Non-parametric Bayes models for mixed scale longitudinal surveys

Tsuyoshi Kunihama,
*Kwansei Gakuin University, Nishinomiya, Japan*

Carolyn T. Halpern
*University of North Carolina at Chapel Hill, USA*

and Amy H. Herring
*Duke University, Durham, USA*

[Received February 2018. Revised February 2019]

**Summary.** Modelling and computation for multivariate longitudinal surveys have proven challenging, particularly when data are not all continuous and Gaussian but contain discrete measurements. In many social science surveys, study participants are selected via complex survey designs such as stratified random sampling, leading to discrepancies between the sample and population, which are further compounded by missing data and loss to follow-up. Survey weights are typically constructed to address these issues, but it is not clear how to include them in models. Motivated by data on sexual development, we propose a novel non-parametric approach for mixed scale longitudinal data in surveys. In the approach proposed, the mixed scale multivariate response is expressed through an underlying continuous variable with dynamic latent factors inducing time varying associations. Bias from the survey design is adjusted for in posterior computation relying on a Markov chain Monte Carlo algorithm. The approach is assessed in simulation studies and applied to the National Longitudinal Study of Adolescent to Adult Health.

**Keywords:** Mixed scale data; Mixture model; Multivariate longitudinal data; Non-parametric Bayes modelling; Sample survey

1. Introduction

The National Longitudinal Study of Adolescent to Adult Health (which is known as ‘Add Health’) has collected vast amounts of information about social, economic and health-related behaviours in adolescence over 20 years. One of the primary aims is to study trajectories of adolescent sexual development into adulthood. In particular, it is of interest to investigate associations between different indicators of sexual orientation, including adolescent sexual behaviours, attraction and identity, changes in the associations over the transition to adulthood and whether those changes may vary by demographic characteristics. Although Add Health data include all the sexual development indicators of interest, several characteristics of the data make it challenging to build a realistic statistical model.

First, the data consist of mixed scale variables, e.g. partner counts or nominal sexual orientation identity. It is not straightforward to model mixed scale multivariate data jointly, especially...
if they include both ordered variables and nominal variables. Second, individuals repeatedly an-
tered a set of questions over time, leading to subject-specific time dependence. Third, the Add
Health study drew its sample via a complex survey design with stratified sampling of schools
and oversampling of numerous groups. Hence, the data are not representative of the population
until sampling weights have been applied. Fourth, missing values are common, and we observe
both design-based and individual-specific missingness. For example, students were not asked as
detailed questions on their sexuality in adolescence as in young adulthood by design. There-
fore, we need to conduct statistical analysis of adolescent sexual development, taking fully into
account all of these challenging characteristics of the Add Health data.

There is a rich literature on modelling of mixed scale data. One approach is to apply gen-
eralized linear models for each outcome in which dependence between variables is induced
through shared latent factors (Sammel et al., 1997; Dunson, 2000; Moustaki and Knott, 2000;
Dunson and Herring, 2005). However, the robustness of the approaches based on generalized
linear mixed models can be weak because of the dual role of the random-effects structure in
controlling the dependence and shape of marginal distributions. Another approach is to use
underlying continuous variables, specified by a Gaussian model (Muthén, 1984) or a Dirichlet
process mixture model (Kottas, 2005; Canale and Dunson, 2011; Kim and Ratchford, 2013).
In this approach, discrete variables can be expressed by thresholding the latent continuous
variables. Avoiding specification of marginal distributions, Hoff (2007) proposed a semipara-
metric Gaussian copula model in which associations between mixed scale variables are induced
by correlations between the latent Gaussian variables. Murray et al. (2013) and Gruhl et al.
(2013) extended this approach by incorporating factor structures but these copula models can
incorporate only ordered variables. Also, Murray and Reiter (2016) proposed a non-parametric
Bayesian joint model for multiple imputation of missing values. McParland et al. (2014) de-
veloped a model-based clustering approach for mixed scale data that combines item response
theory models for ordered variables with factor analysis models for nominal variables.

Recently, several references have addressed the analysis of multivariate longitudinal data
(Bandyopadhyay et al., 2011; Verbeke et al., 2014). Dunson (2003) proposed dynamic latent
trait models in which auto-regressive Gaussian latent factors incorporate subject-specific time
dependence and generalized linear models describe mixed scale outcomes. It is routine to incor-
porate time effects in multivariate models through time varying covariates, such as polynomial
functions of age, with random coefficients (Gueorguieva and Sanacora, 2006; Fieuws and Ver-
beke, 2006; Luo and Wang, 2014; Baghfalakia et al., 2014). Liu et al. (2009) developed a joint
model for longitudinal binary and continuous variables, consisting of a correlated probit model
and a regression model with the Bartlett decomposition of a covariance matrix. Ghosh and Han-
son (2010) proposed a semiparametric approach with a mixture of Polya trees for random-effect
distributions. Das and Daniels (2014) developed a semiparametric model for bivariate sparse
longitudinal data, applying a matrix stick breaking process for a residual covariance matrix.
However, it is not clear how to handle survey bias and missing values in these approaches. There-
fore, none of these methods can capture the exact nature of longitudinal surveys in Add Health.

In the literature on survey data analysis, two major approaches are used: the design-based
approach, which treats outcomes as fixed quantities, and the model-based approach, which
models outcomes and effectively predicts values for the non-sampled subjects in a popula-
tion (Little, 2004; Levy and Lemeshow, 2008; Rao, 2011). Little (2004) and Gelman (2007)
showed difficulties with current methods in practice and highlighted the importance of inclu-
ding survey weights in model-based analyses. Zheng and Little (2003, 2005) proposed a non-parametric method that flexibly models the outcome given inclusion probability by using a
penalized spline, and Chen et al. (2010) extended the model for binary variables. Si et al. (2015)
proposed a non-parametric Bayesian model, which jointly models an outcome and survey weights based on a Gaussian process regression. However, these approaches were developed mainly for inference of the population mean of a univariate response, and extensions to mixed scale longitudinal data are not straightforward.

Further, Zangeneh et al. (2011) proposed estimation methods for finite population quantities, in which unobserved design information is imputed for non-sampled units by using non-parametric mixture models. Chen et al. (2012) developed Bayesian inference of finite population quantiles of continuous variables from unequal probability samples. Dong et al. (2014) applied a Bayesian bootstrap for generation of synthetic populations from complex survey data. Accounting for survey designs, Zhou (2014) and Zhou et al. (2016) extended the bootstrap approach for multiple imputation of missing data by generating draws from a posterior predictive distribution. Zangeneh and Little (2015) developed methods for inference on finite population totals with a probability proportional to size sample by using penalized spline models with heteroscedastic errors. Chen et al. (2017) widely reviewed the literature on inferences of finite population quantities such as means or totals in the population by using complex survey data.

There has been an increase in the use of mixed mode surveys, which collect data by communicating with the respondents by using multiple modes such as face-to-face interview, mail, phone and/or web (Biemer and Lyberg, 2003; Dillman and Christian, 2005; De Leeuw, 2005; De Leeuw et al., 2008; Dillman et al., 2009; Groves et al., 2009; Dillman, 2011). They are cost effective and may lead to higher response rates because individuals have different preferences for modes of reporting their information. Another advantage is that mixed mode designs may improve coverage of the population of interest. For example, those who have no Internet access are not represented in web surveys, but one may correct the bias if respondents are contacted by a letter and then choose the web or mail to report their answers. In addition, effects of the modes on the responses have been studied; for example, the modes with an interviewer may lead to socially desirable answers in sensitive topics (De Leeuw, 1992).

The aim of this paper is to investigate trajectories of associations between adolescent sexual development variables by using the Add Health longitudinal survey. However, none of the existing methods can fully capture the challenging characteristics of the data such as multivariate response with different scales, individual-specific time dependence, complex sampling designs and missing values. Therefore, there is a need to develop a new framework to address the task of our data analysis. We employ a model-based approach because various statistical models have been developed in the vast literatures, some of which partly satisfy the features of the Add Health survey, whereas the design-based literature is relatively limited. The novelty of this paper is to develop estimation methods for a scale-free measure of associations in mixed scale longitudinal surveys under complex sampling designs by combining different ideas from existing model-based approaches in the literature.

In the approach proposed, we first estimate parameters in the model and then compute the scale-free association between mixed scale variables in Goodman and Kruskal (1954, 1959, 1963, 1972) by generating random samples from posterior predictive distributions in target populations. More specifically, mixed scale variables are expressed through transformation of latent continuous variables, for which a Dirichlet process mixture of Gaussian factor models is developed. For unordered categorical variables, we employ the concept of utilities in multinomial probit models, in which the nominal outcome is a manifestation of underlying continuous utility variables. The subject-specific dynamic variability can be captured by time varying latent factors via Gaussian processes which can easily incorporate irregular time intervals for each respondent. In addition, we express effects of covariates and age on associations between the response variables based on the idea of the covariance regression by Hoff and Niu (2012) and
Fox and Dunson (2015). Inferences that are valid for the source population are obtained by adjusting the mixture weights in the Dirichlet process mixtures by using the survey weights. Because we build a joint model of the response variables and covariates, missing values can be easily imputed assuming that they are missing at random. For posterior computation, we develop an efficient Markov chain Monte Carlo (MCMC) algorithm in which we modify the Dirichlet process mixture with survey weights, taking into account uncertainty in the adjustment process.

2. National Longitudinal Study of Adolescent to Adult Health data

Add Health is a nationwide longitudinal study of adolescents in grades 7–12 in the USA in the 1994–1995 academic year (http://www.cpc.unc.edu/projects/addhealth). The cohort study has been following up respondents with four waves of in-home interviews conducted in 1994–1995, 1996, 2001–2002 and 2008–2009. It collects data on respondents’ social, behavioural, psychological and biological information, allowing researchers to study developmental trajectories across the life course of adolescence into adulthood.

The Add Health study design selects a stratified sample of 80 high schools and 52 middle schools from the USA with unequal probability of selection based on region, urbanicity, school type, ethnic mix and size. Then, students in each school are stratified by grade and sex. In addition, the study drew supplemental samples, oversampling groups of particular interest based on ethnicity, genetic relatedness to siblings, adoption status, disability and parental education. Taking into account non-response as well as the sample selection, survey weights have been constructed for population-representative inferences via weighting adjustments. The full study design has been described by Harris et al. (2009).

To understand adolescent sexual orientation development, our primary goal is studying associations between measures of sexual attraction, behaviour and identity in adolescence and trends in the transition to adulthood. We use three waves of surveys which were conducted when participants were in grades 7–12 in wave 1, young adults in wave 3 and adults in wave 4. Table 1 shows the list of variables in our data set. They are mixed scale, and the sexual orientation identity is missing by design in wave 1. For longitudinal studies with these three waves, survey

| Sexual development variables | Type            | Categories                                                                 |
|------------------------------|-----------------|----------------------------------------------------------------------------|
| Attraction to opposite sex   | Binary          | 0, no; 1, yes                                                             |
| Attraction to same sex       | Binary          | 0, no; 1, yes                                                             |
| Total number of opposite sex partners | Count       | 1, heterosexual; 2, mostly heterosexual; 3, mostly homosexual; 4, homosexual; 5, bisexual |
| Total number of same-sex partners | Count       | 1, heterosexual; 2, mostly heterosexual; 3, mostly homosexual; 4, homosexual; 5, bisexual |
| Sexual identity (missing at wave 1) | Nominal | 1, heterosexual; 2, mostly heterosexual; 3, mostly homosexual; 4, homosexual; 5, bisexual |

| Covariates | Type            | Categories                                          |
|------------|-----------------|-----------------------------------------------------|
| Gender     | Binary          | 0, male; 1, female                                  |
| Race       | Nominal         | 1, Caucasian; 2, Hispanic; 3, African American; 4, Asian; 5, other race |
| Parental education | Ordered categorical | 1, less than high school; 2, high school diploma or general educational development; 3, some college or vocational education; 4, college graduate or higher |
weights are constructed for \( n = 12288 \) respondents. The histograms of the sexual development variables, age and survey weights are in the on-line supplementary materials.

Because the survey weight indicates the number of people that the respondent represents (Horvitz and Thompson, 1952), one way of obtaining adjusted (weighted) data is to resample students in the sample according to their survey weights. In the resampled data, oversampled groups with small weights are less likely to appear whereas majorities with large weights are observed more often. We resampled \( n = 12288 \) respondents from the original data set with probability \( \tilde{w}_i = w_i / \sum_{j=1}^{n} w_j \) where \( w_i \) is the survey weight for the \( i \)th respondent. Fig. 1 reports the plot of the weighted and unweighted covariates, indicating the gap that is induced by the sampling design. For example, racial minorities are oversampled in the survey, leading to the relatively large decreased percentage of Caucasians in the raw data. Although the resampling method works for computing basic summary statistics, it may be inefficient in terms of discarding information of the unselected respondents in the original data set. Also, it is not clear how to incorporate uncertainty about the unsampled subjects in the population. We discuss an alternative adjustment method in Section 3.3, which takes fully into account the information of all respondents in our data set and reflects uncertainty about the unobserved subjects in the population.

3. Proposed modelling of mixed scale longitudinal surveys

The main contribution of this paper is to develop model-based methods for population level estimation of a scale-free measure of associations in mixed scale longitudinal surveys under
informative sampling. By borrowing ideas from existing statistical tools, the method addresses characteristics of the Add Health survey data: mixed scale response, respondent-specific time dependence, sample bias adjustment and abundant missing values. This section shows a new Bayesian framework which flexibly incorporates these features.

3.1. Modelling of mixed scale data with latent continuous variables

Let \( y = (y_1, \ldots, y_p)' \) be a multivariate response with mixed scale margins. We introduce a latent continuous variable \( y^* = (y^*_1, \ldots, y^*_p)' \in \mathbb{R}^p \) and express the response \( y \) by transforming the underlying variable \( y^* \). For example, a binary variable \( y_k \in \{0, 1\} \) can be induced by

\[
y_k = 1(y^*_k > 0),
\]

where ‘1(·)’ denotes an indicator function. A count variable \( y_k \in \{0, \ldots, \infty\} \) can be expressed as

\[
y_k = \sum_{r=0}^{\infty} \{r 1(a_r < y^*_k \leq a_{r+1})\},
\]

where \(-\infty = a_0 < a_1 < a_2 < \ldots < \infty\). For a nominal variable with \( d \) categories \( y_k \in \{1, \ldots, d\} \), we introduce \( d \) latent variables \( y^*_k, \ldots, y^*_k \) such that

\[
y_k = l \iff y^*_k = \max_{1 \leq l \leq d} y^*_k, \quad l = 1, \ldots, d.
\]

In this expression, \( y^*_k \) can be interpreted as the utility for the \( l \)th category as in multinomial probit models (McCulloch and Rossi, 1994; Imai and van Dyk, 2005; Burgette and Nordheim, 2012). Because the order of \( y^*_k, \ldots, y^*_k \) is unchanged with respect to adding any constant and multiplying any positive scale by all of them, without loss of generality we may assume that one of the utilities is 0.

3.2. Proposed non-parametric Bayes model

We apply the above framework to mixed scale longitudinal data. Let \( y_{ij} = (y_{ij1}, \ldots, y_{ijp})' \) be the response variable for respondent \( i \) at age \( t_{ij} \) for \( i = 1, \ldots, n \) and \( j = 1, \ldots, n_i \) with \( y_{ijk} \in \mathbb{Y}_k \) allowed to be any type of univariate random variable for \( k = 1, \ldots, p \). We introduce \( y^*_{ij} = (y^*_{ij1}, \ldots, y^*_{ijp})' \in \mathbb{R}^p \) as a latent continuous variable, which induces \( y_{ij} \) through a function \( g \) such that \( y_{ij} = g(y^*_{ij}) \) as in Section 3.1. Also, \( x_i = (x_{i1}, \ldots, x_{iL})' \in \mathcal{X} \) denotes static covariates for the \( i \)th respondent, from which target subpopulations may be constructed. If we use time varying covariates, we simply replace \( x_i \) by \( x_{ij} \) below.

We first consider a statistical model of \( y^*_{ij} \) given \( x_i \) and \( t_{ij} \). One approach may be a standard linear model where a covariance is constant for all respondents over time. However, because our interest is in investigating associations of the response varying with \( x_i \) and \( t_{ij} \), it is not appropriate. Also, the individual-specific variability needs to be incorporated in the model. Therefore, we consider the following factor model:

\[
y^*_{ij} = Bx_i + \Omega \mu_{tij} + \Lambda \eta_{ij} + \epsilon_{ij}, \quad \epsilon_{ij} \sim N(0, \Sigma), \quad \Sigma = \text{diag}(\sigma_1^2, \ldots, \sigma_p^2),
\]

where \( \mu_{tij} = (\mu_{tij1}, \ldots, \mu_{tijQ})' \) is a dynamic effect for age \( t_{ij} \), \( \eta_{ij} = (\eta_{ij1}, \ldots, \eta_{ijQ})' \) is a factor vector for student \( i \) at age \( t_{ij} \) with \( Q < p^* \), \( B \) is a \( p^* \times L \) matrix of regression coefficients, \( \Omega \) is a \( p^* \times Q \) coefficient matrix for the time effect and \( \Lambda \) is a \( p^* \times Q \) factor loading.

In equation (2), \( Bx_i \) and \( \Omega \mu_{tij} \) represent effects of the covariates and age on the mean of \( y^*_{ij} \),
and η_{ij} describes the respondent-specific variability. In addition, we assume that the variability depends on the covariates and age,

\[ \eta_{ij} = V_x \eta_i^* + \xi_{il} \tilde{\eta}_i, \quad \eta_i^* \sim N(0, 1), \quad \tilde{\eta}_i \sim N(0, I_Q), \tag{3} \]

where V is a Q × L coefficient matrix, ξ_i = \{ξ_{qil}\} is a Q × Q matrix for t = 1, . . . , T and η_i^* and \tilde{\eta}_i are independent random effects. Here, V_x η_i^* and ξ_{il} \tilde{\eta}_i represent variabilities that are related to the covariates and age respectively. This modelling comes from the idea of the covariance regression by Hoff and Niu (2012) and Fox and Dunson (2015). Equation (3) implies that η_{ij} \sim N(0, V_x x_i^t V' + ξ_{il} \tilde{\eta}_i) and, after integrating it out in expression (2), we obtain a model of \(y_{ij}^*\) where both the mean and the covariance vary with the covariates and age:

\[ y_{ij}^* \sim N(B x_i + \Omega \mu_{ij}, \Lambda V_x x_i^t V' + \Lambda \xi_{il} \xi_{il}' + \Sigma). \tag{4} \]

For the terms that are related to time effects, let \( \mu_q = \{\mu_{tq}, t \in T\} \) and \( \xi_{ql} = \{\xi_{tql}, t \in T\} \) be stochastic processes of the time effects with time space T, for which we apply Gaussian processes:

\[ \mu_q \sim \text{GP}(0, c_\mu), \quad c_\mu(\mu_{tq}, \mu_{t'q}) = \exp(-\kappa_\mu |t - t'|^2), \tag{5} \]

\[ \xi_{ql} \sim \text{GP}(0, c_\xi), \quad c_\xi(\xi_{tql}, \xi_{t'ql}) = \exp(-\kappa_\xi |t - t'|^2), \tag{6} \]

where \( \kappa_\mu > 0, \kappa_\xi > 0 \) and \( \mu_q \) and \( \xi_{ql} \) are mutually independent with respect to q and l.

Next, relying on the conditional density of the latent variable (2)–(4), we develop a flexible joint model of the response and covariates. The approach of the response–covariate joint modelling has been widely studied in the literature (Müller et al., 1996; Müller and Quintana, 2010; Hannah et al., 2011; Taddy and Kottas, 2012; DeYoreo and Kottas, 2015, 2018). In this framework, we can derive a marginal density \( f(y) \) and a conditional density \( f(y|x \in C) \) given \( C \subset \mathcal{X} \) with covariate space \( \mathcal{X} \) from the joint model \( f(y, x) \). Also, we can easily impute missing values from the conditional distribution given observed information assuming missingness at random.

We construct a joint density of \( y^*_i \) and \( x_i \) with \( y^*_i = \{y^*_i, \ j = 1, \ldots, n_i\} \), relying on Dirichlet process mixtures (Lo, 1984; West et al., 1994; Escobar and West, 1995):

\[ f(y^*_i, x_i) = \int f(y^*_i | x_i, \theta^v) f(x_i | \theta^s) dP(\theta), \quad P \sim \text{DP}(\alpha P_0), \tag{7} \]

\[ = \sum_{h=1}^\infty \pi_h f(y^*_i | x_i, \theta^v_h) \prod_{l=1}^L f(x_{il} | \theta^s_{hl}), \tag{8} \]

\[ \pi_h = v_h \prod_{b<h} (1 - v_b), \quad v_b \sim \text{beta}(1, \alpha), \tag{9} \]

where \( P \) denotes a prior distribution for \( \theta = (\theta^v, \theta^s) \), \( \alpha > 0 \) is a precision parameter and \( P_0 \) is a base probability measure for the Dirichlet process, and \( f(y^*_i | x_i, \theta^v_h) \) denotes the conditional density that is induced by expression (2)–(4). With respect to the density for the covariate \( f_i(x_{il} | \theta^s_{hl}) \) in equation (8), we assume a function depending on the type of the covariate, such as a multinomial distribution for a categorical variable. Equations (8) and (9) correspond to the stick breaking representation of Dirichlet process mixtures (Sethuraman, 1994; Mulliere and Tardella, 1998). The model proposed is the Dirichlet process mixture of Gaussian factor models. Finally, integrating out the latent variable over the space derived from the function in Section 3.1, we obtain the joint density of the response and covariates:

\[ f(y_i, x_i) = \int_{R(y_i)} f(y^*_i, x_i) dy^*_i, \tag{10} \]

where \( R(y_i) = \{y^*_i \in \mathbb{R}^{p^*+n_i} : y_{ij} = g(y^*_i), j = 1, \ldots, n_i\} \).
3.3. Population-representative inference on scale-free association

To adjust for the sample bias from complex survey designs, we incorporate survey weights in the proposed joint model. Let \( D \) be the finite population with \( N \) people, from which \( n \) subjects are sampled with inclusion probability \( \nu_i \) for subject \( i \in D \). The inverse of the inclusion probability \( \nu_i^{-1} \) can be interpreted as the number of people whom respondent \( i \) represents in the population (Horvitz and Thompson, 1952). Also, the survey weights can be defined as \( w_i \propto \nu_i^{-1} \). For the finite approximation of the Dirichlet process mixtures with \( H \) components, Kunihama et al. (2016) developed an adjustment method, which modifies the mixture weights \( \{ \pi_i \} \) in equation (8) by using the survey weights. In the posterior computation, to obtain distributions for a target population, we generate adjusted mixture weights \( \tilde{\pi} = (\tilde{\pi}_1, \ldots, \tilde{\pi}_H)' \) from

\[
\tilde{\pi} \sim \text{Dirichlet}\left(a_1 + \sum_{i : s_i = 1} \frac{w_i}{c}, \ldots, a_H + \sum_{i : s_i = H} \frac{w_i}{c}\right),
\]

where \( c = \sum_{i = 1}^n \frac{w_i}{N} \) and \( s_i \) is the cluster index variable for the \( i \)th respondent. With the prior distribution Dirichlet\((a_1, \ldots, a_H)\), the posterior distribution (11) consists of the prior sample size and the enlarged observations based on the survey weights. Compared with the posterior distribution Dirichlet\(\{a_1 + \sum_{i} I(s_i = 1), \ldots, a_H + \sum_{i} I(s_i = H)\}\) in a standard Bayesian mixture model, the weight of subject \( i \) is changed from 1 to \( w_i/c \approx \nu_i^{-1} \). Although method (11) was originally developed under a single-stage stratified sampling design, it can be applied to more broad cases, relying on the idea of post-stratification (Si et al., 2015). Strata can be constructed after the survey based on the unique values of the survey weights; then the population can be divided into mutually exclusive subpopulations, each of which has a different inclusion probability. The post-stratification assumes that all unique values of the survey weights appear in the sample, but the number of unique values in the sample may be less than that in the population (Si et al., 2015; Vandendijck et al., 2016). Kunihama et al. (2016) took into account the uncertainty about non-sampled units in the population by adjusting the prior sample size in the Dirichlet distribution (11). In our study, we follow the default setting in Kunihama et al. (2016) with prior sample size 1–2% of the population size \( N \).

Let \( \tilde{f}(y, x) \) be the adjusted distribution with the mixture weights \( \tilde{\pi} \) from distribution (11), which corresponds to the posterior predictive distribution for a target population. Hence, we can estimate characteristics of the target population by generating a simple random sample from the distribution \( \tilde{f} \). As a scale-free measure of dependence between two variables, we employ Goodman and Kruskal’s gamma (Goodman and Kruskal, 1954, 1959, 1963, 1972):

\[
\gamma = \frac{N_c - N_d}{N_c + N_d},
\]

where \( N_c \) is the number of concordant pairs, and \( N_d \) is the number of discordant pairs. A concordant pair can be defined as a pair of variables \( (X_1, Y_1) \) and \( (X_2, Y_2) \) such that \( \text{sgn}(X_2 - X_1) = \text{sgn}(Y_2 - Y_1) \). In contrast, a discordant pair means \( \text{sgn}(X_2 - X_1) = -\text{sgn}(Y_2 - Y_1) \). The measure \( \gamma \) ranges from \(-1\) (100% negative association) to 1 (100% positive association), and 0 indicates the absence of association. Although other popular rank correlations such as Spearman’s \( \rho \) need an adjustment for the ties, which are neither concordant nor discordant pairs, Goodman and Kruskal’s gamma is resistant to them.

To estimate associations for the subpopulation that is defined by \( C \subset \mathcal{X} \), we generate \( \tilde{y}_r \) from \( \tilde{f}(y|x \in C) = \tilde{f}(y, x \in C)/\tilde{f}(x \in C) \) with \( r = 1, \ldots, R \) where \( \tilde{y}_r = \{\tilde{y}_{rt}, t \in T\} \) and \( \tilde{y}_{rt} = (\tilde{y}_{rt1}, \ldots, \tilde{y}_{rtp})' \) is a mixed scale response for time \( t \). Then, we compute the concordant and discordant pairs for \( \tilde{y}_{rtj} \) and \( \tilde{y}_{rtj'} \) with \( j \neq j' \) and calculate Goodman and Kruskal’s gamma.
through equation (12) for each time point. For the overall population, we can estimate the
gamma similarly by setting $C = \mathcal{X}$.

We develop an efficient MCMC algorithm for the model proposed, and the details of the
posterior computation are in Appendix A. Rue et al. (2009) developed Bayesian computation
using integrated nested Laplace approximations for a popular class of regression models based
on latent Gaussian variables. The methods estimate parameters relying on approximated pos-
terior distributions whereas the proposed MCMC algorithm generates samples from the exact
posterior distributions. Lumley and Scott (2013) proposed design-based rank tests with com-
plex sampling, which compare distributions of a random variable between two different samples
such as a targeted sample and a large national survey. However, the methods do not measure
associations between two different random variables with different scales; therefore it is not
straightforward to extend the framework for our analysis of the Add Health data.

4. Simulation study

We assess the performance of the proposed method with simulation data by comparing with
existing approaches. We assume that the response $y_{it} = (y_{it1}, y_{it2}, y_{it3}, y_{it4})'$ consists of continu-
ous, binary, count and three-category nominal variables, inducing $p^* = 6$ latent variables, and
the covariate $x_i$ is a binary variable. All respondents participate in the survey three times out of
$T = 9$ time points, randomly taking $t_{ij} \in \{3j - 2, 3j - 1, 3j\}$ with equal probabilities for $j = 1, 2, 3$.
With respect to the groups with $x_i = 0$ and $x_i = 1$, we estimate the gamma measure (12) for all
pairs of variables in the mixed scale response except those of categories in the nominal variable
because the value is analytically equal to $-1$. For the priors in the method proposed, we use
$\alpha \sim \text{gamma}(0.25, 0.25)$ and $\sigma^2_{kk} \sim \text{IG}(2, \tilde{s}_k/200)$ where $\tilde{s}_k$ is the sample variance of the $k$th re-
spone for the ordered variables and is 1 for the nominal variable. For the count variable, we use
non-negative integers as cut points. Also, we set $H = 60$ and put 25 grids on the $(0, 1]$ interval
for $\kappa_\mu$ and $\kappa_\xi$, and the prior sample size is 1% of the population size for the bias adjustment
in Section 3.3. As for the number of factors, we set $Q = 4$. In general, it is not straightforward
to decide a best value analytically, and we need to set $Q$, taking a balance of parsimony and
flexibility. The model proposed applies a shrinkage-type prior with heavy tails, so effects of
redundant elements will be reduced in the estimate. Therefore, we may choose a not-too-small
$Q$, considering the computation time, and then check robustness of the estimation result.

We assume that a finite population with $N = 1200000$ consists of 15 districts with
$N_m = 10000m$ students for the $m$th district with $m = 1, \ldots, 15$. Each district has majority and minority
groups where the ratio of the minority randomly takes a value in $\{0.05, 0.1, 0.15, 0.2 \}$. We
consider a mutistage sampling design that four districts are firstly selected at random; then 500
students are randomly sampled from each of the majority and minority in each of the selected
districts, leading to $n = 4000$ students in the sample. To adjust for the sample bias, survey weights
are constructed by $w_i = N^* / \sum_{j \in S} N_j^*$ and $w_i N_m^*/500$ for respectively the majority and minority
student $i$ in the $m$th district in the sample where $N^*$ and $N_m^*$ respectively represent the numbers of the
majority and minority in the population and the $m$th district, and $S$ is the set of the selected
districts. For this case, the post-stratification for the proposed method is discussed in the on-line
supplementary materials. For the generation of $y_{it}$, we first introduce the latent variable $y_{it}^* \in \mathbb{R}^6$
for which we assume

$$y_{it}^* = \mu + F_{zt} \eta_i + \epsilon_{it}, \quad \eta_i \sim N(0, 1), \quad \epsilon_{it} \sim N(0, \Sigma),$$

where $z_t$ is the standardized logarithm of time $t$ obtained by subtracting the average and dividing
by the standard deviation, $\Sigma = \text{diag}(1, 1, 4, 1, 1, 1)$ and
By integrating out the random effect $\eta_i$, the distribution can be expressed as $y_{it}^* \sim N(\mu, z_t^2 FF' + \Sigma)$. For the majority and minority, the covariance corresponds to

$$
\begin{pmatrix}
    z_t^2 + 1 & 0 & 0 & 0 & 0 \\
    0 & 1 & 0 & 0 & 0 \\
    z_t^2 & 0 & z_t^2 + 4 & 0 & 0 \\
    0 & 0 & 0 & 1 & 0 \\
    0 & 0 & 0 & 0 & 1
\end{pmatrix},
$$

$$
\begin{pmatrix}
    1 & 0 & 0 & 0 & 0 \\
    0 & 0 & 0 & 4 & 0 \\
    0 & z_t^2 & 0 & z_t^2 + 1 & 0 \\
    0 & 0 & 0 & 0 & 1 \\
    0 & 0 & 0 & 0 & 0
\end{pmatrix},
$$

where non-zero off-diagonal elements are for the pair of continuous and count variables for the majority and that of binary and nominal (first category) variables for the minority. Then, we generate $y_{it}$ by transforming the latent variables as follows:

$$
y_{it1} = y_{it1}^*,
$$

$$
y_{it2} = 1(y_{it2}^* > 0),
$$

$$
y_{it3} = \sum_{r=0}^{\infty} \{ r1(a_r < y_{it3}^* \leq a_{r+1}) \},
$$

$$
y_{it4} = \sum_{k=4}^{6} [(k - 3)1(y_{itk}^* = \max\{y_{it4}^*, y_{it5}^*, y_{it6}^*\})],
$$

where $a_0 = -\infty$ and $a_{r+1} = r$ for $r \geq 0$. For the binary covariate, we assume that $P(x_i = 1 | \text{majority}) = 0.7$ and $P(x_i = 1 | \text{minority}) = 0.3$. The associations in $y_{it}$ given $x_i$ depend on all parameters and the functions (14)–(15); therefore it is not straightforward to compute true values of the gamma measure analytically. Hence, we approximate them by generating 10000 samples from the true data-generating function.

As for the competitors, there is no existing method for the estimation of associations in the mixed scale responses that fully incorporates the features of the Add Health survey. In this simulation, we consider a covariance regression with a polynomial function of time (Hoff and Niu, 2012) and a dynamic latent trait model (Dunson, 2003). Details of the model specification and prior information are in the on-line supplementary materials. Also, we applied the proposed model without the bias adjustment as well. These competitors can be applied to the analysis of mixed scale longitudinal data but do not take into account the sampling designs. For all methods, we generated 10000 samples after the initial 5000 burn-in, and every 10th sample was saved. To compute Goodman and Kruskal's gamma, we generated 4000 samples from the posterior predictive distribution. We observed that the sample paths were stable and the sample auto-correlations dropped smoothly.

To investigate the repeated sampling behaviour of the method proposed, we repeated the simulation 100 times and Fig. 2 reports boxplots of posterior means of the logarithm of mean absolute errors, mean coverage rates and the mean width of 95\% credible intervals. In both $x = 0$ and $x = 1$, the method constantly produces small errors over time compared with the competitors, and the difference between ‘new’ (with adjustment) and ‘NA’ (without adjustment) indicates effects of the bias adjustment (11). Figs 2(c)–2(f) show that the 95\% intervals by the proposed method tend to be wider but cover true values with high rates whereas the competitors
Fig. 2. Boxplots of (a), (b) log(mean absolute errors), (c), (d) mean coverage rates and (e), (f) mean widths of 95% credible intervals by using 100 simulations for the proposed model with (new) or without (NA) adjustment, covariance regression (CR) and dynamic latent trait model (DLT) with (a), (c), (e) $x = 0$ and (b), (d), (f) $x = 1$: the x-axis shows time points and the average.
show relatively low coverage rates, indicating that the method more sufficiently incorporates uncertainty in the estimation. Then, to study how well the method captures trajectories of associations, we show one simulation result in Fig. 3 as an illustrative example using the two associations that are related to the dependent structure (13). Fig. 3 indicates that the method captures the shape of the true associations well whereas the competitors miss the trajectories
mainly because they ignore the sampling designs. We also computed the log(mean absolute errors) by using the proposed method with $Q = 2, 3, 5$ for checking the robustness of the result (see the on-line supplementary materials). The model with $Q = 2$ shows slightly smaller errors, but we do not observe a big difference in these results.

In addition, we consider another scenario with a non-informative sampling design in which the design does not cause a gap between the sample and the population. Data are collected with the above sampling design, but we assume that the distribution of the response for the minority group is equal to that for the majority group. As a result, oversampling of the minority does not lead to a sample bias and the adjustment is not necessary to estimate characteristics of the target population. The estimation results are in the on-line supplementary materials, showing that the method proposed works as well as the competitors though it includes the redundant adjustment with the survey weights.

5. Analysis of adolescent sexual development data

This section applies the proposed method to the adolescent sexual development data. For the function $f_1(x)\mid \theta^x$ for the covariates in expression (8), we assume multinomial distributions using Dirichlet priors with all concentration parameters 1. For the count variables, because we observe high right skewness in these data, we use log-cut-points instead of non-negative integers, which enables the Dirichlet process mixtures to approximate such distributions efficiently. Also, as the prior of the count variables, we assume $\sigma_{hk}^2 \sim IG(2, \tilde{s}_k/200)$ where $\tilde{s}_k$ is the sample variance of $\log(y_{ijk} + 0.5)$. For the other parameters, we use the same prior settings as in Section 4. In the sampling bias adjustment, we set the hyperparameter in the Dirichlet prior such that the prior sample size is equal to 1% of the population size $N = 14677347$. We estimated trajectories of associations between sexual development variables for the overall population and within subpopulations with respect to gender, race, parents’ education and interactions of gender and race. We generate 10000 samples after the initial 5000 samples have been discarded as a burn-in period, and every 10th sample is saved. At each MCMC iteration, we computed Goodman and Kruskal’s gamma by generating 2500 subjects from the posterior predictive distribution. We observed that the sample paths were stable, and the sample auto-correlations dropped smoothly.

We study the association between attraction, behaviour and sexual orientation identity, the last of which is included starting at age 18 years, because this variable was not included in wave 1. Also, we omitted the result for age 11 and 34 years from the display below because of the small number of observations, which is 5 and 2 respectively. We estimated the gamma statistic in the overall population for opposite sex attraction $A_o$, same-sex attraction $A_s$, opposite sex partner count $#o$, same-sex partner count $#s$ and sexual orientation identities of heterosexual $H_e$, mostly heterosexual $MMe$, mostly homosexual $MHo$, homosexual $Ho$ and bisexual $Bi$. We observed various patterns of trajectories of the associations (the figure is in the on-line supplementary materials). Note that whereas there is no strong association between the attraction variables in early adolescence, with age a strongly negative association emerges and continues into adulthood. As we might expect, same-sex attraction and same-sex partner count have a strong positive association around 0.9, and, although the association between opposite sex attraction and partner count is positive, it is more modest, with estimated gamma values around 0.5. Whereas the association is essentially null in the early teens, with age a strong negative association between opposite sex attraction and same-sex partner count becomes apparent. However, the associations between the sexual orientation identity variables (not measured in the early to mid-teens) and partner counts tend to be relatively constant over time.
Whereas we saw few differences in the associations, or their changes over time, by race, we observed several strong differences by gender. Figs 4 and 5 report the comparison of the posterior means and 95% credible intervals for several pairs of sexual development variables among males and females. For example, whereas the association between counts of same- and opposite sex partners is null or positive in the early teens, this association remains significantly positive among females into adulthood but becomes negative among males by the late teens. We observe a similar trend in the gender-specific associations between same-sex attraction and opposite sex
Fig. 5. Comparison of posterior means and 95% credible intervals of associations for males (---, ■) and females (— — —, □, △): (a) #o–MHo; (b) #o–Ho; (c) #o–Bi; (d) #s–Bi; (e) #s–Ho; (f) #s–Ho; (g) #s–MHe; (h) #s–He

partner counts. This pattern is consistent with other findings that lesbian and bisexual-identified females are more likely than heterosexuals to experience both early vaginal intercourse and to engage in sexual risk behaviours such as more sexual partners and inconsistent contraception (Garofalo et al., 1998; Case et al., 2004; Charlton et al., 2011; McCabe et al., 2011; Herrick et al., 2013; Riskind et al., 2014; McCauley et al., 2014, 2015). We see opposite trends by gender between heterosexual sexual orientation identity and opposite sex partner counts, with this identity being associated with greater counts among men and lower counts among women from the late teens through early adulthood. In contrast, a mostly heterosexual identity has a null
association with opposite sex partner counts for men, but a positive association among women. Among both men and women, homosexual identity is associated with lower opposite sex partner counts. Same-sex partner counts are strongly associated with all identities except heterosexual among both men and women; the negative association between heterosexual identity and same-sex partner count is stronger among men than among women.

On the basis of the developmental literature, we also examined whether changes in associations over time varied by gender within race but we did not find strong evidence of this interaction in the association parameters. Similarly, we did not identify strong trends according to parental education (the data are not shown).

For comparison, we also applied the covariance regression model (Hoff and Niu, 2012) and obtained similar results of the differences by gender but observed some discrepancies; for example associations between the attraction variables are constantly negative in the competitor whereas there is little association in early adolescence in the method proposed. This is probably because the competitor does not capture all features of the Add Health survey, such as the sample bias. In addition, we investigated the effect of the number of factors on the estimation result with $Q = 2, 3, 5$ and observed similar patterns of trends of associations within biological sex. These results are in the on-line supplementary materials.

6. Discussion

Using data from a longitudinal, nationally representative sample of adolescents followed into adulthood, we could study how the associations between three components of sexual orientation, attraction, behaviour and sexual orientation identity, evolve with age. To study changes with time in the associations between these variables, measured longitudinally on a variety of categorical scales, we developed a flexible dynamic latent factor model allowing time varying associations between the multivariate responses. This model also allowed the associations to depend on covariates of interest, such as gender. Population-representative inferences are obtained via an MCMC algorithm that is used for posterior computation. Previous related work either did not allow longitudinal data (Fox and Dunson, 2015), required variables to be ordinal (Hoff and Niu, 2012), did not allow dependence of the association on covariates (Fox and Dunson, 2015) or did not incorporate survey weights.

With respect to the summary measure of associations between diverse types of variables, one alternative may be a covariance matrix in the Gaussian copula method by Hoff (2007). Although it relies on the assumption that data consist of ordered variables, the model can flexibly describe associations of variables by using a rank likelihood without any assumption about the marginal distributions. Hence, it may be an interesting future direction to extend the framework for incorporation of nominal variables.

Acknowledgements

This work was supported by Japan Society for the Promotion of Science KAKENHI grant JP26245028. The work of Halpern was supported by grant R01HD57046, awarded to Dr Halpern of the Carolina Population Center at the University of North Carolina at Chapel Hill by the Eunice Kennedy Shriver National Institute of Child Health and Human Development. The results were generated mainly by using Ox (Doornik, 2007), and the programming code for the method proposed is available from

https://rss.onlinelibrary.wiley.com/hub/journal/14679876/series-c-datasets.
References

Baghfalakia, T., Ganjalia, M. and Berridge, D. (2014) Joint modeling of multivariate longitudinal mixed measurements and time to event data using a Bayesian approach. *J. Appl. Statist.*, 41, 1934–1955.

Bandyopadhyay, S., Ganguli, B. and Chatterjee, A. (2011) A review of multivariate longitudinal data analysis. *Statist. Meth. Med. Res.*, 20, 299–330.

Biemer, P. P. and Lyberg, L. E. (2003) *Introduction to Survey Quality*. New York: Wiley.

Burgert, L. F. and Nordheim, E. V. (2012) The trace restriction: an alternative identification strategy for the Bayesian multinomial probit model. *J. Bus. Econ. Statist.*, 30, 404–410.

Canale, A. and Dunson, D. B. (2011) Bayesian kernel mixtures for counts. *J. Am. Statist. Ass.*, 106, 1528–1539.

Case, P., Austin, S. B., Hunter, D. J., Manson, J. E., Malspeis, S., Willett, W. C. and Spiegelman, D. (2004) Sexual orientation, health risk factors, and physical functioning in the Nurses' Health Study II. *J. Womns Hlth*, 13, 1033–1047.

Charlton, B. M., Corliss, H. L., Missner, S. A., Frazier, A. L., Rosario, M., Kahn, J. A. and Austin, S. B. (2011) Reproductive health screening disparities and sexual orientation in a cohort study of U.S. adolescent and young adult females. *J. Adolesc. Hlth*, 49, 505–510.

Chen, Q., Elliott, M. R., Haziza, D., Yang, Y., Ghosh, M., Little, R. J. A., Sedransk, J. and Thompson, M. (2017) Approaches to improving survey-weighted estimates. *Statist. Sci.*, 32, 227–248.

Chen, Q., Elliott, M. R. and Little, R. J. A. (2010) Bayesian penalized spline model-based inference for finite population proportion in unequal probability sampling. *Surv. Methodol.*, 36, 23–34.

Chen, Q., Elliott, M. R. and Little, R. J. A. (2012) Bayesian inference for finite population quantiles from unequal probability samples. *Surv. Methodol.*, 38, 203–214.

Das, K. and Daniels, M. J. (2014) A semiparametric approach to simultaneous covariance estimation for bivariate sparse longitudinal data. *Biometrics*, 70, 33–43.

De Leeuw, E. D. (1992) *Data Quality in Mail, Telephone and Face-to-face Surveys*. Amsterdam: TT-Publishers.

De Leeuw, E. D. (2005) To mix or not to mix data collection modes in surveys. *J Off. Statist.*, 21, 233–255.

De Leeuw, E. D., Hox, J. J. and Dillman, D. A. (2008) Mixed-mode surveys: when and why. In *International Handbook of Survey Methodology* (eds E. D. de Leeuw, J. Hox and D. Dillman), pp. 299–316. New York: Routledge.

DeYoreo, M. and Kottas, A. (2015) A fully nonparametric modeling approach to binary regression. *Biometrics*, 10, 821–847.

DeYoreo, M. and Kottas, A. (2018) Bayesian nonparametric modeling for multivariate ordinal regression. *J. Computat. Graph. Statist.*, 27, 71–84.

Dillman, D. A. (2011) *Mail and Internet Surveys: the Tailored Design Method—2007 Update with New Internet, Visual, and Mixed-mode Guide*. New York: Wiley.

Dillman, D. A. and Christian, L. M. (2005) Survey mode as a source of instability in responses across surveys. *Fild Meth.*, 17, 30–52.

Dillman, D. A., Phelps, G., Tortora, R., Swift, K., Kohrell, J., Berck, J. and Messer, B. L. (2009) Response rate and measurement differences in mixed-mode surveys using mail, telephone, interactive voice response (IVR) and the internet. *Socl Sci. Res.*, 38, 1–18.

Dong, Q., Elliott, M. R. and Raghunathan, T. E. (2014) A nonparametric method to generate synthetic populations to adjust for complex sampling design features. *Surv. Methodol.*, 40, 29–46.

Doornik, J. (2007) *Object-oriented Matrix Programming using Ox*, 3rd edn. London: Timberlake.

Dunson, D. B. (2000) Bayesian latent variable models for clustered mixed outcomes. *J. R. Statist. Soc. B*, 62, 355–366.

Dunson, D. B. (2003) Dynamic latent trait models for multidimensional longitudinal data. *J. Am. Statist. Ass.*, 98, 555–563.

Dunson, D. B. and Herring, A. H. (2005) Bayesian latent variable models for mixed discrete outcomes. *Biostatistics*, 1, 11–25.

Escobar, M. D. and West, M. (1995) Bayesian density estimation and inference using mixtures. *J. Am. Statist. Ass.*, 90, 577–588.

Fieuws, S. and Verbeke, G. (2006) Pairwise fitting of mixed models for the joint modeling of multivariate longitudinal profiles. *Biometrics*, 62, 424–431.

Fox, E. B. and Dunson, D. B. (2015) Bayesian nonparametric covariance regression. *J. Mach. Learn. Res.*, 16, 2501–2542.

Garofalo, R., Wolf, R. C., Kessel, S., Palfrey, J. and DuRant, R. H. (1998) The association between health risk behaviors and sexual orientation among a school-based sample of adolescents. *Pediatrics*, 101, 895–902.

Gelman, A. (2007) Struggles with survey weighting and regression modeling. *Statist. Sci.*, 22, 153–164.

Ghosh, P. and Hanson, T. (2010) A semiparametric Bayesian approach to multivariate longitudinal data. *Aust. New. Zeal. J. Statist.*, 52, 275–288.

Goodman, L. A. and Kruskal, W. H. (1954) Measures of association for cross classifications. *J. Am. Statist. Ass.*, 49, 732–764.
Goodman, L. A. and Kruskal, W. H. (1959) Measures of association for cross classifications, ii: Further discussion and references. J. Am. Statist. Ass., 54, 123–163.

Goodman, L. A. and Kruskal, W. H. (1963) Measures of association for cross classifications, iii: Approximate sampling theory. J. Am. Statist. Ass., 58, 310–364.

Goodman, L. A. and Kruskal, W. H. (1972) Measures of association for cross classifications, iv: Simplification of asymptotic variances. J. Am. Statist. Ass., 67, 415–421.

Groves, R. M., Fowler, F. J., Couper, M. P., Lepkowski, J. M., Singer, E. and Tourangeau, R. (2009) Survey Methodology. New York: Wiley.

Gruhl, J., Erosheva, E. A. and Crane, P. K. (2013) A semiparametric approach to mixed outcome latent variable models: estimating the association between cognition and regional brain volumes. Ann. Appl. Statist., 7, 2361–2383.

Gueorguieva, R. V. and Sanacora, G. (2006) Joint analysis of repeatedly observed continuous and ordinal measures of disease severity. Statist. Med., 25, 1307–1322.

Hannah, L. A., Blei, D. M. and Powell, W. B. (2011) Dirichlet process mixtures of generalized linear models. J. Mach. Learn. Res., 12, 1923–1953.

Harris, K. M., Halpern, C. T., Whitasel, E., Hussey, J., Tabor, J., Entzel, P. and Udry, J. R. (2009) The national longitudinal study of adolescent health: Research design. University of North Carolina, Chapel Hill. (Available from http://www.cpc.unc.edu/projects/addhealth/design.)

Herrick, A., Kuhns, L., Kinsky, S., Johnson, A. and Garofalo, R. (2013) Demographic, psychosocial, and contextual factors associated with sexual risk behaviors among young sexual minority women. J. Am. Psychiatr. Nurs. Ass., 19, 345–355.

Hoff, P. D. (2007) Extending the rank likelihood for semiparametric copula estimation. Ann. Appl. Statist., 1, 265–283.

Hoff, P. D. and Niu, X. (2012) A covariance regression model. Statist. Sin., 22, 729–753.

Horvitz, D. G. and Thompson, D. J. (1952) A generalization of sampling without replacement from a finite universe. J. Am. Statist. Ass., 47, 663–685.

Imai, K. and van Dyk, D. (2005) A Bayesian analysis of the multinomial probit model using marginal data augmentation. J. Econometr., 124, 311–334.

Kim, J. S. and Ratchford, B. T. (2013) A Bayesian multivariate probit for ordinal data with semiparametric random-effects. Comput. Statist. Data Anal., 64, 192–208.

Kottas, A., Müller, P. and Quintana, F. A. (2005) Nonparametric Bayesian modeling for multivariate ordinal data. J. Comput. Graph. Statist., 14, 610–625.

Kunihama, T., Herring, A. H., Halpern, C. T. and Dunson, D. B. (2016) Nonparametric Bayes modeling with sample survey weights. Statist. Probab. Lett., 113, 41–48.

Levy, P. S. and Lemeshow, S. (2008) Sampling of Populations: Methods and Applications. New York: Wiley.

Little, R. J. A. (2004) To model or not to model?: Competing modes of inference for finite population sampling. J. Am. Statist. Ass., 99, 546–556.

Liu, X., Daniels, M. J. and Marcus, B. (2009) Joint models for the association of longitudinal binary and continuous processes with application to a smoking cessation trial. J. Am. Statist. Ass., 104, 429–438.

Lo, A. Y. (1984) On a class of Bayesian nonparametric estimates: I, density estimates. Ann. Statist., 12, 351–357.

Lumley, T. and Scott, A. J. (2013) Two-sample rank tests under complex sampling. Biometrika, 100, 831–842.

Luo, S. and Wang, J. (2014) Bayesian hierarchical model for multiple repeated measures and survival data: an application to Parkinson’s disease. Statist. Med., 33, 4279–4291.

McCabe, J., Brewster, K. L. and Tillman, K. H. (2011) Patterns and correlates of same-sex sexual activity among U.S. teenagers and young adults. Perspect. Sexl Reprod. Hlth, 43, 142–150.

McCabe, H. L., Dick, R. N., Tancredi, D. J., Goldstein, S., Blackburn, S., Silverman, J. G., Monasterio, E., James, L. and Miller, E. (2014) Differences by sexual minority status in relationship abuse and sexual and reproductive health among adolescent females. J. Adolesc. Hlth, 55, 652–658.

McCabe, H. L., Silverman, J. G., Decker, M. R., Agénor, M., Borroto, S., Tancredi, D. J., Zelazny, S. and Miller, E. (2015) Sexual and reproductive health indicators and intimate partner violence victimization among female family planning clinic patients who have sex with women and men. J. Womns Hlth, 24, 621–628.

McCulloch, R. and Rossi, P. (1994) An exact likelihood analysis of the multinomial probit model. J. Econometr., 64, 207–240.

McParland, D., Gormley, I. C., McCormick, T. H., Clark, S. J., Kabudula, C. W. and Collinson, M. A. (2014) Clustering South African households based on their asset status using latent variable models. J. Econometr., 43, 391–411.

Muliere, P. and Tardella, L. (1998) Approximating distributions of random functionals of Ferguson Dirichlet priors. Can. J. Statist., 26, 283–297.

Müller, P., Erkanli, A. and West, M. (1996) Bayesian curve fitting using multivariate normal mixtures. Biometrika, 83, 67–79.

Müller, P. and Quintana, F. A. (2010) Random partition models with regression on covariates. J. Statist. Planing Inf., 140, 2801–2808.
Murray, J. S., Dunson, D. B., Carin, L. and Lucas, J. E. (2013) Bayesian Gaussian copula factor models for mixed data. *J. Am. Statist. Ass.*, **108**, 656–665.

Murray, J. S. and Reiter, J. P. (2016) Multiple imputation of missing categorical and continuous values via Bayesian mixture models with local dependence. *J. Am. Statist. Ass.*, **111**, 1466–1479.

Muthén, B. (1984) A general structural equation model with dichotomous, ordered categorical and continuous latent variable indicators. *Psychometrika*, **49**, 115–132.

Rao, J. N. K. (2011) Impact of frequentist and Bayesian methods on survey sampling practice: a selective appraisal. *Statist. Sci.*, **26**, 240–256.

Riskind, R. G., Tornello, S. L., Younger, B. C. and Patterson, C. J. (2014) Sexual identity, partner gender, and sexual health among adolescent girls in the United States. *Am. J. Publ. Hlth*, **104**, 1957–1963.

Rue, H., Martino, S. and Chopin, N. (2009) Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations (with discussion). *J. R. Statist. Soc. B*, **71**, 319–392.

Sammel, M. D., Ryan, L. M. and Legler, J. M. (1997) Latent variable models for mixed discrete and continuous outcomes. *J. R. Statist. Soc. B*, **59**, 667–678.

Sethuraman, J. (1994) A constructive definition of Dirichlet priors. *Statist. Sin.*, **4**, 639–650.

Si, Y., Pillai, N. and Gelman, A. (2015) Bayesian nonparametric weighted sampling inference. *Bayes Anal.*, **10**, 605–625.

Taddy, M. and Kottas, A. (2012) Mixture modeling for marked Poisson processes. *Bayes Anal.*, **7**, 335–362.

Vandendijck, Y., Fues, C., Kirby, R., Lawson, A. and Hens, N. (2016) Model-based inference for small area estimation with sampling weights. *Spat Statist.*, **18**, 455–473.

Verbeke, G., Fieuws, S., Molenberghs, G. and Davidian, M. (2014) The analysis of multivariate longitudinal data: a review. *Statist. Meth. Med. Res.*, **23**, 42–59.

West, M., Müller, P. and Escobar, M. D. (1994) Hierarchical priors and mixture models, with application in regression and density estimation. *Aspects of Uncertainty: a Tribute to DV Lindley* (eds P. R. Freeman and A. F. M. Smith), pp. 363–386. Chichester: Wiley.

Zangeneh, S. Z., Keener, R. and Little, R. J. A. (2011) Bayesian nonparametric estimation of finite population quantities in absence of design information on nonsampled units. In *Proc. Jt Statist. Meet*. Alexandria: American Statistical Association.

Zangeneh, S. Z. and Little, R. J. A. (2015) Bayesian inference for the finite population total from a heteroscedastic probability proportional to size sample. *J. Surv. Statist. Methodol.*, **3**, 162–192.

Zheng, H. and Little, R. J. A. (2003) Penalized spline model-based estimation of finite population total from probability-proportional-to-size samples. *J. Off. Statist.*, **19**, 99–117.

Zheng, H. and Little, R. J. A. (2005) Inference for the population total from probability-proportional-to-size samples based on predictions from a penalized spline nonparametric model. *J. Off. Statist.*, **21**, 1–20.

Zhou, H. (2014) Accounting for complex sample designs in multiple imputation using the finite population Bayesian bootstrap. *Dissertation*. University of Michigan, Ann Arbor.

Zhou, H., Elliott, M. R. and Raghunathan, T. E. (2016) A two-step semiparametric method to accommodate sampling weights in multiple imputation. *Biometrics*, **72**, 242–252.

**Supporting information**

Additional ‘supporting information’ may be found in the on-line version of this article:

‘Web-based supporting materials for Nonparametric Bayes models for mixed-scale longitudinal surveys’.