new value for $\boldsymbol{\beta}_{c,m}^{(t)}$ from the posterior distribution based on the prior G_0 and all observations currently associated with latent class c, $\boldsymbol{\beta}_{c,m}\Big|(Y, \Sigma_m, D, c)$.

6. For each component of Σ_0, draw from $\Sigma_0[j]\Big|(Y, \boldsymbol{\beta}_{i,m}^{(t)})$.

C. Determining the Number of Joint Clusters

The final number of joint clusters is decided by identifying an iteration with "least-squares distance", a procedure adapted from Dahl (2006):

1. After the MCMC burn-in, continue the MCMC simulations for an additional B iterations. Let A denote an $n \times n \times K$ matrix. The $(i, j, k)^{\text{th}}$ entry of A is the proportion of iterations such that subjects i and j $(i, j = 1, ..., n)$ for the k^{th} variable are in the same cluster. The matrix A is referred as an averaged clustering matrix.

2. Continue to run an additional D_0 iterations of the MCMC simulations. For each iteration,

   (a) form an $n \times n \times K$ matrix composed of indicators of clustering for that particular iteration. For instance, if subjects i and j for the k^{th} variable are in one cluster, then the (i, j, k) entry is 1; otherwise, it is zero.

   (b) Calculate the Euclidean distance between the matrix formed above and the averaged clustering matrix A.

3. Sort the Euclidean distances obtained from the D_0 iterations, and the final selection on the number of joint clusters is in favor of simpler clusters and relatively small Euclidean distances.

References

Antoniak, C. E. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The Annals of Statistics 2*, 1152–1174.

Baladandayuthapani, V., B. K. Mallick, and R. J. Carroll (2005). Spatially adaptive Bayesian penalized regression splines (P-splines). *Journal of Computational and Graphical Statistics 14*(2), 378–394.

Bush, C. A. and S. N. MacEachern (1996). A semiparametric bayesian model for randomised block designs. *Biometrika 83*, 275–285.

Cheng, K.-O., N.-F. Law, W.-C. Siu, and T. Lau (2007). Bivisu: software tool for bicluster detection and visualization. *Bioinformatics 23*, 2342–2344.

Cheng, Y. and G. M. Church (2000). Biclustering of expression data. In *Ismb*, Volume 8, pp. 93–103.

Dahl, D. (2006). Model-based clustering for expression data via a Dirichlet process mixture model*, in Bayesian Inference for Gene Expression and Proteomics,* Do, K., Müller, P., Vannucci, M. (Eds.). Cambridge University Press, Cambridge.

Dorazio, R. M., B. Mukherjee, L. Zhang, M. Ghosh, H. L. Jelks, and F. Jordan (2008). Modeling unobserved sources of heterogeneity in animal abundance using a dirichlet process prior. *Biometrics 64*, 635–644.

Doss, H. (2008). Estimation of Bayes factors for nonparametric Bayes problems via Radon-Nikodym derivatives. Technical report, Technical Report, University of Florida, Department of Statistics.

Doss, H. (2012). Hyperparameter and model selection for nonparametric bayes problems via radon–nikodym derivatives. *Statistica Sinica 22*, 1–26.

Eilers, P. H. and B. D. Marx (1996). Flexible smoothing with B-splines and penalties. *Statistical Science 11*, 89–102.

Freitas, A. V., W. Ayadi, M. Elloumi, J. Oliveira, J. Oliveira, and J.-K. Hao (2013). *Survey on Biclustering of Gene Expression Data*, pp. 591–608. John Wiley & Sons, Inc.

Good, I. J. (1965). *The estimation of probabilities: An essay on modern Bayesian methods*, Volume 30. MIT press.

Gu, J. and J. S. Liu (2008). Bayesian biclustering of gene expression data. *BMC genomics 9*, S4.

Hartigan, J. A. (1972). Direct clustering of a data matrix. *Journal of the American statistical association 67*(337), 123–129.

Hide, D., S. Matthews, S. Tariq, and S. Arshad (1996). Allergen avoidance in infancy and allergy at 4 years of age. *Allergy 51*(2), 89–93.

Kaiser, S. and F. Leisch (2008). A toolbox for bicluster analysis in R. Technical report, Department of Statistics,University of Munich.

Kass, R. E. and L. Wasserman (1995). A reference bayesian test for nested hypotheses and its relationship to the schwarz criterion. *Journal of the american statistical association 90*, 928–934.

Kyung, M., J. Gill, G. Casella, et al. (2010). Estimation in dirichlet random effects models. *The Annals of Statistics 38*, 979–1009.

Lee, J., P. Müller, Y. Zhu, and Y. Ji (2013). A nonparametric Bayesian model for local clustering with application to proteomics. *Journal of the American Statistical Association 108*(503), 775–788.

Liu, J. S. (1996). Nonparametric hierarchical bayes via sequential imputations. *The Annals of Statistics 24*, 911–930.

Madeira, S. C. and A. L. Oliveira (2009). A polynomial time biclustering algorithm for finding approximate expression patterns in gene expression time series. *Algorithms for Molecular Biology 4*, 8.

McAuliffe, J. D., D. M. Blei, and M. I. Jordan (2006). Nonparametric empirical Bayes for the Dirichlet process mixture model. *Statistics and Computing 16*, 5–14.

Meeds, E. and S. Roweis (2007). Nonparametric Bayesian biclustering. Technical report, Department of Computer Science, University of Toronto.

Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of computational and graphical statistics 9*, 249–265.

Prelić, A., S. Bleuler, P. Zimmermann, A. Wille, P. Bühlmann, W. Gruissem, L. Hennig, L. Thiele, and E. Zitzler (2006). A systematic comparison and evaluation of biclustering methods for gene expression data. *Bioinformatics 22*, 1122–1129.

Ruppert, D., M. P. Wand, and R. J. Carroll (2003). *Semiparametric regression*. Number 12. Cambridge university press.

Santamaría, R., R. Therón, and L. Quintales (2008). A visual analytics approach for understanding biclustering results from microarray data. *BMC bioinformatics 9*, 247.

Spiegelhalter, D. J., N. G. Best, B. P. Carlin, and A. Van Der Linde (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 64*, 583–639.

Wu, C.-J. and S. Kasif (2005). Gems: a web server for biclustering analysis of expression data. *Nucleic acids research 33*, W596–W599.