Biostatistics (2008)
XXXXX/biostatistics/XXXX
Advance Access publication on XXXXX

Bayesian Profile Regression with an Application to the National Survey of Children's Health

JOHN MOLITOR[1*], MICHAIL PAPATHOMAS[1], MICHAEL JERRETT[2] and SYLVIA RICHARDSON[1]

[1]*Centre for Biostatistics, Imperial College, London*
Department of Epidemiology and Public Health
Imperial College, St Mary's Campus Norfolk Place London W2 1PG
Email: john.molitor@imperial.ac.uk
[2]*School of Public Health, University of California, Berkeley*

SUMMARY

Standard regression analyses are often plagued with problems encountered when one tries to make meaningful inference going beyond main effects, using datasets that contain dozens of variables that are potentially correlated. This situation arises, for example, in epidemiology where surveys or study questionnaires consisting of a large number of questions, yield a potentially unwieldy set of inter-related data from which teasing out the effect of multiple covariates is difficult. We propose a method that addresses these problems for categorical covariates by using, as its basic unit of inference, a profile, formed from a sequence of covariate values. These covariate profiles are clustered into groups and associated via a regression model to a relevant outcome. The Bayesian clustering aspect of the proposed modeling framework has a number of advantages over traditional clustering approaches in that it allows the number of groups to vary, uncovers subgroups and examines their association with an outcome of interest and fits the model as a unit, allowing an individual's outcome to potentially influence cluster membership. The method is demonstrated with an

*To whom correspondence should be addressed.

analysis of survey data obtained from The National Survey of Children's Health (NSCH). The approach has been implemented using the standard Bayesian modeling software, WinBUGS, with code provided in the supplementary material. Further, interpretation of partitions of the data is helped by a number of post-processing tools that we have developed.

Keywords: Profile Regression; Dirichlet Process, clustering, Bayesian analysis; MCMC

1. INTRODUCTION

A common problem encountered in a regression setting is the difficulty involved in making meaningful inference from data containing a large number of inter-related explanatory variables, such as data arising from detailed questionnaires. The covariates in these datasets are often confounded (aliased) with each other, meaning that the association between the outcome and one specific covariate, x_p, may achieve a high level of statistical significance by itself, but not in the presence of many other related covariates. Additionally, the effect of a particular covariate on the outcome might only be revealed in the presence of other covariates. Therefore, the overall pattern of joint effects may be elusive, and hard to capture by traditional analyses that include main effects and interactions of increasing order, as the model space becomes soon unwieldy and power to find any effects beyond simple two-way interactions quickly vanishes.

One way to deal with the above mentioned problems is to adopt a more global point of view, where inference is based on clusters representing covariate patterns as opposed to individual risk factors. This general approach has been suggested in epidemiology in recently published commentaries as a possible method for examining aging profiles (Wang, 2006) and dietary patterns (Tucker, 2007; van Dam, 2005). In that spirit, we use as the main unit of inference an individual's covariate profile, where a profile consists of a particular sequence of categorical covariate values, $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$, and associate the entire

profile pattern with the outcome.

The idea of utilizing clustering to profile correlated data is not new, and many techniques have been proposed (see, for example, Forgy, 1965; Hartigan and Wong, 1979). For instance, an analysis of dietary data using Latent Class Analysis (LCA) in a frequentist context was demonstrated by Patterson et al. (2002). Recent developments in LCA techniques to analyze correlated data can be found in Desantis et al. (2008, 2009). However, the modeling framework introduced here combines many recent developments and offers a number of advantages over traditional approaches. First, it utilizes a Bayesian mixture-model framework (Diebolt and Robert, 1994; Richardson and Green, 1997) that takes into account the uncertainty associated with cluster assignments, i.e. it employs model-based stochastic clustering as opposed to traditional distance-metric "hard" clustering. Appropriately, the model is fitted using Markov chain Monte Carlo (MCMC) sampling methods (see, for example, Gilks et al., 1996), and outputs a different clustering or *partition* of the data at each iteration of the sampler, thus coherently propagating uncertainty. Second, the method allows the number of clusters to be variable. Third, the method links clusters to an outcome of interest via a regression model so that the outcome and the clusters mutually inform each other. Finally, the method allows the analyst to examine the "best" or most typical partition of the data obtained from the algorithm (as described in Dahl, 2006), and then utilizes model-averaging techniques to assess, using the posterior output obtained from the sampler, the uncertainly associated with subgroups contained within this "best" partition. This last point is especially important, since Bayesian clustering models produce rich output and interpretation of results from such models can be challenging.

In this manuscript, we first describe the method with special emphasis paid to interpretation of model output. We then provide a brief simulation section demonstrating the performance of the model both in the presence and absence of a well-defined signal in the data. We then demonstrate the utility of the method

on an analysis of an epidemiological dataset obtained from the National Survey of Children's Health (NSCH) (`www.childhealthdata.org`). Finally we discuss model limitations and outline areas of future research.

2. METHODS

Our approach consists of an *assignment sub-model*, which assigns individual profiles to clusters, and a *disease sub-model*, which links clusters of profiles to an outcome of interest via a regression model. As is typical with Bayesian methods, both sub-models will be fitted jointly using Markov chain Monte Carlo methods (Gilks et al., 1996), so, for example, allocation of individual profiles to clusters will depend on both the covariate data in the assignment sub-model, and the outcome information in the disease sub-model. Both these sub-models will be addressed in turn.

2.1 *Assignment sub-model*

We first construct an allocation sub-model of the probability that an individual is assigned to a particular cluster. The basic model we use to cluster profiles is a standard discrete mixture model, the kind described in Jain and Neal (2004) or Neal (2000). Our mixture model incorporates a Dirichlet process prior on the mixing distribution. The use of the Dirichlet process in statistical modelling has been thoroughly examined in Walker et al. (1999). A good overview of Dirichlet process mixture models can be found in West et al. (1994), while a biomedical example of their application can be found in Mueller and Rosner (1997). For further background information regarding mixture models with Dirichlet process priors, see Escobar and West (1995); Green and Richardson (2001); MacEachern and Muller (1998); Neal (2000).

Mathematically, we denote, for individual i, a covariate profile as, $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{iP})$. Profiles

are clustered into groups, and an allocation variable, $z_i = c$, indicates the c^{th} cluster to which individual, i, belongs. We restrict our approach to categorical covariates with M_p categories for the p^{th} covariate. We denote with ψ_c the probability of assignment to the c^{th} cluster and let $\phi_c^p(x)$ denote the probability that the p^{th} covariate in cluster c is equal to x. In other words, for each cluster, c, the parameters, $\phi_c^p,\ p = 1, \ldots, P$ define the prototypical profile for that cluster. Our basic mixture model for assignment is,

$$\begin{aligned} \Pr(\boldsymbol{x}_i) &= \sum_{c=1}^{C} \Pr(z_c = c) \prod_{p=1}^{P} \Pr(x_{ip} | z_c = c) \\ &= \sum_{c=1}^{C} \psi_c \prod_{p=1}^{P} \phi_{z_i}^p(x_{ip}). \end{aligned} \tag{2.1}$$

Note that as is typical with discrete mixture models, covariates are assumed to be independent conditional on cluster assignment. Unconditionally, they are of course dependent as a profile's overall covariate pattern will affect the cluster to which the profile is assigned, and thus the probability that a particular covariate takes on a certain value. In this manuscript, we only analyze datasets with binary covariates, and so we use the notation ϕ_c^p to indicate the probability that a variable belonging to cluster, c, takes a value of 1.

The mixture weights corresponding to a maximum of C clusters, denoted as $\boldsymbol{\psi} = (\psi_c,\ c = 1, \ldots, C)$, will be modeled according to a "stick-breaking" prior (Ishwaran and James, 2001; Ohlssen et al., 2007) on the mixture weights, $\boldsymbol{\psi}$, using the following construction. We define a series of independent random variables, $V_1, V_2, \ldots, V_{C-1}$, each having distribution $V_c \sim Beta(1, \alpha)$. This generative process is referred to as a stick-breaking formulation since one can think of V_1 as representing the breakage of a stick of length 1, leaving a remainder of $(1 - V_1)$, and then a proportion V_2 begin broken off leaving $(1 - V_1)(1 - V_2)$, etc. Since we have little *a priori* information regarding the specification of α, we place a uniform prior on the $(0.3, 10)$ interval. This parameter is important, since it determines the degree of clustering that takes place, and we want this to be driven by the data as opposed to prior beliefs. An interval bounded on

the left by 0.3 was suggested in Ohlssen et al. (2007), so that potential computational traps in WinBUGS (Spiegelhalter et al., 2003) are avoided. In our analyses, the sampled values for α were always well away from the chosen bounds of this prior specification.

By considering a maximum number of clusters, C, we have approximated the infinite cluster model with a finite one. We need to set C to a large enough value so that the approximation is good, however, we want to choose a value that is not too large to avoid having to estimate a large number of unnecessary cluster parameters and allocation probabilities for very small clusters. To obtain some insight on what an appropriate value for C may be, we proceed along the lines of Ohlssen et al. (2007), where C is set to a value so that the probability assigned to ψ_C is small. To make sure that we allow for enough clusters we always specify $C = 20$. This corresponds to a relatively large value of $\alpha = 3.6$, while posterior values of α obtained from analyses performed in this manuscript were generally in the range of $\alpha \in (0.5, 2.5)$. Thus, the upper bound chosen imposes little structure in practise.

2.2 *Disease sub-model*

The previously described assignment model clusters individuals into groups and these cluster assignments can be simultaneously used as categorical predictors of an outcome. As above, we define allocation variables for each individual as $z_i = c,\ c = 1, \ldots, C$, which indicates the c^{th} cluster to which individual i belongs. The c^{th} cluster is assigned a parameter that measures its influence on the outcome (on the logistic scale) denoted as θ_c. Since it is possible for a particular θ_c to be associated with an empty cluster, these parameters must be assigned a proper prior. Therefore we assign to each θ_c a proper t density function with 7 degrees of freedom and scale 2.5 as a prior, as discussed in Gelman et al. (2008), which corresponds to the baseline case of one-half of a success and one-half of a failure for a single binomial

trial with probability $p = \text{logit}^{-1}(\theta_c)$. Below, we build a disease model which links the clusters with the outcome.

The general form of our disease model not only quantifies the association between the health outcome and the cluster profiles, but also allows for a number of fixed covariates to be included, as would be needed in order to adjust for known confounders. We denote, for individual i, $i = 1, \ldots, N$, confounding covariates $\boldsymbol{w}_i$, $(w_{ip},\ p = 1, \ldots, P)$. Given a binary outcome, y_i, and a corresponding probability $p_i = \Pr(y_i = 1)$, our disease model is then,

$$\text{logit}(p_i) = \theta_{z_i} + \boldsymbol{\beta}\boldsymbol{w}_i, \tag{2.2}$$

where logit denotes the standard logistic link function, and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_P)$ denotes the regression parameter coefficients associated with the confounding covariates $\boldsymbol{w}_i = (w_{i1}, \ldots, w_{iP})$. Note that in this model, θ_{z_i} is an individual-level intercept term which can be interpreted as the baseline log odds for individual i, which is the log odds obtained when all confounders are set to their "reference" value of zero. These individual-level intercepts are identifiable because they are smoothed via the clustering modeled defined in equation (2.1). Due to the clustering aspect of the model, at each iteration of the sampler individuals assigned to the same cluster will be assigned the same baseline log odds. However, each individual will have its own unique distribution for θ_{z_i} when the sampler is complete. Further, for a prospective study, we can calculate an individual-level baseline risk for disease for individual, i, as $p_{z_i} = \exp(\theta_{z_i})/[1 + \exp(\theta_{z_i})]$.

The model is fitted via Markov chain Monte Carlo (MCMC) methods (Gilks et al., 1996), where, at each iteration of the MCMC sampler, individual profiles are assigned to clusters, and each individual is assigned the risk associated with the cluster to which the individual belongs. Code for the software package WinBUGS (Spiegelhalter et al., 2003), used to perform the MCMC parameter estimation, is provided in

Section 1 of the supplementary material (available at Biostatistics online, http://www.biostatistics.oxfordjournals.org).

3. EXAMINING CLUSTERING OUTPUT

Our model implementation allows the number of groups to change from iteration to iteration of the sampler, and this added flexibility leads to a rich output that requires careful interpretation. Below we develop methods to process the output of our method to make useful, interpretable inference. There are two main areas of interest, namely (i) find the partition (grouping) that is most supported by the data, and (ii) assess uncertainty associated with subgroups of this best partition in a manner which exploits the MCMC output of the sampler. We discuss these issues in turn.

3.1 *Finding the best partition*

We wish to find the general, "typical" way in which the stochastic algorithm groups subjects into clusters. This problem has been addressed in the literature by many authors in the context of mixture models; see, for example, Dahl (2006); Medvedovic and Sivaganesan (2002). The starting point is to construct, at each iteration of the sampler, a score matrix with each element of the matrix set equal to 1 if individuals i and j belong to the same cluster, and zero otherwise. At the end of the estimation process, a probability matrix, $\boldsymbol{S}$, is formed by averaging the score matrices obtained at each iteration, so element S_{ij} denotes the probability that individuals i and j are assigned to the same cluster. The task is then to find the partition, $\boldsymbol{z}^{best}$, that best represents the final average probability matrix, $\boldsymbol{S}$. Dahl (2006) suggests an approach to finding the best partition by choosing among all the partitions generated by the sampler the one which minimizes the least-squared distance to the matrix $\boldsymbol{S}$. We have found this approach useful, however, it requires one to choose one of the observed partitions as optimal, resulting in a choice that is somewhat

susceptible to Monte-Carlo error. We find that a more robust approach is to process the similarity matrix, $\boldsymbol{S}$, through a deterministic clustering procedure such as the Partitioning Around Medoids (PAM) (Kaufman and Rousseeuw, 2005), a deterministic clustering method available in R (R Development Core Team, 2006), where an optimal number of clusters can be chosen by maximizing an associated clustering score. This clustering method robustly provides a set of assignments of individuals to clusters that can be used to summarize the pairwise similarity matrix, $\boldsymbol{S}$. Note that we found that the PAM approach and the approach of Dahl (2006) often produce very similar results.

3.2 *Evaluating uncertainty associated with best partition - a model averaging approach*

It is important to examine, with proper consideration for uncertainty, the characteristics associated with the subgroups present in any chosen *partition* of the dataset. For clarity, we demonstrate this for the partition $\boldsymbol{z}^{best}$ described above, but the concepts that we define apply to any given partition. The basic idea is to take the partition $\boldsymbol{z}^{best}$, representing the "best" clustering of the data, and to examine by post-processing whether or not the model consistently clusters individuals in a manner similar to $\boldsymbol{z}^{best}$. Consistent clustering will be associated with greater certainty regarding subgroup parameter estimates, such as disease risk, leading to narrower posterior credible intervals. For example, in a dataset with a strong clustering "signal", the model may cluster individuals slightly differently at each iteration of the MCMC sampler, but, due to the strength of the signal in the data, will generally cluster individuals with a good degree of repeatability over the iterations of the sampler. However, if the data is "noisy" in that individuals do not tend to group into clusters, the clustering obtained from the model will tend to be haphazard and highly variable. While even noisy data will exhibit a "best" clustering, a re-examination of the entire MCMC output will reveal little confidence in this clustering as it will not generally coincide with the way individuals are clustered

at each iteration of the sampler. Thus evaluating uncertainty is important for interpretation.

We wish to obtain a distribution of the baseline risks for each subgroup defined by $\boldsymbol{z}^{best}$. We do this by simply computing, at each iteration of the sampler, the average of baseline risks, p_{z_i}, for all individuals within a particular subgroup, k, of the best partition. This average baseline risk for subgroup k is computed as,

$$\bar{p}_k = \frac{1}{n_k} \sum_{i:z_i^{best}=k} p_{z_i}, \tag{3.3}$$

where n_k denotes the number of individuals in subgroup k of $\boldsymbol{z}^{best}$. Note that if $\boldsymbol{z}^{best}$ coincides with a partition found at a particular iteration of the sampler, then at this particular iteration, all individuals in a subgroup of z^{best}, k, all belong to one cluster, say, cluster, c, and $\bar{p}_k = p_c$. However, at other iterations of the sampler, computation of $\bar{p}_k$ allows different subgroups to borrow strength from one another in computing subgroup parameter values. Subgroup parameter values corresponding to covariate probabilities, $\bar{\phi}_k^p$, can be computed similarly as,

$$\bar{\phi}_k^p = \frac{1}{n_k} \sum_{i:z_i^{best}=k} \phi_{z_i}^p. \tag{3.4}$$

Note that instead of using means in equations (3.3) and (3.4), one could use medians if it is believed that the posterior distribution of the particular parameter is skewed, which is likely to happen if, for example, the posterior mass for a particular $\bar{\phi}_k^p$ is close to zero or one.

For interpretation purposes, we define new centered baseline risk parameters, $\bar{p}_k^* = \bar{p}_k - \bar{\bar{p}}$, so that $\sum_{k=1}^K \bar{p}_k^* = 0$, and define similar centered covariate parameters, $\bar{\phi}_k^{p*}$. These centered parameters are computed easily at each step of the MCMC sampler via the post-processing steps. A useful summary derived from the sampled values for the $\bar{p}_k^*$ parameters is the probability $P(\bar{p}_k^* > 0)$ (for high risk groups) or $P(\bar{p}_k^* < 0)$ (for low risk groups). This probability is calculated by considering the frequency of positive $\bar{p}_k^*$ in the sample. The closer these posterior probabilities are to one, the stronger evidence there is that the

particular subgroup has high or low risk for disease. Similar summaries are derived for each $\bar{\phi}_k^{p*}$.

Regardless of the procedure used for choosing $\boldsymbol{z}^{best}$, we stress that the groups in this partition, as any partition, should be post-processed through the output of the sampler in this way in order to properly assess uncertainty for group parameters. This post-processing approach represents a compromise between an examination of interpretable "hard groupings", as exemplified by $\boldsymbol{z}^{best}$, and inspection of raw output from a random mixture model. In other words, while we may choose a "best" partition for interpretation purposes, we utilize a model averaging approach to process this partition through the rich MCMC output to characterize its uncertainty.

3.3 *Illustration of model performance using simulated data*

We performed a small simulation exercise to illustrate the performance of the model both in the presence of a strong signal and in the case where two of the sub-populations are nearly identical. We first created a simulated data set of sample size $N = 600$ using model (2.2) with a binary outcome, $y \in \{0, 1\}$, and $P = 10$ binary covariates, $\boldsymbol{x} = (x_1, \ldots, x_{10})$ with no confounders. For purposes of comparison, we report the cluster parameters, $\bar{p}_k$ and $\bar{\phi}_k^p$ as both the mean and the median of relevant individual-level subgroup parameters stimulated at each iteration of the sampler. In Table 1 of the supplementary material we give the simulation parameters and parameter estimates obtained from the MCMC sampler, implementing model (2.2) with no confounders. We simulated five subgroups with equal prior assignment probabilities, namely $\psi_1 = \psi_2 = \ldots = \psi_5 = 1/5$, and profiles with clearly distinct patterns. In this case, our method found a best partition, z^{best}, with the correct number of subgroups (5) and estimated relevant cluster parameters well.

Next, in order to examine the situation where the signal was not as strong, we simulated data were

individuals fell into one of three subgroups with equal probability, however, two of the profiles had very similar covariate parameter values $\bar{\phi}_k^p$, as shown in Table 2 in the Supplementary material. We analysed the simulated output, again under model (2.2) with no confounders, with results for covariate parameter estimates displayed in Table 2. Note here that the model only found two subgroups, effectively collapsing the similar clusters into one and averaging the parameter estimates. This demonstrates that when the signal is weak, the clustering will effectively group together profiles with similar patterns of covariates. We investigated a range of simulated cases (results not reported) and found that our model and algorithm was able to capture adequately main patterns, with a degree of 'blur' related to group size. When no true structure is present in the covariates, typically only one cluster is formed using Dahl's partition method, whereas the default of the PAM method can only support a minimum of two clusters. Hence, with very weak structure, Dahl's method may be preferable.

4. DATA ANALYSIS - NATIONAL SURVEY OF CHILDREN'S HEALTH

The data analyzed in this section come from The National Survey of Children's Health (NSCH), a U.S. national survey that was conducted by telephone in English and Spanish during 2003-2004. This survey was conducted as part of the Child and Adolescent Health Measurement Initiative (CAHMI) (`www.childhealthdata.org`) . CAHMI is a national initiative based out of the Oregon Health and Science University in the Department of Pediatrics in Portland, OR.

The dataset consists of responses to a wide variety of health-related questions. In addition to the raw survey questions, an enhanced dataset is available that includes indicators developed by the Data Resource Center in collaboration with the National Center for Health Statistics and a national expert panel of child health researchers and policy makers. These indicators are derived variables, often made up of responses

to two or more related questions. As a way to demonstrate the utility of the approach described in this manuscript, we analyzed profiles made up of indicators together with a few basic variables. (See Table 3 in the Supplementary material for a detailed description of the data.) We eliminated variables consisting of follow-up questions. We further eliminated any variables that contained more than 40% missing data.

The data cover children from different age groups in all 50 U.S. states. For our illustrative analysis, we restricted ourselves to children in the age category of $6 - 17$ years, and to children residing in the state of California. We chose, as an outcome of interest, the mental health of a child, a derived variable where an observed value of one indicates that at least one child in the household had ongoing emotional, developmental, or behavioral conditions that required treatment or counseling. Since this outcome was derived from mental health variables, we eliminated the other mental health variables from the list of predictors. This reduced our dataset to 34 variables (listed in Table 3 of the Supplementary material). Including all individuals living in California with a complete profile of these 34 variables, we obtain a sample size of $N = 642$. This dataset is sufficient, given space considerations, to demonstrate the utility of the method, though a more complete analysis, perhaps the subject of a separate substantive paper, would be desirable.

The data were analyzed using the methods described in this paper and the provided WinBUGS code (Section 1 of the supplementary material), along with additional post-processing code written in R (R Development Core Team, 2006). For all real data analyses performed in this paper, the algorithm was run for a $50,000$ iterations with $10,000$ iterations discarded for burn-in. Visual inspection of posterior time-series plots for the $\bar{\phi}$'s and $\vec{p}$'s indicated that the model mixed well, and shorter runs gave very similar results, indicating that convergence was not an issue.

4.1 *Results*

Here, we provide data analysis results from two different approaches; standard logistic regression combined with stepwise variable selection and profile regression.

Logistic Regression

We first examined the data using traditional logistic regression analysis methods implementing the software package, R (R Development Core Team, 2006). As a first step, we ran forward stepwise selection, forcing four variables as confounders, three demographic variables, "young_school_age", "non_white" and "male", as well as "mother" which indicates that mother was the respondent. The stepwise procedure did produce a final model, however, due to the highly correlated nature of the covariates, R gave warnings suggesting that the maximum likelihood estimates may not be reliable. Problems associated with using standard maximum likelihood approaches for analyzing correlated data are well known; see MacLehose et al. (2007). Therefore, we trimmed the stepwise results by only keeping covariates with $p < 0.05$ and then refit the model with results listed in Table 1. We next formed a model consisting of all covariates and all two-way interaction terms made up of these covariates. Again, a final model was obtained by running forward selection but as previously, warnings were produced, requiring that we refit the model after eliminating all non-significant covariates and interactions, with results given in Table 2.

Results displayed in Tables 1 and 2 highlight the influence of family habits and health access on the risk of mental health problems for the child. Psychological problems of the mother, smoking in the household and not getting enough sleep were detrimental. With regards to health access, the covariate "medical home" was highly significant, which is defined by the American Academy of Pediatrics as, "accessible, continuous, comprehensive, family centered, coordinated, compassionate, and culturally effective" (The

medical home, 2002). The coefficient for this variable was negative, indicating that children who live in medical homes have reduced risk of having mental health problems. On the opposite, emergency admission and spending a lot of time with your personal doctor were, as could be expected, associated with higher risk. Other variables reducing the risk were the variable "activity", which is related to physical or social activity of the child, and "rep-grade" (repeating a grade) which could be both interpreted as indicating less child stress. Note that the coefficient for language was negative, suggesting that children whose primary language is not English have, in this data set, lower risk of having mental health problems. This seemingly contradictory result also comes up in the subsequent profile analyses, and will be discussed later in Section 4.1.

Profile regression

We analysed the data using the profile approach described in this manuscript, using the model corresponding to equation (2.2). As with the standard logistic regression analysis, we included covariates, "young_school_age", "non_white", "male" and "mother", as confounders, and then included all environmentally-influenced covariates as clustering variables.

The post-processing methods described in Section 3 revealed a "best" partition of six subgroups. Two of these subgroups were "statistically significant" in that they were associated with high posterior probabilities that centered values for $\bar{p}^*$ were away from zero. One of these subgroups was associated with low mental health risk while the other was associated with high risk. The other four subgroups can be thought of as "baseline subgroups", representing different combinations of covariate values associated with mental health risks closer to average.

The subgroup strongly associated with low mental health risk for the child ($\bar{p}^* = -0.25$ and $\Pr(\bar{p}^* < 0) = 0.92$) is depicted in Figure 1(a). Somewhat surprisingly, this low-risk subgroup contains a large

number of non-English speaking individuals and is characterized by such seemingly detrimental characteristics as low education, low level of activity, poor maternal health, and poor medical care access. The low risk of mental health problems associated with this subgroup suggest the possibility that cultural factors are influencing parents' attitudes towards mental-health medical care that could potentially induce under-reporting. This phenomenon has been described in the literature before, for example, Yeh et al. (2003) reported findings of a study where they found that "ethnic minority youth had higher levels of unmet need" though it was suggested that certain portions of the sample, such as Latinos "did not want to use mental health services due to a culturally severe stigma associated with such service use".

The other statistically significant subgroup is associated with high mental health risk for the child, with $\bar{p}^* = 0.36$ and $\Pr(\bar{p}^* > 0) \approx 1.00$, and is depicted in Figure 1(b). This subgroup is mostly English-speaking and exhibits, as could be expected, a coherent combination of behavioral and medical problems for the child (high values for such variables as "miss_school" and "c_asthma") and maternal health problems along with high levels of maternal smoking. In Figure 2, we display two English-speaking subgroups, both of which have risks that are close to the average. Comparing Figure 1(b) with Figures 2(a) and 2(b), we see that these subgroups are associated with "healthier" communities (high values for "support neighbor") and family structures ("twoparent") and are associated with lower levels of maternal smoking together with lower risk of maternal health problems. Hence the pattern of covariate values are clearly contrasted with those of Figure 1(b), a subgroup associated with an above average risk of mental health problems. Note that the two profiles in Figure 2 differ mainly with respect to health access, but that this is not reflected in any difference in the risks of mental health problems.

5. DISCUSSION

We have described a new analytical strategy which uses a covariate pattern, or profile, as the basic unit of inference, and examines associations between these profiles and an outcome of interest. Some of the ingredients of our approach are well established, but, to our knowledge, have not all been put together in the manner described in this manuscript to create a unified, easy-to-implement method for analyzing data with a sizable number of interactive variables. Our method groups profiles into clusters, and the number of clusters is allowed to be random. Post-processing techniques help determine interesting partitions of the data, and allow the analyst to construct interpretable inference based on these partitions. Parameters are associated with clusters, and these are used in turn in a regression model of an outcome of interest.

We have used a simple formulation of the mixture model with conditionally independent cluster probabilities for each binary covariate given cluster membership. Extensions of the model to allow for additional dependence as well inclusion of continuous covariates could be envisioned. Such multilevel extensions will be the subject of future research. We have focused on epidemiological interpretation of the profiles in our analysis of the NSCH data, but the method could be applicable for classification problems in other contexts, for example to characterize deprivation and neighborhood conditions in social studies of small area characteristics, going beyond the simple summary indices typically used.

The method was implemented using standard Bayesian modeling software (provided), along with simple post-processing scripts, making the method easy to implement and accessible to a wide audience. While our WinBUGS implementation makes the method more transparent and user friendly, we have simultaneously developed MATLAB code that will allow larger number of variables with several categories to be efficiently analysed, as well as incorporating model extensions like ordinal covariate modeling using an underlying probit model. Note also that while we analysed data with full covariate information, missing

values can be accommodated simply in our Bayesian setting and with our WinBUGS implementation by denoting each missing value as 'NA', causing it to be multiply imputed throughout the sampler using full model information (see Spiegelhalter et al., 2003).

Our model was formulated in a Dirichlet Process framework allowing for flexibility in the number of cluster used. However, our general profiling approach, including the post-processing steps incorporating Bayesian model averaging, could be formulated using mixture weights that follow a different mixture model, for example a model that allows for a flexible number of clusters estimated via Reversible Jump Markov chain Monte Carlo (RJMCMC) techniques as in Green and Richardson (2001). The DP approach has the advantage of being easy to implement in standard Bayesian modeling software such as WinBUGS, and thus provides a convenient way to model heterogeneity (Ohlssen et al., 2007). However, as the sampler progresses, clusters containing only one or two individuals are sometimes observed, which could lead to estimation problems for small sized datasets. The finite mixture model approach does not tend to have this problem, but requires nonstandard split/merge moves as part of the RJMCMC estimation procedure.

Covariate selection in multivariate regression for health modeling is often problematic because of the wide range of possible predictors, collinearity and the potential for interactions. The proposed modeling framework sidesteps this traditional approach and proposes instead to cluster covariates of interest into subgroups, which can avoid problems of instability in picking out a small number of so-called significant covariates among a larger set of multicollinear ones as is commonly done in epidemiological practise. Indeed, using forward selection in our case study was problematic and needed to be followed by somewhat arbitrary trimming in order to produce interpretable results. Of course, more sophisticated statistical approaches could be employed to stabilize multivariate regression (MacLehose et al., 2007). Alternatively, one could utilize the Localized Regression techniques proposed by Tutz and Binder (2005) where splines

are used to allocate similar individuals to clusters. However, this latter approach focuses more on cluster assignment and variable selection, and not on interpretation of subgroups and their associated risks with an outcome of interest. In contrast, our approach embraces multicollinearity by highlighting coherent patterns and combination of variables influencing the health outcome.

Variables within the profile that explain the contrast between the subgroups with high probability can be highlighted, thus adding to the interpretability of the clusters. Using this framework on a population health survey, we have demonstrated some benefits of using the approach presented over the traditional logistic regression. However, we note that logistic regression aims at estimation of main effects and interaction terms, whilst the profile approach described in this manuscript is aimed at the examination of a combination of variables that structure the variability of the data. Since the two approaches address different characteristics of association, both should be used in a complementary fashion to progress our understanding of the association between an outcome and a set of correlated covariates.

REFERENCES

The medical home. *Pediatrics*, 110(1.1):184–6, 2002.

Child and A. H. M. I. (CAHMI). *2003 National Survey of Children's Health Indicator Data Set*. URL `www.childhealthdata.org`.

D. Dahl. *Bayesian Inference for Gene Expression and Proteomics*, chapter Model-Based Clustering for Expression Data via a Dirichlet Process Mixture Model. Cambridge University Press, 2006.

S. M. Desantis, E. A. Houseman, B. A. Coull, A. Stemmer-Rachamimov, and R. A. Betensky. A penalized latent class model for ordinal data. *Biostatistics*, 9(2):249–62, 2008.

S. M. Desantis, E. A. Houseman, B. A. Coull, D. N. Louis, G. Mohapatra, and R. A. Betensky. A latent class model with hidden markov dependence for array cgh data. *Biometrics*, 2009.

J. Diebolt and C. P. Robert. Estimation of finite mixture distributions through Bayesian sampling. *J. Roy. Statist. Soc. B*, 56:363–375, 1994.

M. Escobar and M. West. Bayesian density estimation and inference using mixtures. *J. Amer. Statist. Assoc.*, 90: 577–588, 1995.

B. Everitt. *An introduction to latent variable models*. Monographs on statistics and applied probability. Chapman and Hall, London ; New York, 1984.

B. Everitt and D. J. Hand. *Finite mixture distributions*. Monographs on applied probability and statistics. Chapman and Hall, London ; New York, 1981.

E. Forgy. Cluster analysis of multivariate data: efficiency vs interpretability of classifications. *Biometrics*, 21:768–769, 1965.

A. Gelman, A. Jakulin, M. Pittau, and Y.-S. Su. A weakly informative default prior distribution for logistic and other regression models. *Annals of Applied Statistics*, 2(4):1360–1383, 2008.

W. Gilks, S. Richardson, and D. Spiegelhalter, editors. *Markov Chain Monte Carlo in Practice*. Chapman & Hall, London, 1996.

P. Green and S. Richardson. Modelling heterogeneity with and without the Dirichlet process. *Scandinavian Journal of Statistics*, 28(2):355–375, 2001.

J. Hartigan and M. Wong. A k-means clustering algorithm. *Applied Statistics*, 28:100–108, 1979.

H. Ishwaran and L. James. Gibbs sampling methods for stick-breaking priors. *J. Amer. Statist. Assoc.*, 96:161–173, 2001.

S. Jain and R. Neal. A split-merge Markov chain Monte carlo procedure for the Dirichlet process mixture model. *Journal of computational and Graphical Statistics*, 13:158–182, 2004.

L. Kaufman and P. J. Rousseeuw. *Finding groups in data : an introduction to cluster analysis*. Wiley series in probability and mathematical statistics. Wiley-Interscience, Hoboken, N.J., 2005.

P. Lazarfeld. *Measurement and Prediction*. Princeton University Press, Princeton, NJ, 1950.

P. F. Lazarsfeld and N. W. Henry. *Latent structure analysis*. Houghton Mifflin, New York, 1968.

S. N. MacEachern and P. Muller. Estimating mixture of dirichlet process models. *Journal of Computational and Graphical Statistics*, 7:223–238, 1998.

R. F. MacLehose, D. B. Dunson, A. H. Herring, and J. A. Hoppin. Bayesian methods for highly correlated exposure data. *Epidemiology*, 18(2):199–207, 2007.

T. McHugh. Efficient estimation and local identification in latent class analysis. *Psychometrika*, 21:331–347, 1956.

G. J. McLachlan and K. E. Basford. *Mixture models : inference and applications to clustering*. M. Dekker, New York, N.Y., 1988.

M. Medvedovic and S. Sivaganesan. Bayesian infinite mixture model based clustering of gene expression profiles. *Bioinformatics*, 18(9):1194–206, 2002.

P. Mueller and G. Rosner. A Bayesian population model with hierarchical mixture priors applied to blood count data. *J. Amer. Statist. Assoc.*, 92:1279–1292, 1997.

R. M. Neal. Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 9(2):249–265, 2000.

D. Ohlssen, L. Sharples, and D. Spiegelhalter. Flexible random-effects models using Bayesian semi-parametric models: applications to institutional comparisons. *Stat Med*, 26:2088–2112, 2007.

B. H. Patterson, C. M. Dayton, and B. I. Graubard. Latent class analysis of complex sample survey data: Application to dietary data. *J. Amer. Statist. Assoc.*, 97(459), September 2002.

R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2006. URL `http://www.R-project.org`.

S. Richardson and P. Green. On Bayesian analysis of mixtures with an unknown number of components. *J. Roy. Statist. Soc. B (with discussion)*, 59:731–792, 1997.

D. Spiegelhalter, A. Thomas, and N. Best. *WinBUGS version 1.4 user manual*. MRC Biostatistics Unit, Cambridge, 2003.

D. M. Titterington, A. F. M. Smith, and U. E. Makov. *Statistical analysis of finite mixture distributions*. Wiley series in probability and mathematical statistics. Probability and mathematical statistics. Wiley, Chichester, 1985.

K. Tucker. Commentary: Dietary patterns in transition can inform health risk, but detailed assessments are needed to guide recommendations. *Int J Epidemiol*, 36(3):610–1, 2007.

G. Tutz and H. Binder. Localized classification. *Statistics and Computer*, 15:155–166, 2005.

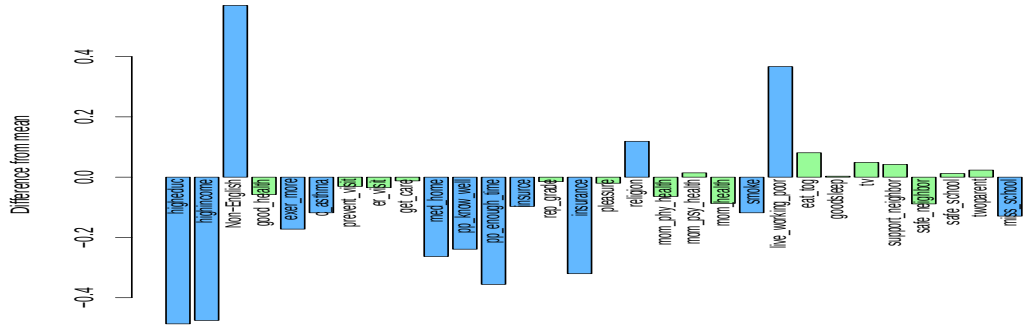
R. M. van Dam. New approaches to the study of dietary patterns. *Br J Nutr*, 93(5):573–4, 2005.

S. Walker, P. Damien, P. Laud, and A. Smith. Bayesian nonparametric inference for random distributions and related functions. *J. Roy. Statist. Soc. B (with discussion)*, 61:485–527, 1999.

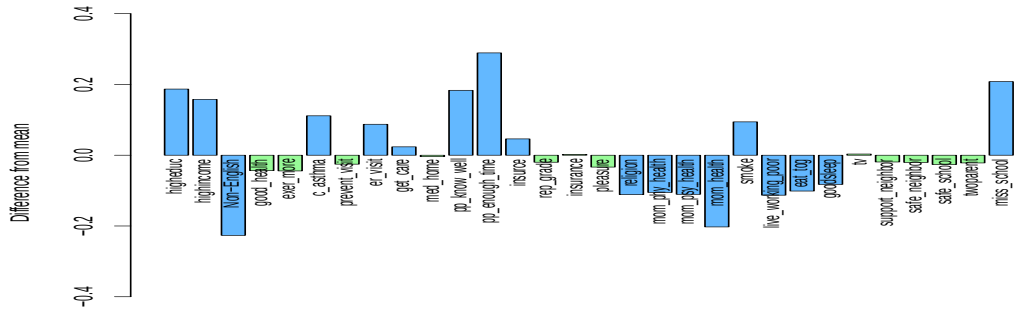
C. Wang. Invited commentary: beyond frequencies and coefficients–toward meaningful descriptions for life course epidemiology. *Am J Epidemiol*, 164(2):122–5; discussion 126–7, 2006.

M. West, P. Mueller, and M. Escobar. *Aspects of Uncertainty: A tribute to D. V. Lindley*, chapter Hierarchical priors and mixture models, with application in regression and density estimation, pages 363–386. Wiley, New York, 1994.

M. Yeh, K. McCabe, R. L. Hough, D. Dupuis, and A. Hazen. Racial/ethnic differences in parental endorsement of barriers to mental health services for youth. *Mental Health Services Research*, 5(2):65–77, 2003.

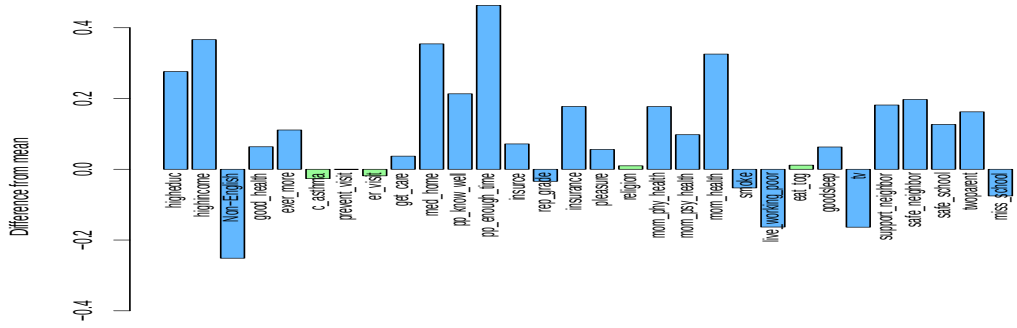


(a) $\bar{p}^* \approx -0.25$; $\Pr(\bar{p}^* < 0) \approx 0.92$

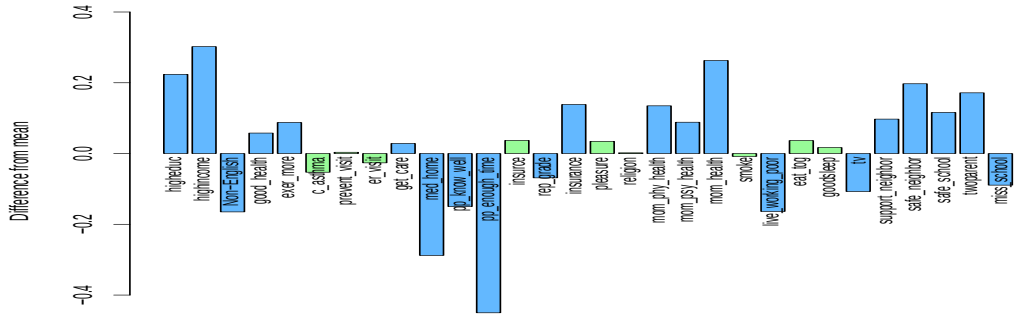


(b) $\bar{p}^* \approx 0.36$; $\Pr(\bar{p}^* > 0) \approx 1.00$

Fig. 1. Profiles values for $\bar{\phi}_k^p$ corresponding to subgroups with low and high risks for mental health problems. Bars in blue correspond to statistically significant values of $\bar{\phi}_k^p$, i.e. parameters for which the posterior probability of being greater (less) than zero is above 0.95. Individual variables are defined in Table 3 of the supplementary material.



(a) $\bar{p}^* \approx 0.01$; $\Pr(\bar{p}^* < 0) \approx 0.50$



(b) $\bar{p}^* \approx -0.00$; $\Pr(\bar{p}^* < 0) \approx 0.51$

Fig. 2. Profiles values for $\bar{\phi}_k^p$ corresponding to English speaking subgroups which do not have statistically risks for mental health problems. Bars in blue correspond to statistically significant values of $\bar{\phi}_k^p$, i.e. parameters for which the posterior probability of being greater (less) than zero is above 0.95. Individual variables are defined in Table 3 of the supplementary material.

Table 1. Main-effects model using forward selection. Note that stepwise procedure gave warnings and model was refit with all covariates significant at $p < 0.05$ level. Individual variables are defined Table 3 of the supplementary material.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.3070	0.0613	5.01	0.0000
young*	0.0153	0.0181	0.85	0.3982
nonwhite*	−0.0139	0.0215	−0.65	0.5171
male*	0.0447	0.0179	2.50	0.0128
mother*	0.0292	0.0231	1.26	0.2070
language	−0.0962	0.0274	−3.51	0.0005
er_visit	0.0522	0.0234	2.23	0.0261
med_home	−0.0961	0.0224	−4.29	0.0000
pp_enough_time	0.0982	0.0249	3.94	0.0001
rep_grade	−0.0612	0.0326	−1.88	0.0611
activity	−0.1205	0.0260	−4.63	0.0000
religion	0.0381	0.0214	1.78	0.0752
mom_psy_health	−0.1460	0.0376	−3.88	0.0001
smoke	0.0583	0.0222	2.62	0.0089
goodsleep	−0.0701	0.0333	−2.10	0.0357
safe_neighbor	−0.0626	0.0253	−2.47	0.0136

*Used as confounders for the logistic regression analysis in Section 4.1.

Table 2. Interaction model using forward selection starting with main-effects listed in Table 1 plus all two-way interaction terms. Note that stepwise procedure gave warnings and model was refit with all covariates significant at $p < 0.05$ level. Individual variables are defined Table 3 of the supplementary material.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.2172	0.0667	3.26	0.0012
young*	0.0208	0.0178	1.17	0.2425
nonwhite*	−0.0169	0.0210	−0.80	0.4223
male*	0.0393	0.0173	2.27	0.0238
mother*	0.0242	0.0226	1.07	0.2843
language	−0.0868	0.0272	−3.19	0.0015
er_visit	0.2144	0.0598	3.59	0.0004
med_home	−0.3322	0.0599	−5.55	0.0000
pp_enough_time	0.3308	0.0562	5.88	0.0000
activity	−0.0064	0.0407	−0.16	0.8743
religion	0.0414	0.0209	1.98	0.0483
mom_psy_health	−0.1454	0.0368	−3.95	0.0001
smoke	0.0564	0.0217	2.59	0.0097
goodsleep	−0.0832	0.0326	−2.56	0.0108
safe_neighbor	−0.0566	0.0247	−2.29	0.0221
activity:med_home	0.2761	0.0643	4.30	0.0000
activity:er_visit	−0.1846	0.0648	−2.85	0.0045
activity:pp_enough_time	−0.2835	0.0611	−4.64	0.0000

*Used as confounders for the logistic regression analysis in Section 4.1.