Biostatistics (2010), **11**, 3, *pp.* 484–498
doi:10.1093/biostatistics/kxq013
Advance Access publication on March 29, 2010

Bayesian profile regression with an application to the National survey of children's health

JOHN MOLITOR*, MICHAIL PAPATHOMAS

Department of Epidemiology and Biostatistics, School of Public Health, Imperial College, St Mary's Campus, Norfolk Place, London W2 1PG, UK
john.molitor@imperial.ac.uk

MICHAEL JERRETT

Division of Environmental Health Sciences, School of Public Health, University of California, Berkeley, CA 94720-7360, USA

SYLVIA RICHARDSON

Department of Epidemiology and Biostatistics, School of Public Health, Imperial College, St Mary's Campus, Norfolk Place, London W2 1PG, UK

SUMMARY

Standard regression analyses are often plagued with problems encountered when one tries to make inference going beyond main effects using data sets 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 interrelated 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 potentially to influence cluster membership. The method is demonstrated with an analysis of survey data obtained from the National Survey of Children's Health. The approach has been implemented using the standard Bayesian modeling software, WinBUGS, with code provided in the supplementary material available at *Biostatistics* online. Further, interpretation of partitions of the data is helped by a number of postprocessing tools that we have developed.

Keywords: Bayesian analysis; Clustering; Dirichlet process; MCMC; Profile Regression.

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 interrelated explanatory variables such as data arising

*To whom correspondence should be addressed.

from detailed questionnaires. The covariates in these data sets 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 soon becomes unwieldy and power to find any effects beyond simple 2-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 (van Dam, 2005; Tucker, 2007). 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, e.g. 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 *and others* (2002). Recent developments in LCA techniques to analyze correlated data can be found in DeSantis *and others* (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 (Richardson and Green, 1997; Diebolt and Robert, 1994) that takes into account the uncertainty associated with cluster assignments, that is, 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, e.g. Gilks *and others*, 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 uncertainty 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 article, we first describe the method with special emphasis paid to interpretation of model output. We then report the results of a simulation study 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 data set 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 submodel," which assigns individual profiles to clusters and a "disease submodel," which links clusters of profiles to an outcome of interest via a regression model. As is typical with Bayesian methods, both submodels will be fitted jointly using MCMC methods (Gilks *and others*, 1996), so, for example, allocation of individual profiles to clusters will depend on both the covariate data in the assignment submodel and the outcome information in the disease submodel. Both these submodels will be addressed in turn.

2.1 *Assignment submodel*

We first construct an allocation submodel 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, as described in Jain and Neal (2004) or Neal (2000). Our mixture model incorporates a Dirichlet process (DP) prior on the mixing distribution. The use of the DP in statistical modeling has been thoroughly examined in Walker *and others* (1999). A good overview of DP mixture models can be found in West *and others* (1994), while a biomedical example of their application can be found in Mueller and Rosner (1997). For further background information regarding mixture models with DP priors, see Neal (2000), Green and Richardson (2001), MacEachern and Muller (1998) and Escobar and West (1995).

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 cluster to which individual, i belongs. We restrict our approach to categorical covariates with M_p categories for the pth covariate. We denote with ψ_c the probability of assignment to the cth cluster, and let $\phi_c^p(x)$ denote the probability that the pth 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 article, we only analyze data sets with binary covariates and 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 *and others*, 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 $V_c \sim \text{Beta}(1, \alpha)$. Since we have little *a priori* information regarding the specification of α, we place a uniform prior on the interval $(0.3, 10)$. 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 *and others* (2007), so that potential computational traps in WinBUGS (Spiegelhalter *and others*, 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 the value of C large enough to give a good approximation but small enough 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 *and others* (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 article were generally in the range of $\alpha \in (0.5, 2.5)$. Thus, the upper bound chosen imposes little structure in practice.

2.2 *Disease submodel*

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 cluster to which individual i belongs. The cth 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 *and others* (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 that 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 that 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 (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 their 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 MCMC methods (Gilks *and others*, 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 *and others*, 2003), used to perform the MCMC parameter estimation, is provided in Section 2 of the supplementary material (available at *Biostatistics* online).

3. Examining clustering output

Our model implementation allows the number of groups to change between iterations 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 2 main areas of interest, namely (i) characterizing the partition (grouping) that is most supported by the data and (ii) assess uncertainty associated with subgroups of this best partition and compute risks associated with specified profiles of interest in a manner which exploits the MCMC output of the sampler. We discuss these issues in turn.

3.1 *Characterizing the typical partition*

For many applications, it is of interest to display 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) and 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 0 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, $\boldsymbol{z}^{\text{best}}$, 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 an alternative approach that provides good results 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 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, $\boldsymbol{S}$. However, we have found that the PAM approach and the approach of Dahl (2006) often produce very similar results. Further, the PAM approach requires a specification of at least 2 clusters when calculating the optimal clustering score, meaning that Dahl's approach may be preferable where the data have only a weak structure.

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 data set. For clarity, we demonstrate this for the partition $\boldsymbol{z}^{\text{best}}$ described above, but the concepts that we define apply to any given partition. The basic idea is to take the partition $\boldsymbol{z}^{\text{best}}$, representing the "best" clustering of the data, and to examine by postprocessing whether or not the model consistently clusters individuals in a manner similar to $\boldsymbol{z}^{\text{best}}$. 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 data set 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 are "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 reexamination 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 $\boldsymbol{z}^{\text{best}}$. 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 follows:

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

where n_k denotes the number of individuals in subgroup k of $\boldsymbol{z}^{\text{best}}$. Subgroup parameter values corresponding to covariate probabilities, $\bar{\phi}_k^p$, can be computed similarly as follows:

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

Note that instead of using means in (3.3) and (3.4), one could use medians if 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 0 or 1.

For interpretation purposes, we define new centered baseline risk parameters, $\bar{p}_k^*$, 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 postprocessing 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. Note that these probabilities are calculated based on θ_k parameters and not on parameters related to confounders. Similar summaries are derived for each $\bar{\phi}_k^{p*}$.

Regardless of the procedure used for choosing $\boldsymbol{z}^{\text{best}}$, we stress that the groups in this partition, as any partition, should be postprocessed through the output of the sampler in this way in order to assess uncertainty for group parameters. This postprocessing approach represents a compromise between an examination of interpretable "hard groupings," as exemplified by $\boldsymbol{z}^{\text{best}}$, 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 series of small simulation studies to illustrate the performance of the model both in the presence and absence of a strong signal, with results depicted in the supplementary material available at *Biostatistics* online. The results demonstrate that the model performs well when the signal is strong (Table 1 of the supplementary material available at *Biostatistics* online), and that the presence of an outcome is useful in differentiating between clusters with highly similar profile parameter values (Tables 2 and 3 of the supplementary material available at *Biostatistics* online). We also compared the Bayesian profiling method with standard LCA in regards to detection of a preset number of "hard" clusters. While recovery of the "correct" number of hard clusters represents only one aspect of the proposed method, results depicted in Table 4 of the supplementary material available at *Biostatistics* online reveal that our method is competitive with LCA in this regard.

4. DATA ANALYSIS—NSCH

The data analyzed in this section come from the NSCH, a US 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 data set consists of responses to a wide variety of health-related questions. In addition to the raw survey questions, an enhanced data set 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 2 or more related questions. As a way to demonstrate the utility of the approach described in this article, we analyzed profiles made up of indicators together with a few basic variables. (See Table 6

in the supplementary material available at *Biostatistics* online 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 US 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 1 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 data set to 34 variables (listed in Table 6 of the supplementary material available at *Biostatistics* online). Including all individuals living in California with a complete profile of these 34 variables, we obtain a sample size of $N = 642$.

The data were analyzed using the methods described in this paper and the provided WinBUGS code (Section 2 of the supplementary material available at *Biostatistics* online), along with additional postprocessing code written in R (R Development Core Team, 2006). For all real data analyses performed in this paper, the algorithm was run for 50 000 iterations with 10 000 iterations discarded for burn-in. Visual inspection of posterior time–series plots for the $\bar{\phi}$'s and $\bar{p}$'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 2 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 4 variables as confounders, 3 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 *and others*, 2007). Therefore, we trimmed the stepwise results by only keeping covariates with $p < 0.05$ and then refitted the model with the results listed in Table 1. We next formed a model consisting of all covariates and all 2-way interaction terms made up of these covariates. Again, a final model was obtained by running forward selection but, as previously, warnings were produced. We therefore refitted the model after eliminating all nonsignificant covariates and interactions, with the 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 regard to health access, the covariate "medical home," which is defined by the American Academy of Pediatrics as "accessible, continuous, comprehensive, family centered, coordinated, compassionate, and culturally effective" (The medical home, 2002), was highly significant. 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 one's 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 both be interpreted as

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* 6 *of the supplementary material available at Biostatistics online*

	Estimate	Standard 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 2-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* 6 *of the supplementary material available at Biostatistics online*

	Estimate	Standard 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.

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 in Section 4.1.

Profile regression. We analyzed the data using the profile approach described in this article, using the model corresponding to (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 postprocessing methods described in Section 3 revealed a "best" partition of 6 subgroups. Four of these subgroups are clearly "statistically significant" in that they are associated with very high posterior probabilities that centered values for $\bar{p}^*$ are away from zero. The other 2 subgroups are closer to the average risk, are contrasted in regard to health care access, yet lack consistency with regard to the statistical significance of the lifestyle and community-related covariates which appear to differentiate subgroups in regard to mental health risk.

The subgroup most strongly associated with low risk of mental health problems for the child ($\bar{p}^* = -0.054$ and $\Pr(\bar{p}^* < 0) = 0.996$) 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 toward mental health medical care that could potentially induce underreporting. This phenomenon has been described in the literature before; for example, Yeh *and others* (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 subgroup most strongly associated with high risk of mental health problems for the child ($\bar{p}^* = 0.174$ and $\Pr(\bar{p}^* > 0) \approx 1.00$) is depicted in Figure 1(b). This subgroup is mostly English-speaking and exhibits a coherent combination of behavioral and medical problems for the child (high values for "miss_school" and "c_asthma") and maternal health problems (low values for "mom_phy_health," "mom_psy_health," "mom_health") along with high levels of maternal smoking. In Figure 2, we display 2 English-speaking subgroups, both of which have risks that are less than average. Comparing Figure 1(b) with Figure 2(a) and (b), we see that these lower-risk 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 incidence 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 2 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.

Profile regression: comparing risks between prespecified profiles. The methodology proposed in this article allows for comparison of risks associated with prespecified profiles defined to address specific substantive questions. This can be accomplished by postprocessing the MCMC output to obtain a posterior distribution for the baseline risk for the prespecified profile in question. To achieve this, one simply calculates, at each iteration of the sampler, the probability that the prespecified profile, i', belongs to each of the K clusters and then samples the appropriate risk, θ_c, with those probabilities. Cluster membership

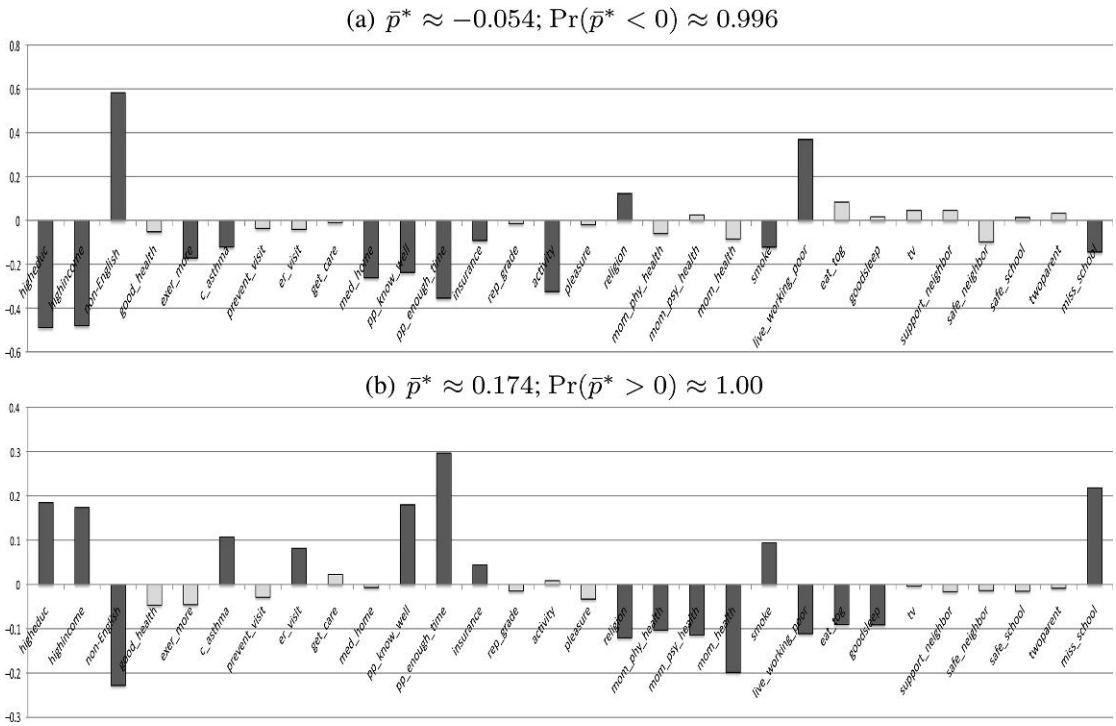


Fig. 1. Profile values for $\bar{\phi}_k^p$ corresponding to subgroups with highest and lowest risks for mental health problems. Bars in blue correspond to statistically significant values of $\bar{\phi}_k^p$, that is, parameters for which the posterior probability of being greater (less) than zero is above 0.95. Individual variables are defined in Table 6 of the supplementary material available at *Biostatistics* online.

probabilities can be easily calculated as follows:

$$\Pr(z_{i'} = c|\boldsymbol{x}, \psi_c, \boldsymbol{\phi}) \propto \Pr(z_{i'} = c) \Pr(\boldsymbol{x}|z_{i'} = c, \psi_c, \boldsymbol{\phi}) = \psi_c \prod_{p=1}^{P} \phi_p^c(x_p). \tag{4.5}$$

The conditional independence assumption of our assignment model (2.1) means that risk distributions for partially specified profiles can also be easily processed by including only relevant parameter values in calculating the probability that the profile is assigned to each cluster.

For example, in the NSCH data set one may wish to compare the baseline risks for mental health problems associated with the 2 profiles specified in Figures 3(a) and (b). These 2 profiles both represent children raised in homes associated with English-speaking, high-income, well-educated family members, but differ in regard to lifestyle variables such as high levels of television watching ("TV") and in regard to variables associated with the communities in which they live ("support_neighbor," "safe_neighbor," "safe_school"). Figure 3(c) shows the posterior distribution of the difference in risks associated with these 2 profiles ("generally unfavorable"' minus "generally favorable"). The bulk of this distribution is above zero, suggesting that children from prosperous families living in homes associated with less favorable environments are more likely to have mental health problems. Methodologically, it is important to note that the results depicted in Figure 3 were obtained via Bayes model averaging techniques, that is, they were calculated by averaging cluster allocations and risk values contained in the rich MCMC output produced

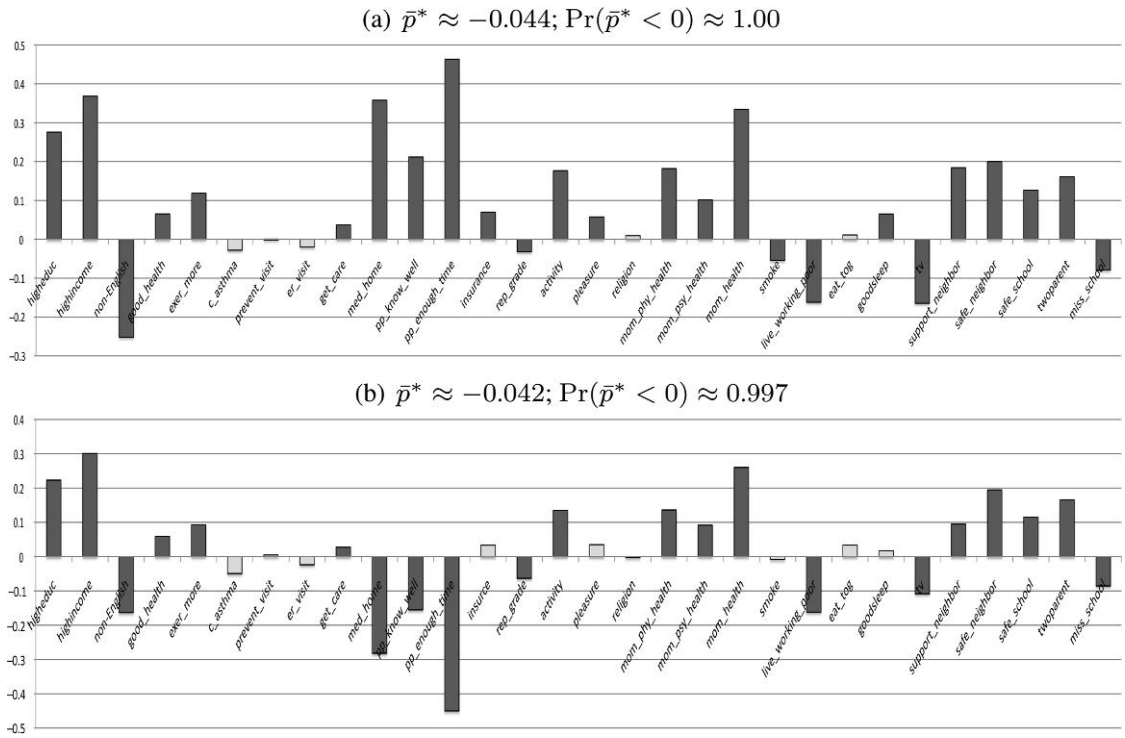


Fig. 2. Profile values for $\bar{\phi}_k^p$ corresponding to 2 English-speaking subgroups which have risks for mental health problems that are lower than average. Bars in blue correspond to statistically significant values of $\bar{\phi}_k^p$, that is, parameters for which the posterior probability of being greater (less) than zero is above 0.95. Individual variables are defined in Table 6 of the supplementary material available at *Biostatistics* online.

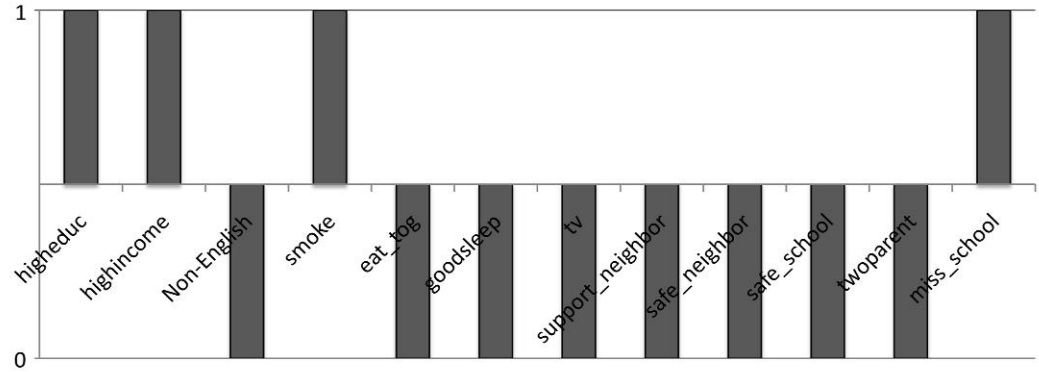
by the model. For this analysis, we did not base the analysis on "typical" clusters, as described in the previous sections, but rather averaged over the possible numbers of clusters and cluster assignments.

5. DISCUSSION

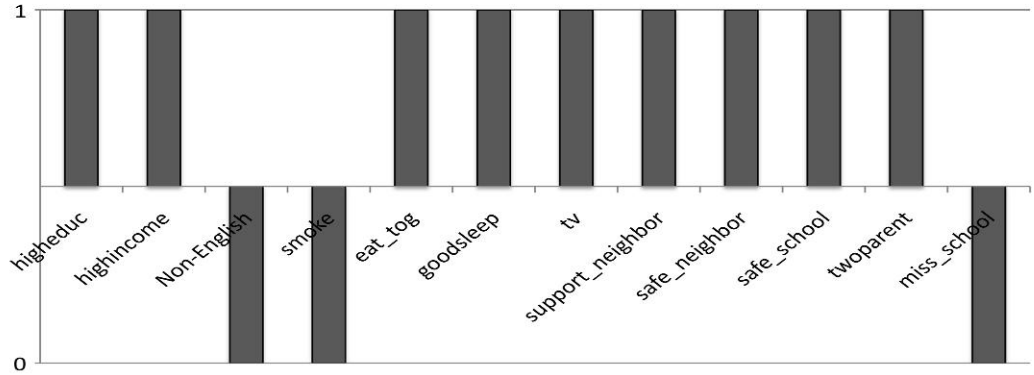
We have described a new analytical strategy that uses a covariate pattern, or profile, as the basic unit of inference, and examines associations between these profiles and an outcome of interest. Some of the ingredients of our approach are well established but, to our knowledge, have not all been put together in the manner described in this article to create a unified, easy-to-implement method for analyzing data with many interactive variables. Our method groups profiles into clusters, and the number of clusters is allowed to be random. Postprocessing 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 and inclusion of continuous covariates could be envisaged. 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

(a) Prosperous background / "generally unfavorable" lifestyle and community variables. Bars represent binary covariate values.



(b) Prosperous background / "generally favorable" lifestyle and community variables. Bars represent binary covariate values.



(c) Prosperous unhealthy community - Prosperous healthy community: $\Pr(\theta_{z_a} - \theta_{z_b} > 0) \approx 0.911$

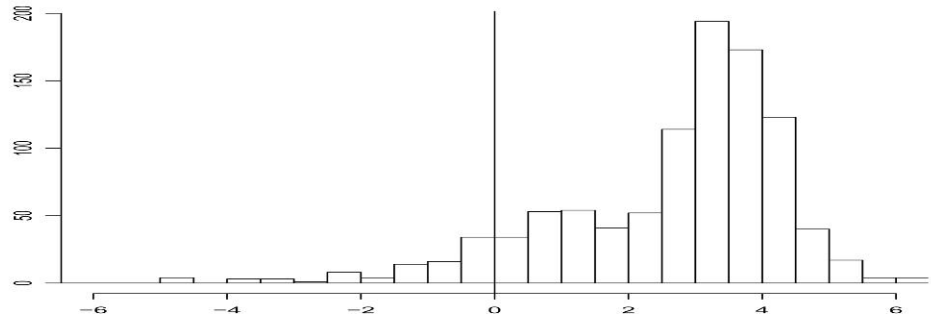


Fig. 3. Comparisons of risks associated with partially specified profiles.

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 postprocessing 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 numbers of variables with several categories to be

efficiently analyzed, as well as incorporating model extensions like ordinal covariate modeling using an underlying probit model. Note also that while we analyzed 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 *and others*, 2003). We further note that it might be tempting for users of the WinBUGS program to add an intercept to model (2.2). However, as an anonymous reviewer pointed out on a previous version of this article, the combination of an intercept term and the cluster random effects would render the model nonidentifiable.

Our model was formulated in a DP framework allowing for flexibility in the number of clusters used. However, our general profiling approach, including the postprocessing 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 truncated 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 *and others*, 2007). However, as the sampler progresses, clusters containing only 1 or 2 individuals are sometimes observed, which could lead to estimation problems for small sized data sets. 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 practice. 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 *and others*, 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 combinations of variables influencing the health outcome.

Variables within the profile that explains 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 traditional logistic regression. However, we note that logistic regression aims at estimation of main effects and interaction terms, while the profile approach described in this article is aimed at the examination of a combination of variables that structure the variability of the data. Since the 2 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.

SUPPLEMENTARY MATERIAL

Supplementary material is available at http://biostatistics.oxfordjournals.org.

ACKNOWLEDGMENTS

We would like to thank the Child and Adolescent Health Measurement Initiative (CAHMI) at the Oregon Health and Science University for being so helpful in providing us with the data. Finally, we would like to

acknowledge the many useful and insightful comments made by the editor, associate editor and reviewers on previous drafts of this paper. These comments have contributed to a much improved final manuscript. *Conflict of Interest:* None declared.

FUNDING

Medical Research Council (G0600609).

REFERENCES

AMERICAN ACADEMY OF PEDIATRICS (2002). Medical Home Initiatives for children with special needs project advisory committee, The Medical Home, *Pediatrics* **110**, 184–186.

DAHL, D. (2006). Model-based clustering for expression data via a Dirichlet process mixture model. In: Kim-Anh, D., Müller, P. and Vannucci, M. (editors), *Bayesian Inference for Gene Expression and Proteomics*. Cambridge: Cambridge University Press, pp. 210–216.

DESANTIS, S. M., HOUSEMAN, E. A., COULL, B. A., LOUIS, D. N., MOHAPATRA, G. AND BETENSKY, R. A. (2009). A latent class model with hidden markov dependence for array CGH data. *Biometrics* **65**, 1296–1305.

DESANTIS, S. M., HOUSEMAN, E. A., COULL, B. A., STEMMER-RACHAMIMOV, A. AND BETENSKY, R. A. (2008). A penalized latent class model for ordinal data. *Biostatistics* **9**, 249–262.

DIEBOLT, J. AND ROBERT, C. P. (1994). Estimation of finite mixture distributions through Bayesian sampling. *Journal of the Royal Statistical Society, Series B* **56**, 363–375.

ESCOBAR, M. AND WEST, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association* **90**, 577–588.

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

GELMAN, A., JAKULIN, A., PITTAU, M. AND SU, Y.-S. (2008). A weakly informative default prior distribution for logistic and other regression models. *Annals of Applied Statistics* **2**, 1360–1383.

GILKS, W., RICHARDSON, S. AND SPIEGELHALTER, D. (editors) (1996). *Markov Chain Monte Carlo in Practice*. London: Chapman & Hall.

GREEN, P. J. AND RICHARDSON, S. (2001). Modelling heterogeneity with and without the Dirichlet process. *Scandinavian Journal of Statistics* **28**, 355–375.

HARTIGAN, J. AND WONG, M. (1979). A k-means clustering algorithm. *Applied Statistics* **28**, 100–108.

ISHWARAN, H. AND JAMES, L. (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association* **96**, 161–173.

JAIN, S. AND NEAL, R. (2004). A split-merge Markov chain Monte carlo procedure for the Dirichlet process mixture model. *Journal of Computational and Graphical Statistics* **13**, 158–182.

KAUFMAN, L. AND ROUSSEEUW, P. J. (2005). *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley Series in Probability and Mathematical Statistics. Hoboken, NJ: Wiley-Interscience.

MACEACHERN, S. N. AND MULLER, P. (1998). Estimating mixture of dirichlet process models. *Journal of Computational and Graphical Statistics* **7**, 223–238.

MACLEHOSE, R. F., DUNSON, D. B., HERRING, A. H. AND HOPPIN, J. A. (2007). Bayesian methods for highly correlated exposure data. *Epidemiology* **18**, 199–207.

MEDVEDOVIC, M. AND SIVAGANESAN, S. (2002). Bayesian infinite mixture model based clustering of gene expression profiles. *Bioinformatics* **18**, 1194–1206.

MÜLLER, P. AND ROSNER, G. (1997). A Bayesian population model with hierarchical mixture priors applied to blood count data. *Journal of the American Statistical Association* **92**, 1279–1292.

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

OHLSSEN, D., SHARPLES, L. AND SPIEGELHALTER, D. (2007). Flexible random-effects models using Bayesian semi-parametric models: applications to institutional comparisons. *Statistics in Medicine* **26**, 2088–2112.

PATTERSON, B. H., DAYTON, C. M. AND GRAUBARD, B. I. (2002). Latent class analysis of complex sample survey data: application to dietary data. *Journal of the American Statistical Association* **97**, 721–728.

R DEVELOPMENT CORE TEAM (2006). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. http://www.R-project.org.

RICHARDSON, S. AND GREEN, P. J. (1997). On Bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society, Series B* **59**, 731–792.

SPIEGELHALTER, D., THOMAS, A. AND BEST, N. (2003). *WinBUGS User Manual*. Version 1.4. Cambridge: MRC Biostatistics Unit.

TUCKER, K. (2007). Commentary: dietary patterns in transition can inform health risk, but detailed assessments are needed to guide recommendations. *International Journal of Epidemiology* **36**, 610–611.

TUTZ, G. AND BINDER, H. (2005). Localized classification. *Statistics and Computer* **15**, 155–166.

VAN DAM, R. M. (2005). New approaches to the study of dietary patterns. *British Journal of Nutrition* **93**, 573–574.

WALKER, S., DAMIEN, P., LAUD, P. AND SMITH, A. (1999). Bayesian nonparametric inference for random distributions and related functions (with discussion). *Journal of the Royal Statistical Society, Series B* **61**, 485–527.

WANG, C. (2006). Invited commentary: beyond frequencies and coefficients—toward meaningful descriptions for life course epidemiology. *American Journal of Epidemiology* **164**, 122–125; discussion 126–127.

WEST, M., MUELLER, P. AND ESCOBAR, M. (1994). Hierarchical priors and mixture models, with application in regression and density estimation. In: Smith, A. F. M. and Freeman, P. R. (editors), *Aspects of Uncertainty: Attribute to D. V. Lindley*. New York: Wiley, pp. 363–386.

YEH, M., MCCABE, K., HOUGH, R. L., DUPUIS, D. AND HAZEN, A. (2003). Racial/ethnic differences in parental endorsement of barriers to mental health services for youth. *Mental Health Services Research* **5**, 65–77.

[*Received June 4, 2009; revised February 16, 2010; accepted for publication February 16, 2010*]