ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Applied Statistics
Series C

Appl. Statist. (2017)

Modelling trends in digit preference patterns

Carlo G. Camarda,

Institut National d'Études Démographiques, Paris, France

Paul H. C. Eilers

Erasmus University Medical Center, Rotterdam, The Netherlands

and Jutta Gampe

Max Planck Institute for Demographic Research, Rostock, Germany

[Received March 2015. Final revision November 2016]

Summary. Digit preference is the habit of reporting certain end digits more often than others. If such a misreporting pattern is a concern, then measures to reduce digit preference can be taken and monitoring changes in digit preference becomes important. We propose a two-dimensional penalized composite link model to estimate the true distributions unaffected by misreporting, the digit preference pattern and a trend in the preference pattern simultaneously. A transfer pattern is superimposed on a series of smooth latent distributions and is modulated along a second dimension. Smoothness of the latent distributions is enforced by a roughness penalty. Ridge regression with an L_1-penalty is used to extract the misreporting pattern, and an additional weighted least squares regression estimates the modulating trend vector. Smoothing parameters are selected by the Akaike information criterion. We present a simulation study and apply the model to data on birth weight and on self-reported weight of adults.

Keywords: Birth weight; Composite link model; L_1-penalty; Penalized likelihood; Self-reported weight

1. Introduction

Digit preference is a well-known phenomenon that occurs when people read analogue scales or recall measurements taken a while back. Certain end digits are preferred and reported substantially more often than others. The preferred end digits are usually multiples of 5 and 10, and even numbers are given preference over odd numbers. Strong digit preference is a concern if accurate measurements are essential, such as in medical applications when diagnosis and treatment decisions are based on these measurements. The replacement of many analogue scales by digital displays raised expectations for removal, or at least strong reduction, of digit preference. However, digit preference still is found frequently, and not only laymen are prone to it. Hence raising awareness among professionals and providing training to reduce this kind of misreporting is important and estimating trends in digit preference is a helpful device to monitor improvements.

Accurate measurements of birth weight are particularly important in neonatal intensive care when drug prescription is weight based, and misreporting of weight can be especially detrimental for very small babies. Emmerson and Roberts (2013) reported a study in which birth weights in

Address for correspondence: Carlo G. Camarda, Institut National d'Études Démographiques, 133 boulevard Davout, 75980 Paris cédex 20, France.
E-mail: carlo-giovanni.camarda@ined.fr

 0035–9254/17/66000

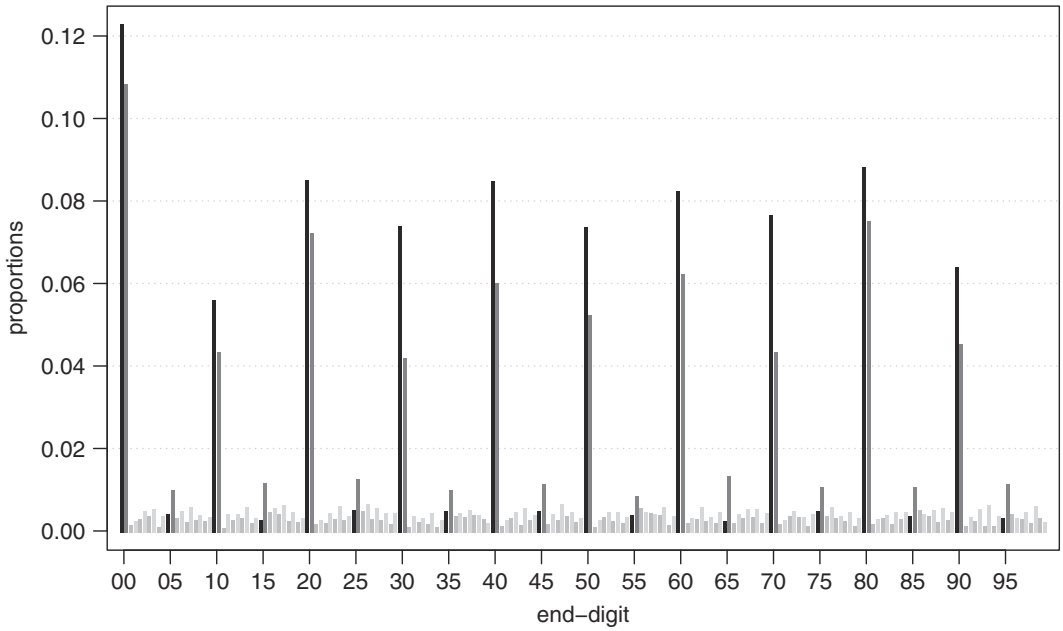

Fig. 1. Proportion of the last two end digits in the birth weight data from Emmerson and Roberts (2013) for the earlier (▮, June 1994–May 2004) and later (▮, June 2004–May 2013) observation period

a neonatal intensive care unit were recorded over 20 years, and birth weights of 9170 newborns were analysed. Weights were obtained by digital scales with 1-g resolution; however, as Fig. 1 shows, considerable digit preference is present. Some reduction of the strong heaping at multiples of 10 is suggested for the second decade of the observation period, but a more detailed analysis of the time trend is needed to identify how the accuracy changed. In this paper we present a model for digit preference and the trend in the preference pattern over a second dimension, which in most cases will be calendar time. We use the same data as Emmerson and Roberts (2013) to estimate the trend in improvement against digit preference in birth weights as well as the preference pattern itself.

As a second application of the approach we analyse self-reported weight (in pounds) recorded in the national health surveys conducted by the US National Center for Health Statistics (e.g. National Health and Nutrition Examination Survey (2013)). In this case the second dimension is the age of the respondent, i.e. we are interested in whether digit preference in self-reported weight varies with age. Fig. 2 shows the weight distribution for a number ages between 20 and 59 years. A strong preference for weights that are multiples of 5 and 10 is obvious, but it is not evident whether older people are more or less prone to digit preference and, consequently, may report their weight more or less accurately.

Digit preference arises because some observations or measurements are not reported at their true value but are rounded upwards or downwards and reported at the preferred end digits. The proportion of observations that are transferred from their actual values to the reported values can vary across the range of observations. Camarda *et al.* (2008) described an approach to model digit preference in a single distribution, which explicitly mimics this data-generating process. The true but unobserved distribution, which is free from any misreporting, is assumed to be smooth, whereas the observed counts at each value are the superposition of the number of correctly reported values and the misreported observations from other end digits. The mis-

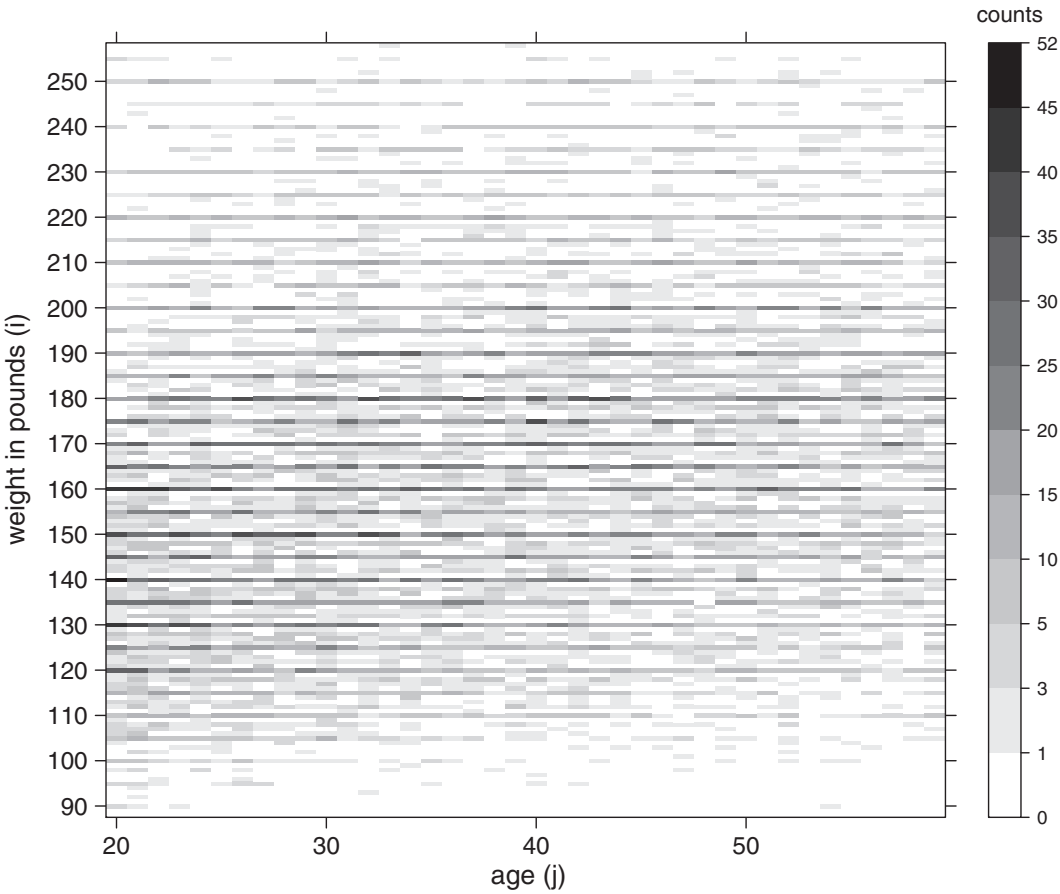

Fig. 2. Distribution of self-reported weight (in pounds) split by the age of the respondents (between 20 and 59 years), from NHANES data: the grey levels represent the number of observations in each cell; the darker horizontal stripes at multiples of 5 and 10 represent the digit preference which is present at all ages

reported proportions are estimated as well. The model can be expressed as a composite link model (CLM) (Thompson and Baker, 1981) with a penalty added to guarantee smoothness of the unobserved true distribution (Eilers, 2007). Regression with an L_1-penalty allows the extraction of the misreporting pattern.

To model trends in digit preference this approach is extended here to a two-dimensional setting. It is assumed that a common misreporting pattern is superimposed on a sequence of true, but latent, distributions. Each of these latent distributions is assumed to be smooth and the change in the distributions along the second dimension (calendar time or age respectively in the birth weight and self-reported weight applications that were described above) is only gradual. If the measurements are recorded closely in time, such as consecutive years, or for adjacent ages, the corresponding distributions will be similar and the smoothness assumption across the second dimension seems natural. The digit preference habit, which is superimposed on the true values, is assumed to have the same structure over time or age, i.e. the relative strength of particular preferences stays the same. However, the overall strength of the pattern can vary over the second dimension. This variation gives the trend in digit preference. For example, in the National Health and Nutrition Examination Survey (NHANES) data, rounding a weight

of 151 lb to 150 lb could consistently be twice as likely as reporting 155 instead of 156 lb, but the general tendency for all rounding preferences may change with the age of the respondent.

The rest of the paper is structured as follows. Section 2 summarizes the essential components of the one-dimensional digit preference model, which are then generalized in Section 3 to allow for a trend in the preference pattern. The model specification and the estimation of the latent distributions are described in Section 3.1, followed by the step to extract the misreporting pattern in Section 3.2. The modelling and estimation of the temporal trend is described in Section 3.3, the selection of smoothing parameters is discussed in Section 3.4 and inference is addressed in Section 3.5. A simulation study and the results for the two applications are presented in Sections 4 and 5 respectively. We close with a discussion and suggestions for possible extensions.

2. Modelling digit preference in one dimension

2.1. The composite link model

Digit preference in a single distribution was modelled by Camarda *et al.* (2008) and the essential steps are described in what follows. Measurement values are denoted by $i = 1, \ldots, I$, and the corresponding observed counts by the vector $\mathbf{y} = (y_1, \ldots, y_I)'$. The elements of $\mathbf{y}$ are assumed to be realizations from Poisson distributions with means $\boldsymbol{\mu} = C\boldsymbol{\gamma}$. The vector $\boldsymbol{\gamma}$ is the unknown latent distribution free from digit preference. This distribution is assumed to be smooth. The matrix $C \in \mathbb{R}^{I \times I}$ embodies the misreporting pattern. Digit preference is conceived as a process that (partly) redistributes counts to neighbouring categories. These can be categories that are more than one step away from the actual value (see Section 5.2) but for simplicity we restrict the presentation here to the case when redistribution occurs only to immediately neighbouring categories. The proportion of counts that is moved from category k to category i is denoted by p_{ik}, which leads (for the particular preference pattern that is considered here) to the following form of C:

$$C = \begin{pmatrix} 1-p_{21} & p_{12} & 0 & 0 & \cdots \\ p_{21} & 1-p_{12}-p_{32} & p_{23} & \cdots & \\ 0 & p_{32} & 1-p_{23}-p_{43} & p_{34} & \vdots \\ 0 & 0 & p_{43} & \ddots & \ddots \\ \vdots & \vdots & \vdots & \ddots & \ddots \\ 0 & \cdots & \cdots & 0 & \ddots \end{pmatrix}.$$

When transfers happen only between neighbouring categories, only the subdiagonal and superdiagonal entries $p_{k,k-1}$ and $p_{k-1,k}$ are non-zero. Different misreporting patterns can be accommodated by modifying the matrix C. The composition matrix can be written as a sum:

$$C = I_I + \begin{pmatrix} -p_{21} & p_{12} & \cdots & 0 \\ p_{21} & -p_{12}-p_{32} & \cdots & \vdots \\ 0 & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & p_{I-1,I} \\ 0 & \cdots & p_{I,I-1} & -p_{I-1,I} \end{pmatrix} = I_I + C_p. \tag{1}$$

In this additive representation I_I denotes the identity matrix of dimension I, and the matrix C_p contains all information about the proportions of counts that are transferred.

The vector $\boldsymbol{\gamma}$ of the latent distribution is given as $\boldsymbol{\gamma} = \exp(X\boldsymbol{\alpha})$ to ensure non-negative elements. In the simplest case X is the identity matrix and $\boldsymbol{\gamma} = \exp(\boldsymbol{\alpha})$. If the number of categories I is large, it is advantageous to express the logarithm of the latent distribution more parsimoniously as a combination of B-splines (Eilers and Marx, 1996). In this case the matrix X is of dimension $I \times R$, where R is the number of B-splines employed. The length of the parameter vector $\boldsymbol{\alpha}$ decreases correspondingly so $\boldsymbol{\alpha} \in \mathbb{R}^R$. Smoothness of $\boldsymbol{\gamma}$ is obtained by enforcing smoothness of $\boldsymbol{\alpha}$. This will be achieved by introducing a difference penalty on the elements of $\boldsymbol{\alpha}$ (see below).

With $\mathbf{y} \sim \text{Poisson}(\boldsymbol{\mu} = C\boldsymbol{\gamma})$, and $\boldsymbol{\gamma} = \exp(X\boldsymbol{\alpha})$, where $\boldsymbol{\alpha}$ is to be estimated, we have a case of the CLM, for which Thompson and Baker (1981) proposed an extension of the iteratively weighted least squares (IWLS) algorithm for generalized linear models. Using $u_{ik} = \Sigma_j c_{ij} x_{jk} \gamma_j / \mu_i$,

$$(U' \tilde{W} U)\boldsymbol{\alpha} = U' \tilde{W} \tilde{\mathbf{z}}, \tag{2}$$

with $\tilde{W} = \text{diag}(\tilde{\boldsymbol{\mu}})$, $\tilde{\mathbf{z}} = \tilde{W}^{-1}(\mathbf{y} - \tilde{\boldsymbol{\mu}}) + U\tilde{\boldsymbol{\alpha}}$, and the tilde denoting current values in the iteration. Smoothness of $\boldsymbol{\alpha}$ can be enforced by subtracting a roughness penalty from the log-likelihood L:

$$L^* = L - \frac{\lambda}{2} \| D\boldsymbol{\alpha} \|^2. \tag{3}$$

D is a difference matrix for the coefficient vector $\boldsymbol{\alpha}$ (Eilers, 2007). The length of the coefficient vector $\boldsymbol{\alpha}$ and, consequently, the size of D depend on the choice of the design matrix X. Third-order differences will be used in what follows.

The penalized log-likelihood (3) leads to a penalized version of the system (2):

$$(U' \tilde{W} U + P)\boldsymbol{\alpha} = U' \tilde{W} \tilde{\mathbf{z}}, \tag{4}$$

where the additional penalty term P is given by $P = \lambda D' D$. The positive regularization parameter λ controls the trade-off between smoothness of $\boldsymbol{\alpha}$ and model fidelity. Once λ has been selected, the system of equations described in equation (4) has a unique solution. The choice of λ is addressed in the following section.

2.2. L_1-penalty for the misreporting pattern

The misreporting proportions, which are contained in the subdiagonal and superdiagonal of C, can be estimated by a constrained weighted least squares regression within the IWLS algorithm.

If $\mathbf{p}$ denotes the proportions concatenated into a vector, we can rewrite the mean $\boldsymbol{\mu} = C\boldsymbol{\gamma}$ as

$$\boldsymbol{\mu} = C\boldsymbol{\gamma} = \boldsymbol{\gamma} + \Gamma\mathbf{p},$$

with

$$\Gamma = \begin{pmatrix} \gamma_2 & 0 & \cdots & 0 & -\gamma_1 & 0 & \cdots & 0 \\ -\gamma_2 & \gamma_3 & & \vdots & \gamma_1 & -\gamma_2 & & \vdots \\ 0 & -\gamma_3 & \ddots & 0 & 0 & \gamma_2 & \ddots & 0 \\ \vdots & & \ddots & \gamma_J & \vdots & & \ddots & -\gamma_{J-1} \\ 0 & \cdots & 0 & -\gamma_J & 0 & \cdots & \cdots & \gamma_{J-1} \end{pmatrix}.$$

Since $\mathbf{y} \sim \text{Poisson}(\boldsymbol{\mu})$, the difference between actual data and the latent distribution can be approximated by the following normal distribution:

$$\mathbf{y} - \boldsymbol{\gamma} \approx \mathcal{N}\{\Gamma\mathbf{p}, \text{diag}(\boldsymbol{\mu})\}. \tag{5}$$

Because of the large number of unknown proportions, Camarda *et al.* (2008) suggested estimating $\mathbf{p}$ by employing an L_1-penalty (Tibshirani, 1996). Such a constrained regression has two advantages: it makes the estimation feasible and it allows us to select only those p_{ik} that exhibit strong effects, whereas shrinking others to 0.

Using the idea of Schlossmacher (1973) the following penalized weighted least squares (PWLS) system can be solved iteratively to obtain estimates of $\mathbf{p}$:

$$(\Gamma' \tilde{V} \Gamma + \tilde{Q})\mathbf{p} = \Gamma' \tilde{V} (\mathbf{y} - \tilde{\gamma}), \tag{6}$$

where $\tilde{V} = \mathrm{diag}(1/\tilde{\boldsymbol{\mu}})$ and

$$\tilde{Q} = \kappa \, \mathrm{diag}\left(\frac{1}{|\tilde{\mathbf{p}}| + \epsilon} \right). \tag{7}$$

Numerical instabilities near zero are avoided by adding a small number, $\epsilon = 10^{-6}$, in the penalty term $\tilde{Q}$.

The size of misreporting proportions p_{ik} is tuned by the smoothing parameter κ. We use a modified version of Akaike's information criterion (AIC) to optimize both λ and κ. Specifically, the effective dimension of the model is computed as the sum of the effective dimensions of the two model components: the penalized CLM in equation (4) and the L_1-penalized weighted least squares (L_1-PWLS) regression in equation (6). For further details, see Camarda *et al.* (2008), section 4.3.

3. Modelling trends in digit preference

The aim of this paper is to provide a tool to study trends in digit preference across several distributions of the same kind of measurements. The number of distributions is denoted by J. Each of the J (latent) distributions of true values, free from digit preference, is assumed to be smooth (as in Section 2). The observations across the second dimension (e.g. calendar time or age) are made in such a way that the assumption of a gradual change in the J true distributions is reasonable (although the number of observations in each of the cross-sections can vary). The latter assumption will allow the exploitation of a smoothness restriction on the second dimension.

The observed data are the result of digit preference patterns superimposed on the true values. The pattern is assumed to have the same structure over the second (trend) dimension; however, the strength of the pattern can vary over time (or age). These assumptions view the structure of the preference pattern as a stable attitude, which can manifest itself to a varying degree over time. Therefore we assume the same misreporting pattern, as expressed in the matrix C, for each of the J cross-sections, but include a vector $\mathbf{g} = (g_j)$, $j = 1, \ldots, J$, whose components act multiplicatively on the elements of the vector $\mathbf{p}$, i.e. one factor g_j per cross-section, and thereby we modulate the strength of the digit preference over the trend dimension.

Several extensions of the unidimensional model are necessary to allow the estimation of trend-modulated digit preference. First the penalized CLM must be extended (Section 3.1) and the L_1-PWLS regression needs to be adjusted accordingly (Section 3.2). The estimation of the additional trend component $\mathbf{g}$ is described in Section 3.3. The second dimension in the latent distributions and the trend component add two additional smoothing parameters to the model and their choice is addressed in Section 3.4. Bootstrap-based inference is discussed in Section 3.5. All estimation procedures were implemented in R (R Core Team, 2016) and the code can be obtained from

`http://wileyonlinelibrary.com/journal/rss-datasets`

3.1. A two-dimensional penalized composite link model

The two-dimensional generalization of penalized IWLS, as presented in equation (4), requires different composition and design matrices, which we shall derive in what follows. We denote by y_{ij} the observed counts for measurement value i, $i = 1, \ldots, I$, in cross-section j, $j = 1, \ldots, J$. The dimension along which the digit preference pattern is modulated will often be time. However, other variables can define the J distributions, and in Section 5.1 we present an application where the trend variable is the age of the respondents. We use the terminology of cross-sections for brevity, no matter which variable defines the trend dimension. For the regression formulation we arrange these count data as a column vector $\mathbf{y} = (y_{11}, \ldots, y_{I1}, \ldots, y_{1J}, \ldots, y_{IJ})' \in \mathbb{R}^{IJ}$. The total numbers of counts in each cross-section j are denoted by $y_{+j} = \Sigma_i y_{ij}$, and we collect these totals in the vector $\breve{\mathbf{n}} = (y_{+1}, \ldots, y_{+J})'$. To match the totals appropriately with the structure of the counts in $\mathbf{y}$, we arrange them in the vector $\mathbf{n} = \text{vec}(\mathbf{1}_{I,1}\breve{\mathbf{n}}')$, where vec$(\cdot)$ converts a matrix into a column vector. The observed data $\mathbf{y}$ are assumed to be Poisson distributed, as in the one-dimensional case. The expected values are $\boldsymbol{\upsilon} = \mathbf{n} * \boldsymbol{\mu}$, where elementwise multiplication is denoted by '$*$'. The expected values incorporate the exposure numbers $\mathbf{n}$ and the vector $\boldsymbol{\mu}$. This vector $\boldsymbol{\mu}$ is linearly composed from the values in a series of latent distributions: $\boldsymbol{\mu} = \breve{C}\boldsymbol{\gamma}$. The elements in $\boldsymbol{\gamma} \in \mathbb{R}^{IJ}$ are arranged in the same order as the elements of $\mathbf{y}$. The matrix $\breve{C}$ is a generalization of the composition matrix C in equation (1) and will be derived in what follows.

As we assume that the general structure of the preference pattern is shared across the second dimension and only the strength of this pattern is modulated, the same composition matrix C_p, as defined in equation (1), is multiplied by a cross-section-specific factor g_j. The matrix C_p contains the transfer proportions across the measurement categories $i = 1, \ldots, I$. Multiplication of all elements in C_p by a factor g_j, which is specific for the cross-section j, modulates the preference pattern in C_p, which is common to all cross-sections. Small values of g_j attenuate the preference pattern, whereas large values amplify it. The sequence of g_j-values defines the trend in digit preference.

If we combine these factors in the J-dimensional vector $\mathbf{g} = (g_1, \ldots, g_J)'$, then we can write the overall new composition matrix $\breve{C}$ efficiently by using the Kronecker product:

$$\breve{C} = I_{IJ} + \text{diag}(\mathbf{g}) \otimes C_p. \tag{8}$$

In the general formulation of the model for the latent distributions we have $\boldsymbol{\gamma} = \exp(\breve{X}\boldsymbol{\alpha})$, where the model matrix $\breve{X}$ now represents a two-dimensional basis for the two-dimensional regression and $\boldsymbol{\alpha}$ is the associated vector of regression coefficients.

For the presentation of the estimation procedure we consider the most simple case and hence will simply use one coefficient for each combination of measurement category and cross-section. This corresponds to a regression matrix $\breve{X}_R = I_I$ for the measurement values and, correspondingly, $\breve{X}_C = I_J$ for the cross-sections. These matrices are combined for the two-dimensional regression model by the Kronecker product, which leads to

$$\breve{X} = \breve{X}_C \otimes \breve{X}_R = I_J \otimes I_I = I_{IJ}.$$

The model can thus be written as

$$\boldsymbol{\mu} = \breve{C}\boldsymbol{\gamma} = \breve{C}\exp(\breve{X}\boldsymbol{\alpha}) = \breve{C}\exp(\boldsymbol{\alpha}). \tag{9}$$

The case when $\breve{X}_R$ and $\breve{X}_C$ are not the identity matrix is discussed at the end of this section.

To adapt the IWLS system (4) for a two-dimensional CLM the design matrix U and the penalty matrix P need to be modified. Owing to the presence of the modulating vector $\mathbf{g}$, the design matrix $\breve{U}$ becomes block diagonal:

$$\breve{U} = \text{diag}(\breve{U}_1, \breve{U}_2, \ldots, \breve{U}_j, \ldots, \breve{U}_J). \tag{10}$$

Matrix $\breve{U}_j$ contains only elements belonging to the jth cross-section. If we denote those components of $\boldsymbol{\mu}$ and $\boldsymbol{\gamma}$ that refer to stratification j by $\boldsymbol{\mu}_j = (\mu_{1j}, \ldots, \mu_{Ij})'$ and $\boldsymbol{\gamma}_j = (\gamma_{1j}, \ldots, \gamma_{Ij})'$ respectively, we can build the matrix

$$E_j = \left(\frac{\gamma_{lj}}{\mu_{kj}}\right)_{k,l} = \text{diag}(\boldsymbol{\mu}_j)^{-1} \cdot (\mathbf{1}_I \otimes \boldsymbol{\gamma}_j') \in \mathbb{R}^{I \times I}, \tag{11}$$

and consequently

$$\breve{U}_j = (I_I + g_j C_p) * E_j. \tag{12}$$

To construct the modified penalty matrix $\breve{P}$ we index the elements of the coefficient vector $\boldsymbol{\alpha}$ by α_{ij} such that the first index i refers to the measurement value and the second index j to the cross-section. We arrange these values in the matrix $\mathcal{A} = (\alpha_{ij}) \in \mathbb{R}^{I \times J}$. The columns of $\mathcal{A}$ will be denoted by $\boldsymbol{\alpha}_{:j}$ and the rows will be written as $\boldsymbol{\alpha}_{i:}'$. The coefficient vector is $\boldsymbol{\alpha}' = (\boldsymbol{\alpha}_{:1}', \ldots, \boldsymbol{\alpha}_{:J}')$.

To enforce smoothness within and across the latent distributions strong differences between neighbouring coefficients both in the rows and in the columns of $\mathcal{A}$ are penalized. This is achieved by the penalty terms

$$\sum_{j=1}^{J} \boldsymbol{\alpha}_{:j}' D_I' D_I \boldsymbol{\alpha}_{:j}$$

and

$$\sum_{i=1}^{I} \boldsymbol{\alpha}_{i:}' D_J' D_J \boldsymbol{\alpha}_{i:},$$

where D_I and D_J are difference matrices of corresponding dimensions (Currie *et al.*, 2004).

These row- and column-specific penalties can be combined into a single penalty matrix $\breve{P}$, which operates on the coefficient vector $\boldsymbol{\alpha}$:

$$\breve{P} = \lambda_R I_J \otimes D_I' D_I + \lambda_C D_J' D_J \otimes I_I, \tag{13}$$

with smoothing parameters λ_R and λ_C.

With these specifications the system of equations for the two-dimensional penalized CLM is

$$(\breve{U}' \tilde{W} \breve{U} + \breve{P}) \boldsymbol{\alpha} = \breve{U}' \tilde{W} \tilde{\mathbf{z}}. \tag{14}$$

The weight matrix is $\tilde{W} = \text{diag}(\tilde{\boldsymbol{\mu}})$ and the model is estimated by iteratively solving the system (14) for the coefficients in $\boldsymbol{\alpha}$.

Using a two-dimensional penalized CLM allows the choice of different smoothing parameters for the measurement values and for the trend dimension, which makes the model quite flexible. Selection of the optimal (λ_R, λ_C) combination will be presented in Section 3.4.

If more general matrices $\breve{X}_R$ and $\breve{X}_C$ are to be used, a few changes to the matrices $\breve{U}$ and $\breve{P}$ are necessary. For example, if we want to model the latent distributions more parsimoniously, we can express the smooth surface by a linear combination of B-splines (Currie *et al.*, 2006). If the numbers of B-splines for each axis are R and C, the regression matrices change to $\breve{X}_R = B_R \in \mathbb{R}^{I \times R}$ and $\breve{X}_C = B_C \in \mathbb{R}^{J \times C}$. Here B_R and B_C incorporate the one-dimensional B-spline basis over each axis, and the final regression matrix $\breve{X}$ for the two-dimensional CLM is

$$\breve{X} = B_C \otimes B_R. \tag{15}$$

The number of columns of $\breve{X}$, and hence the length of the coefficient vector $\boldsymbol{\alpha}$, is now RC.

As a consequence the matrix $\breve{U}$ no longer has a block diagonal structure. If we define $E = \text{diag}(E_1, \ldots, E_J) \in \mathbb{R}^{IJ \times IJ}$ as the matrix with the individual $E_j \in \mathbb{R}^{I \times I}$ along the diagonal (see equation (11)), then we have

$$\breve{U} = (\breve{C} * E) \cdot \breve{X},$$

which for $\breve{X} = I_{IJ}$ reduces to equations (12) and (10).

The coefficients in $\boldsymbol{\alpha}$ now can be arranged in a matrix $\mathcal{A} \in \mathbb{R}^{R \times C}$, and the penalty is applied to both rows and columns of $\mathcal{A}$, which leads to the penalty matrix

$$\breve{P} = \lambda_R I_C \otimes D_R' D_R + \lambda_C D_C' D_C \otimes I_R.$$

The difference matrices D_R and D_C act on the columns and rows of the coefficients in $\mathcal{A}$ and smoothness is controlled by λ_R and λ_C respectively.

3.2. Finding the common misreporting pattern

Besides the series of latent distributions, given by the coefficients $\boldsymbol{\alpha}$, we need to estimate the common misreporting proportions in the composition matrix (8). Similarly to the one-dimensional setting, we use the Poisson assumption and the corresponding normal approximation:

$$\begin{aligned} E(\mathbf{y}) &= \mathbf{n} * (\breve{C}\boldsymbol{\gamma}) = \mathbf{n} * \gamma + \breve{\Gamma}\mathbf{p} \\ &\Rightarrow \mathbf{y} - \mathbf{n} * \gamma \approx \mathcal{N}\{\breve{\Gamma}\mathbf{p}, \text{diag}(\boldsymbol{\mu})\}. \end{aligned} \tag{16}$$

Because of the large number of misreporting proportions, some constraints need to be added to the simple weighted least squares setting. As in the one-dimensional case we choose an L_1-penalty, which allows us to extract the most relevant misreporting proportions.

To enforce non-negative transfer proportions, we introduce an asymmetric penalty within the L_1-PWLS regression.

To estimate the $2(I-1)$ proportions, which are modulated by $\mathbf{g}$, but otherwise shared across cross-sections, the system of equations (6) is modified in the following way: the new model matrix $\breve{\Gamma}$ takes the form

$$\breve{\Gamma} = \begin{pmatrix} g_1 \breve{\Gamma}_1 \\ g_2 \breve{\Gamma}_2 \\ \vdots \\ g_J \breve{\Gamma}_J \end{pmatrix}, \tag{17}$$

where

$$\breve{\Gamma}_j = \begin{pmatrix} -\gamma_{2j} & 0 & \cdots & 0 & \gamma_{1j} & 0 & \cdots & 0 \\ \gamma_{2j} & -\gamma_{3j} & & \vdots & -\gamma_{1j} & \gamma_{2j} & & \vdots \\ 0 & \gamma_{3j} & \ddots & 0 & 0 & -\gamma_{2j} & \ddots & 0 \\ \vdots & & \ddots & -\gamma_{I,j} & \vdots & & \ddots & \gamma_{I-1,j} \\ 0 & \cdots & \cdots & \gamma_{I,j} & 0 & \cdots & \cdots & -\gamma_{I-1,j} \end{pmatrix}.$$

The penalty to enforce non-negative p_{ik} is defined as

$$P_p = \kappa_p \text{diag}(\breve{w}_{ik})$$

with asymmetric weights

$$\breve{w}_{ik} = \begin{cases} 0 & \text{if } p_{ik} \geqslant 0, \\ 1 & \text{otherwise.} \end{cases}$$

This weight becomes effective only if p_{ik} turns negative and the parameter κ_p controls the strength of the penalization. A value of $\kappa_p = 10^6$ works well in our case.

With the additional asymmetric penalty and the generalized model matrix $\breve{\Gamma}$ the system of equations for the transfer proportions $\mathbf{p}$ is

$$(\breve{\Gamma}' \tilde{V} \breve{\Gamma} + \tilde{Q} + P_p)\mathbf{p} = \breve{\Gamma}' \tilde{V}(\mathbf{y} - \mathbf{n} * \tilde{\boldsymbol{\gamma}}), \tag{18}$$

where the shrinkage matrix $\tilde{Q}$ is defined as in the one-dimensional approach; see equation (7). Here the weight matrix V includes the expected values $\boldsymbol{\mu}$ for each measurement value and cross-section, arranged as a column vector. Also, in the two-dimensional setting, the smoothing parameter κ within $\tilde{Q}$ controls the size of the misreporting proportions p_{ik}, shrinking the less important proportions towards 0.

3.3. The temporal trend

To estimate the modulating vector $\mathbf{g}$ we rearrange the normal approximation (16) for $\mathbf{y} - \mathbf{n} * \boldsymbol{\gamma}$. Taking into account that each g_j appears for only a specific j (see equation (17)), we obtain separate weighted least squares equations of the form

$$\min_{g_j} \left\| \frac{\mathbf{y}_j - \mathbf{n}_j * \boldsymbol{\gamma}_j - g_j \mathbf{n}_j * \boldsymbol{\theta}_j}{\sqrt{\boldsymbol{\mu}_j}} \right\|^2 \qquad j = 1, \ldots, J.$$

Here $\mathbf{y}_j$, $\mathbf{n}_j$, $\boldsymbol{\gamma}_j$ and $\boldsymbol{\mu}_j$ represent the I-dimensional subvectors of $\mathbf{y}$, $\mathbf{n}$, $\boldsymbol{\gamma}$ and $\boldsymbol{\mu}$ that correspond to cross-section j. The vectors $\boldsymbol{\theta}_j$ are

$$\boldsymbol{\theta}_j = \begin{pmatrix} -p_{21}\gamma_{1j} + p_{12}\gamma_{2j} \\ p_{21}\gamma_{1j} - p_{12}\gamma_{2j} - p_{32}\gamma_{2j} + p_{23}\gamma_{3j} \\ p_{32}\gamma_{2j} - p_{23}\gamma_{3j} - p_{43}\gamma_{3j} + p_{34}\gamma_{4j} \\ \vdots \\ p_{I,I-1}\gamma_{I-1,j} - p_{I-1,I}\gamma_{I,j} \end{pmatrix} \in \mathbb{R}^I.$$

If the components g_j of the modulating vector can vary freely then we solve J independent equations to obtain

$$\hat{g}_j = s_j / r_j, \tag{19}$$

where

$$s_j = \sum_{k=1}^{I} \frac{y_{kj} - n_{kj}\gamma_{kj}}{\mu_{kj}} \theta_{kj}$$

and

$$r_j = \sum_{k=1}^{I} \frac{\theta_{kj}^2}{\mu_{kj}}.$$

In matrix form we can write $\text{diag}(\mathbf{r}) \cdot \mathbf{g} = \mathbf{s}$.

If the modulating vector is assumed to be smooth we add a difference penalty that operates on neighbouring elements of $\mathbf{g}$ and solve the following system:

$$\{\text{diag}(\mathbf{r}) + P_g\}\mathbf{g} = \mathbf{s}, \tag{20}$$

where

$$P_g = \lambda_g D_g' D_g.$$

The smoothing parameter λ_g tunes the smoothness of $\mathbf{g}$ and it needs to be optimized together with the previously introduced smoothing and shrinkage parameters in equations (13) and (18).

Unless we add some constraints the solutions for $\mathbf{p}$ and $\mathbf{g}$ are identified only up to a multiplicative constant. To obtain a unique solution we chose to normalize $\mathbf{g}$ so that its mean equals 1.

3.4. Selection of the smoothing parameters

The model that has been presented so far assembles three components, namely a two-dimensional penalized CLM, an L_1-PWLS regression and a weighted least squares regression, eventually penalized. Each of these components depends on specific smoothing and regularization parameters. A schematic summary of the model is given in Table 1.

Specifically, if we opt for a smooth modulating vector $\mathbf{g}$, the estimating equations (14), (18) and (20) depend on the combination of the four smoothing and shrinkage parameters λ_R, λ_C, κ and λ_g. To optimize these parameters we minimize the AIC:

$$\mathrm{AIC}(\lambda_R, \lambda_C, \kappa, \lambda_g) = \mathrm{Dev}(\mathbf{y}|\boldsymbol{\mu}) + 2\,\mathrm{ED}, \tag{21}$$

where $\mathrm{Dev}(\mathbf{y}|\boldsymbol{\mu})$ is the deviance of the Poisson model. As effective dimension ED we use the sum of the effective dimension of the three model components: $\mathrm{ED} = \mathrm{ED}_1 + \mathrm{ED}_2 + \mathrm{ED}_3$.

The first term ED_1 denotes the effective dimension of the two-dimensional penalized CLM and ED_2 refers to the L_1-PWLS regression for the common misreporting pattern. The effective dimension for the third component ED_3 is equal to either the length of modulating vector in the case of an unrestricted $\mathbf{g}$, or to the effective dimension of the system in equation (20). In formulae, we have

$$\begin{aligned}
\mathrm{ED}_1 &= \mathrm{tr}\{\breve{U}(\breve{U}'\hat{W}\breve{U} + \breve{P})^{-1}\breve{U}'\hat{W}\}, \\
\mathrm{ED}_2 &= \mathrm{tr}\{\breve{\Gamma}(\breve{\Gamma}'\hat{V}\breve{\Gamma} + \hat{Q} + P_p)^{-1}\breve{\Gamma}'\hat{V}\}, \\
\mathrm{ED}_3 &= \begin{cases} J & \text{general } \mathbf{g}, \\ \mathrm{tr}[\{\mathrm{diag}(\mathbf{r}) + P_g\}^{-1}\mathrm{diag}(\mathbf{r})] & \text{smooth } \mathbf{g}. \end{cases}
\end{aligned}$$

The values of the AIC in equation (21) are explored over a three- or four-dimensional grid of values for λ_R, λ_C, κ and possibly λ_g, to find the optimal combination of smoothing and shrinkage parameters.

Table 1. Summary of the model components, estimation methods and associated smoothing and shrinkage parameters

Model component	*Method*	*Smoothing or shrinkage parameters*
Latent distributions γ_{ij}	Two-dimensional penalized CLM	λ_R, λ_C
Misreporting pattern p_{ik}	L_1-PWLS	κ
Modulating vector g_j	(Penalized) weighted least squares	(λ_g)

3.5. Inference

Each model component is represented as a regression model; therefore, component-specific asymptotic confidence intervals could be formulated. However, the penalized CLM for estimating the latent distributions interacts with the two weighted least squares systems that are used for estimating the misreporting pattern and its modulating trend. We thus opted for confidence intervals constructed via a bootstrap approach.

Specifically, we carried out a non-parametric resampling of our data without making assumptions concerning the distributions of the estimated $\hat{\boldsymbol{\gamma}}$, $\hat{\mathbf{p}}$ and $\hat{\mathbf{g}}$. In each of the J cross-sections we take a random sample of size $\check{n}_j$ from the multinomial distribution with parameters $\pi_{ij} = y_{ij}/\check{n}_j$, $i = 1, \ldots, I$. The resulting counts give the bootstrap sample $y_j^* = (y_{1j}^*, \ldots, y_{Ij}^*)'$ in cross-section j. These are combined for all cross-sections to give the bootstrap sample $\mathbf{y}^* = (y_{11}^*, \ldots, y_{I1}^*, \ldots, y_{1J}^*, \ldots, y_{IJ}^*)'$. We then apply our model to this $\mathbf{y}^*$ data set and calculate the bootstrap version of the latent distribution $\hat{\boldsymbol{\gamma}}^*$, the misreporting pattern $\hat{\mathbf{p}}^*$ and the common modulating trend vector $\hat{\mathbf{g}}^*$. We repeat sampling and modelling steps 500 times to obtain bootstrap distributions of each model component.

The bootstrap latent distributions, misreporting proportions and modulating trends can then be used to form the 95% non-parametric bootstrap confidence interval for $\hat{\boldsymbol{\gamma}}$, $\hat{\mathbf{p}}$ and $\hat{\mathbf{g}}$. Among several options, we preferred the bootstrap percentile interval, since it avoids the normal assumption and is more reliable than the standard normal interval (Efron and Tibshirani, 1998). For the latent distributions, we simply take the empirical percentiles from the bootstrap distribution of $\boldsymbol{\gamma}$: $(\boldsymbol{\gamma}_{(\alpha/2)}^*; \boldsymbol{\gamma}_{(1-\alpha/2)}^*)$, where $\boldsymbol{\gamma}_{(1-\alpha/2)}^*$ denotes the $(1-\alpha/2)$-percentile of the bootstrap estimates $\boldsymbol{\gamma}^*$. In the same way, the 95% bootstrap confidence intervals for $\hat{\mathbf{p}}$ and $\hat{\mathbf{g}}$ are obtained.

4. Simulation study

To demonstrate the performance of our approach we applied it to a simulated scenario. There were $I = 38$ categories and $J = 15$ cross-sections. The overall sample size was $\Sigma_{i,j} y_{ij} = 6770$ (this was chosen to be similar to the size of the data that are analysed in Section 5.2). The true latent distributions γ_{ij} are shown in Fig. 3(a). They were created as a sequence of Gaussian densities with shifting means and changing variances. The digit preference pattern was set so that the values 10, 20 and 30 receive additional observations from their two neighbours (i.e. from 9 and 11 to 10, etc.). The proportion of counts that go to the target digits was set to 0.6 for $g_j = 1$. The trend vector $\mathbf{g}$ was a sum of a linear function and a sine wave.

The data $\mathbf{y}$, containing $IJ = 38 \times 15 = 570$ counts, were simulated from the Poisson distribution with mean $\boldsymbol{\mu} = \breve{C}\boldsymbol{\gamma}$, where $\breve{C}$ was constructed as in equation (8). These counts are portrayed in Fig. 3(b). Fig. 4 presents the fitted values $\hat{\gamma}_{ij}$ for the $J = 15$ distributions, together with 95% bootstrap confidence intervals, which were determined as described in Section 3.5. Note that the $\hat{\gamma}_{ij}$ were estimated as a two-dimensional surface and are given as individual cross-sections only to show the confidence bands. The optimal $(\lambda_I, \lambda_J, \kappa)$ combination of the smoothing parameters was obtained by minimizing the AIC as described in Section 3.4. The grid extended over $5 \times 5 \times 5 = 125$ values for λ_I, λ_J and κ (computing time 4.2 min on a portable personal computer, Intel i5-3320M processor, 2.6 GHz and 4 Gbytes random-access memory).

Fig. 5 shows the estimated misreporting proportions (Fig. 5(a)) and the estimated trend $\hat{\mathbf{g}}$ that modulates the digit preference (Fig. 5(b)). The results for the common digit preference pattern are presented in the following way: in the centre is the measurement axis (of length I). Many categories could receive additional observations due to digit preference; however, in the

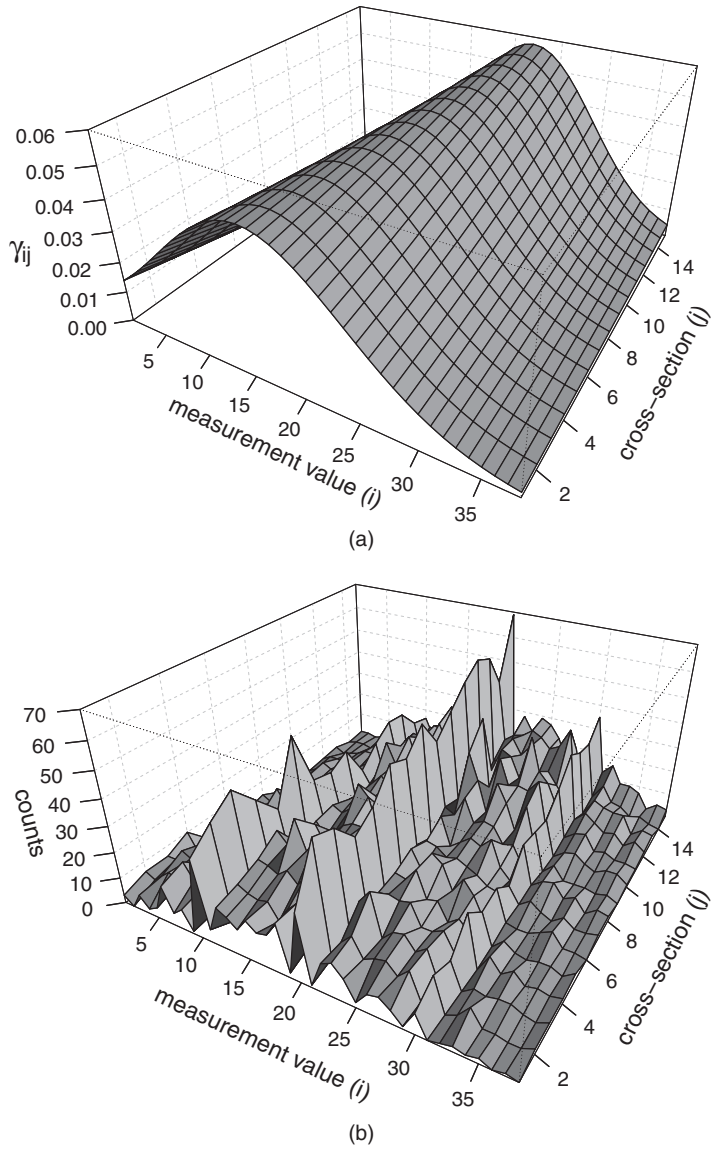

Fig. 3. Simulated data: (a) true latent distribution γ_{ij} ($I = 38$ measurement values and $J = 15$ cross-sections); (b) simulated counts following a common digit preference pattern ($\mathbf{y} = p(\tilde{\mathbf{C}}\gamma\mathbf{n})$) that is modulated over the cross-sections (see also Fig. 5)

simulation study only the values 10, 20 and 30 were receiving extra counts. At the same vertical level are the estimated proportions of counts that go from the left and right neighbour to the target digit. For example, at target digit 30 these are the proportion of counts at 29 and 31 that are transferred to 30. As can be seen only the proportions for target values 10, 20 and 30 have intervals bounded away from zero. All the other proportions were correctly shrunk to 0. The true values that were used in the simulation are indicated by triangles. The intervals are wider at the low and high end of the measurement axis, where the true γ_{ij} are low and hence uncertainty is higher.

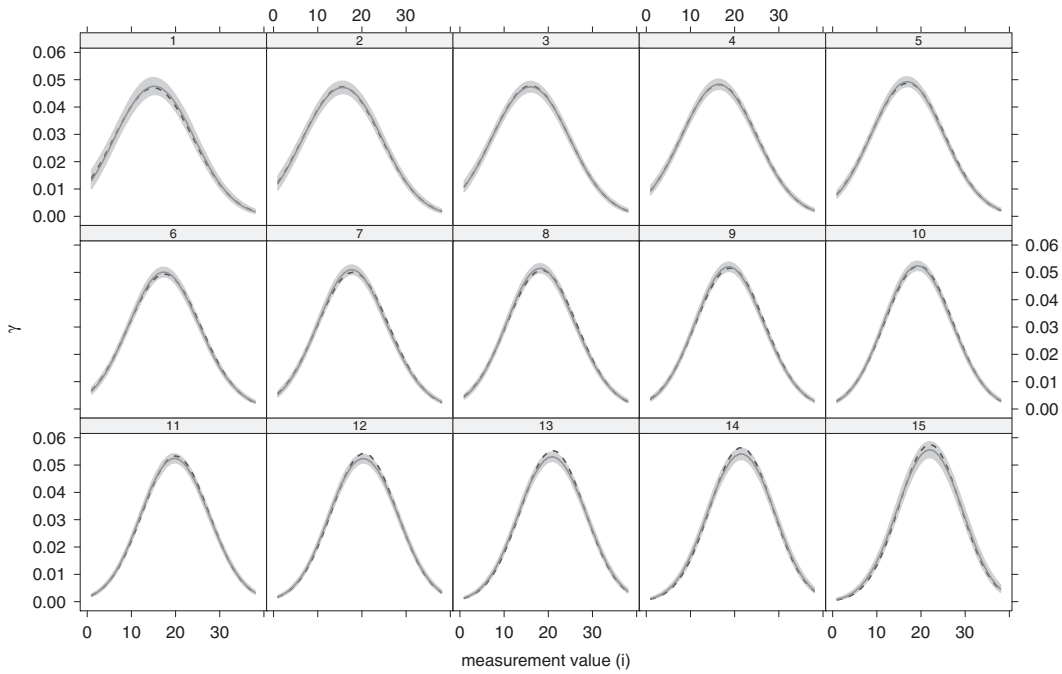

Fig. 4. Fitted latent distributions $\hat{\gamma}_{ij}$ for the simulation study: each panel shows one cross-section; the true (– – –) and estimated (———) values are shown with 95% bootstrap intervals added

Note that the data were simulated with a smooth trend vector $\mathbf{g}$, but to estimate the modulating trend no smoothness assumption was employed. Nevertheless the estimated trend vector $\hat{\mathbf{g}}$ captures the assumed trend function well.

This simulation is limited in that it considers only the case of transfers that occur to immediately neighbouring categories, but it demonstrates that the three model components—the sequence of true distributions that are unaffected by digit preference, the preference pattern itself and the trend in the strength of the preference habit along the second axis—can be determined simultaneously.

5. Applications

5.1. Self-reported weight across age

The self-reported weights from the NHANES data were presented in Fig. 2. The weights range from 88 to 258 lb and the age range is from 20 to 59 years. The overall sample size is $N = 24\,000$. This leads to $I = 171$ measurement values (weights in pounds) and $J = 40$ different weight distributions. Of the $171 \times 40 = 6840$ categories 2072 contain zero observations. The size of the problem motivates the use of (cubic) B-splines to represent the two-dimensional surface of the true distributions. We chose $R = 37$ and $C = 11$, which reduce the length of the parameter vector from $IJ = 6480$ to $RC = 407$. The choice of R and C was such that we placed equidistant knots every 5 lb and 5 years of ages. The composition matrix $\breve{C}$ was constructed such that all values potentially can attract observations from their immediate left-hand and right-hand neighbours. The trend vector $\mathbf{g}$ was assumed to change gradually and hence was estimated by using a smoothness penalty; see equation (20). For optimizing the amount of smoothness in the

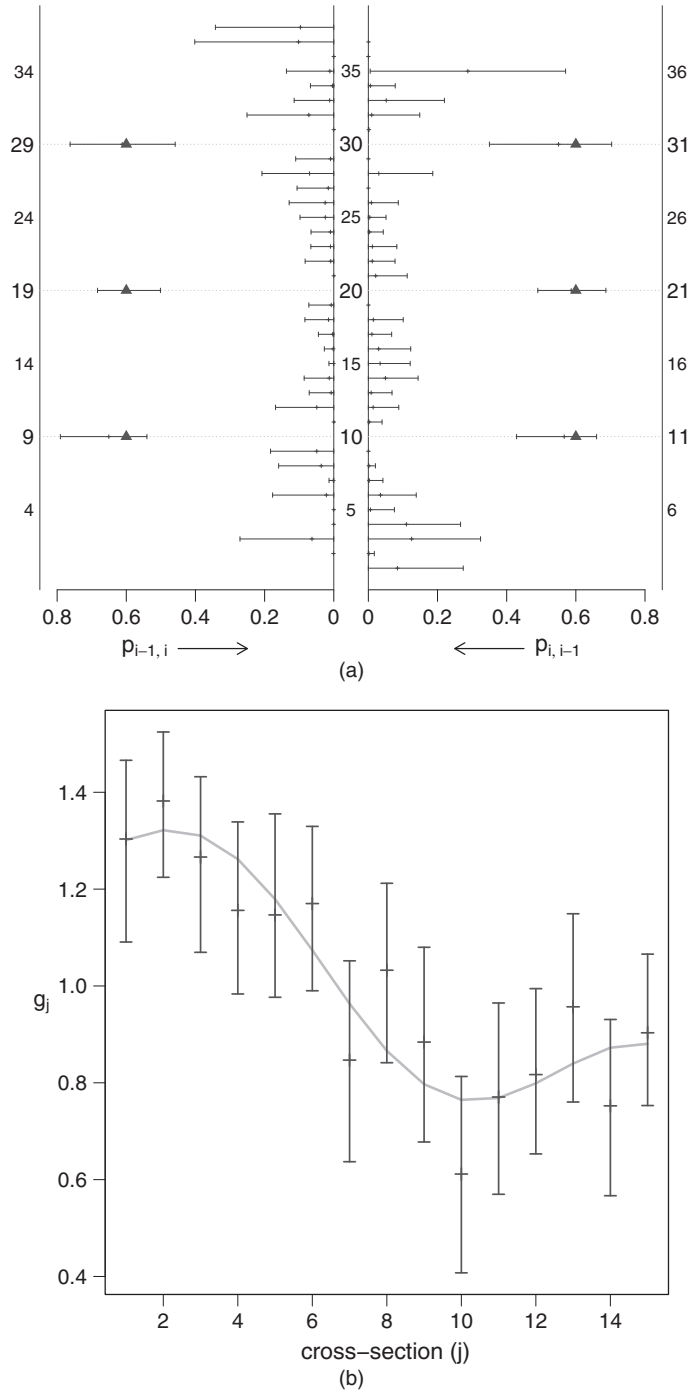

Fig. 5. (a) Estimated digit preference pattern p_{ik} with $g_j = 1$ (▲, true $p \neq 0$; ⊢●⊣, $\hat{p}$ + 95% confidence interval) and (b) trend vector (——, true $\mathbf{g}$; ⟟, $\hat{\mathbf{g}}$ + 95% confidence interval), which multiplies the common misreporting pattern, with bootstrap intervals added: all misreporting proportions, except those that transfer observations to the receiving values 10, 20 and 30, are correctly shrunk to 0; the trend vector was estimated without assuming smoothness

latent distributions and modulating vector as well as the number of non-zero p_{ik}, we minimize the AIC by performing a grid search over a total of $3^4 = 81$ combinations of smoothing and shrinkage parameters. This search was completed in 24 min (portable personal computer, Intel i5-3320M processor, 2.6 GHz and 4 Gbytes random-access memory).

Fig. 6 shows the self-reported weights for selected ages and the corresponding estimated latent distributions. The model produces smooth distributions free from digit preference, with modes that gradually shift upwards with age. For comparison, we provide in Appendix A outcomes from a standard smoothing technique such as *P*-splines (Eilers and Marx, 1996): ignoring digit preference, estimates clearly overfit the data when the amount of smoothness is selected automatically.

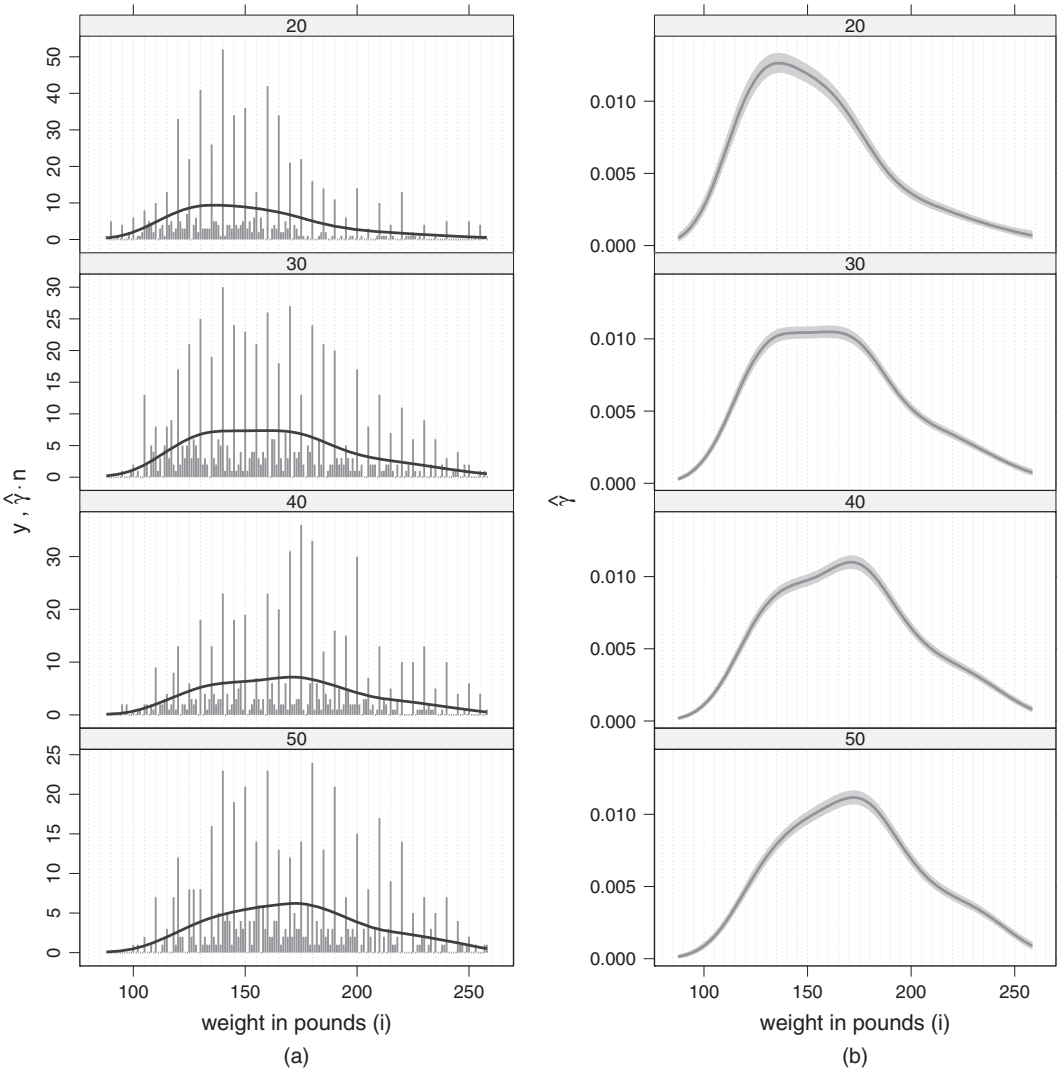

Fig. 6. Self-reported weights (in pounds) and fitted latent distributions for selected ages for the NHANES data: (a) estimates (———) superimposed on the counts that are affected by digit preference (▮); (b) the same estimated distributions (———), but on the same density scale with 95% bootstrap intervals (▮)

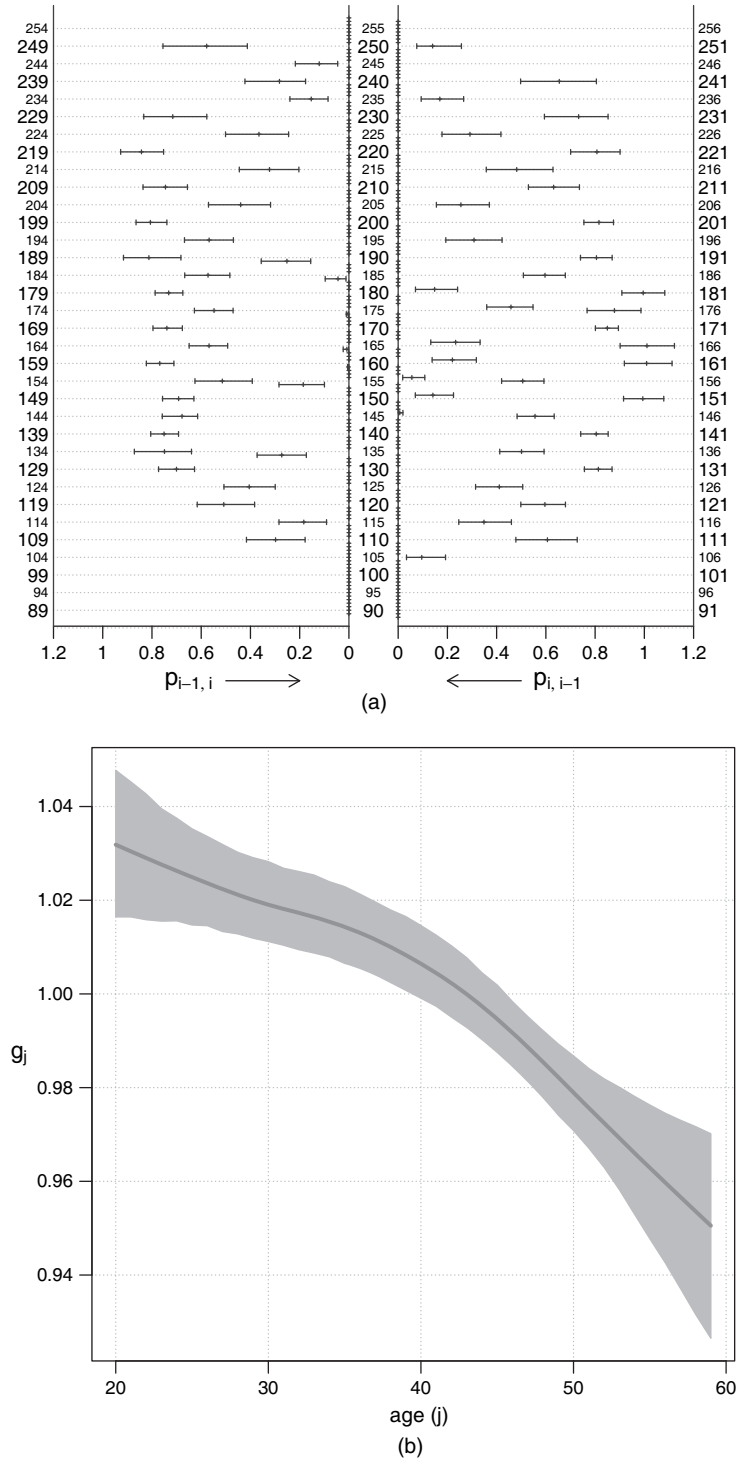

Fig. 7. NHANES weight data: (a) fitted misreporting proportions (p_{ik} with $g_i = 1$; ⊢●⊣, $\hat{\mathbf{p}}$ + 95% confidence interval); (b) estimated trend vector (———) for the preference pattern (■, 95% confidence interval)

Fig. 7(a) shows the estimated misreporting proportions (for $g_j = 1$). As is already clear from Fig. 6, digit preference is very strong in these data, so the high proportions of observations that are transferred to neighbouring multiples of 5 and 10 are not surprising. Misreporting proportions that are greater than 1 are consequences of the extremely high level of digit preference in these data. Weights ending in 0 and 5 are often surrounded by categories with very few counts; therefore the associated p_{ik} can result in values that are bigger than 1 since p_{ik} is defined as proportions of counts in k redistributed to i. If multiples of these (low) numbers must be transferred to fit to the observed counts, values of p_{ik} that are greater than 1 result. Most of the other $p_{i-1,i}$ and $p_{i,i-1}$ are shrunk to 0.

The smooth modulating vector $\mathbf{g}$ describes the changes in the strength of digit preference over age, as shown in Fig. 7. This is a decreasing trend and it accelerates after about age 40 years. Confidence intervals show that differences in the strength of the digit preference between the young and the old are significant, though not strong, ranging from $\hat{g}_1 = 1.03$ at age 20 to $\hat{g}_{40} = 0.95$ at age 59 years. The decline in digit preference with age could be explained with increasing health awareness or more frequent visits to the doctor, with weight being measured exactly more regularly and consequently recalled more accurately. Certainly older individuals, in the age range that is considered here, are not more prone to digit preference; however, the general level of digit preference is high for self-reported weight in these data.

5.2. Birth weights across time

The data from Emmerson and Roberts (2013) contain birth weights between 500 and 4500 g, collected between June 1993 and May 2013. The number of observations that were available for the first year is only about a fifth of the numbers in the following 19 years and may be in some way selective, so we incorporate only the period from June 1994 to May 2013 in the analysis. The weights are available at 1-g resolution, so the dimensions are $I = 4001$ measurement values and $J = 19$ years, and consequently the length of the vector of counts $\mathbf{y}$ is $4001 \times 19 = 76\,019$. There are 9080 observations in total and 91.4% of the elements of $\mathbf{y}$ are 0. The size of the problem combined with the sparsely available counts creates some extra challenges, which we discuss in what follows.

Following the results that were reported in Emmerson and Roberts (2013) we focus on a model for the heapings at multiples of 10. We assume that several neighbouring categories, not only the immediate two, contribute to the heaping. Counts at, for example, ⌟⌟20 arise from misreporting of a proportion of counts at from ⌟⌟16 to ⌟⌟19 as well as from ⌟⌟21 to ⌟⌟24. The proportion depends on the distance to the target. These proportions are denoted by p_1^{20} (for ⌟⌟19 and ⌟⌟21) to p_4^{20} (for ⌟⌟16 and ⌟⌟24). We allow that different multiples of 10 have different p_w^d (d for 'deca' as shorthand for multiples of 10, ranging from 00 to 90; $w = 1, \ldots, 4$). However, we do not discriminate between, say, 3420 and 3520. Observations at multiples of 5 are not transferred.

The composition matrix is constructed from C_0, which incorporates this misreporting pattern. C_0 is a block diagonal matrix over i,

$$C_0 = \text{diag}(\ldots, C^{00}, C^{10}, \ldots, C^{d}, \ldots, C^{90}, C^{00}, C^{10}, \ldots),$$

where the superscript denotes the multiple of 10 attracting counts from the neighbouring four categories on both sides. A generic C^d is given by

$$C^d = \begin{pmatrix} \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & -p_4^d & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & -p_3^d & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & -p_2^d & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & -p_1^d & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & p_4^d & p_3^d & p_2^d & p_1^d & \cdot & p_1^d & p_2^d & p_3^d & p_4^d \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & -p_1^d & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & -p_2^d & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & -p_3^d & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & -p_4^d \end{pmatrix}.$$

In this way, 40 different misreporting probabilities, four for each multiple of 10, constitute the misreporting pattern. The misreporting pattern in C_0 is then modulated over the years by the trend vector $\mathbf{g} = (g_1, \ldots, g_J)'$ so the final composition matrix is

$$\breve{C} = I_{IJ} + \mathrm{diag}(\mathbf{g}) \otimes C_0.$$

Again the size of the problem immediately suggests the use of a B-splines representation of the latent distributions $\boldsymbol{\gamma}$, to reduce the size of the system of equations to be solved. We placed 100 equidistant knots over the range of birth weights, leading to a total number of $R = 103$ cubic B-splines. Furthermore we chose $C = 6$ cubic B-splines over the 19 years, i.e. we placed a knot at every fifth data point along the trend axis.

The latent $\boldsymbol{\gamma}$ are needed at a 1-g resolution to estimate the misreporting probabilities; however, reliable estimates of the γ_{ij} can be achieved even when the weights are binned in intervals of length 100 g. This only changes the composition matrix to $C_G = Q\breve{C}$, where $Q = I_{IK} \otimes \mathbf{1}_{1,100}$, with K denoting the number of 100-g intervals. The elements q_{il} of Q are equal to 1, if weight i is contained in class l, and to 0 otherwise. Again $\boldsymbol{\mu} = C_G\boldsymbol{\gamma}$ and a CLM results. Therewith we still can estimate the γ_{ij} at 1-g resolution but can reduce the computation time by a factor of 20.

As in the previous application, both the misreporting probabilities in C_0 and the modulating vector $\mathbf{g}$ are estimated by weighted least squares systems. However, since we work at different levels of resolution in the penalized CLM and in the estimation of the misreporting pattern, equation (16) is modified as follows:

$$\mathbf{y} - \mathbf{n} * \boldsymbol{\gamma} \approx \mathcal{N}\{\breve{\Gamma}\mathbf{p}, \mathrm{diag}(\breve{\boldsymbol{\mu}})\},$$

where $\breve{\boldsymbol{\mu}} = \breve{C}\boldsymbol{\gamma}$ holds the expected values at a 1-g resolution. The design matrix $\breve{\Gamma}$ maintains the structure as in equation (17), but for the specific preference pattern the year-specific components are given by

$$\breve{\Gamma}_j = \begin{pmatrix} \breve{\Gamma}_j^{500} & & & \\ & \breve{\Gamma}_j^{510} & & \\ & & \ddots & \\ & & & \breve{\Gamma}_j^{590} \\ \breve{\Gamma}_j^{600} & & & \\ & \breve{\Gamma}_j^{610} & & \\ & & \ddots & \\ & & & \breve{\Gamma}_j^{690} \\ \vdots & \vdots & \vdots & \vdots \end{pmatrix}.$$

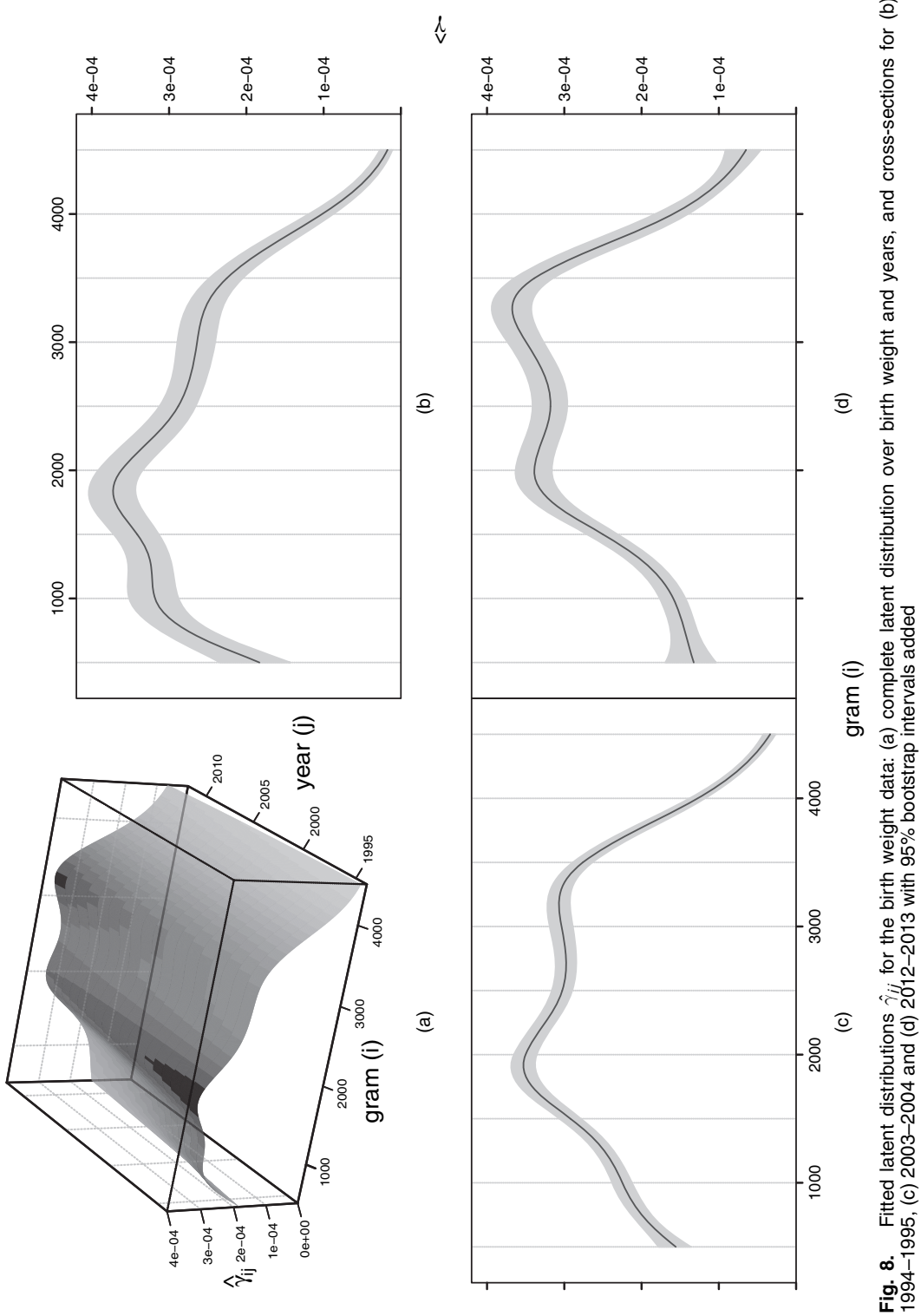

Fig. 8. Fitted latent distributions $\hat{\gamma}_{ij}$ for the birth weight data: (a) complete latent distribution over birth weight and years, and cross-sections for (b) 1994–1995, (c) 2003–2004 and (d) 2012–2013 with 95% bootstrap intervals added

Each of the submatrices $\breve{\Gamma}^i_j$ has the form

$$\breve{\Gamma}^i_j = \begin{pmatrix} \cdot & \cdot & \cdot & \gamma_{i-4,j} \\ \cdot & \cdot & \gamma_{i-3,j} & \cdot \\ \cdot & \gamma_{i-2,j} & \cdot & \cdot \\ \gamma_{i-1,j} & \cdot & \cdot & \cdot \\ \gamma_{i-1,j}+\gamma_{i+1,j} & \gamma_{i-2,j}+\gamma_{i+2,j} & \gamma_{i-3,j}+\gamma_{i+3,j} & \gamma_{i-4,j}+\gamma_{i+4,j} \\ \gamma_{i+1,j} & \cdot & \cdot & \cdot \\ \cdot & \gamma_{i+2,j} & \cdot & \cdot \\ \cdot & \cdot & \gamma_{i+3,j} & \cdot \\ \cdot & \cdot & \cdot & \gamma_{i+4,j} \end{pmatrix}.$$

In this application, since we estimate relatively few misreporting proportions, there is no need to have an additional penalty for ensuring positive estimates. Furthermore we aim to estimate all 40 p_{ik}; therefore no L_1-penalty is needed for shrinking some of the proportions to 0. As a consequence of the structure of the composition matrix a modification of the set-up for the estimation of $\mathbf{g}$, as given in equations (19) and (20), is also required. Specifically, we have $\boldsymbol{\theta}'_j = (\boldsymbol{\theta}^{500}_j, \boldsymbol{\theta}^{510}_j, \ldots, \boldsymbol{\theta}^{4500}_j)$, where

$$\boldsymbol{\theta}^i_j = \begin{pmatrix} -\gamma_{i-4,j}\, p^d_4 \\ -\gamma_{i-3,j}\, p^d_3 \\ -\gamma_{i-2,j}\, p^d_2 \\ -\gamma_{i-1,j}\, p^d_1 \\ (\gamma_{i-1}+\gamma_{i+1})\, p^d_1 + (\gamma_{i-2}+\gamma_{i+2})\, p^d_2 + (\gamma_{i-3}+\gamma_{i+3})\, p^d_3 + (\gamma_{i-4}+\gamma_{i+4})\, p^d_4 \\ -\gamma_{i+1,j}\, p^d_1 \\ -\gamma_{i+2,j}\, p^d_2 \\ -\gamma_{i+3,j}\, p^d_3 \\ -\gamma_{i+4,j}\, p^d_4 \end{pmatrix},$$

for a category i which belongs to the multiple of 10 d. We assume a smooth change of $\mathbf{g}$ over time; we thus have a smoothness penalty (see equation (20)). In this application we chose a grid of size 60 for minimizing the AIC. The search was completed in 15 min (with the same hardware as before).

When we fit the model to the birth weight data we obtain the results that are presented in Figs 8 and 9. The distribution of birth weight free from digit preference is bimodal with higher values just before 2 kg and for weights slightly over 3 kg. Whereas the first peak is the result of numerous low weight births within a neonatal intensive care unit, the second peak emerges only in the last years.

We analysed the full data set as well as the subset of low weight babies below 2500 g. Multiples of 100 g (end digits 00) clearly attract more observations than the other multiples of 10 (Fig. 9(a)). For all groups of proportions, except for end digits 00, misreporting probabilities are higher for categories next to the target multiple of 10, with a surprising slightly higher value of p^d_w at distance $w = 4$ than at distance $w = 3$. We address this issue in Section 6. If considered across the different multiples of 10, they are lowest for 10 and 90, whereas for weights ending with 20 and 80 the p^d_w are noticeably higher. The misreporting pattern for infants with low birth weight is slightly different from the pattern for the complete data set: all estimated p^d_w are lower than for the full data, and misreporting towards multiples of 100 g does not stand out so clearly from the other multiples of 10. Measurements of low weight babies are presumably

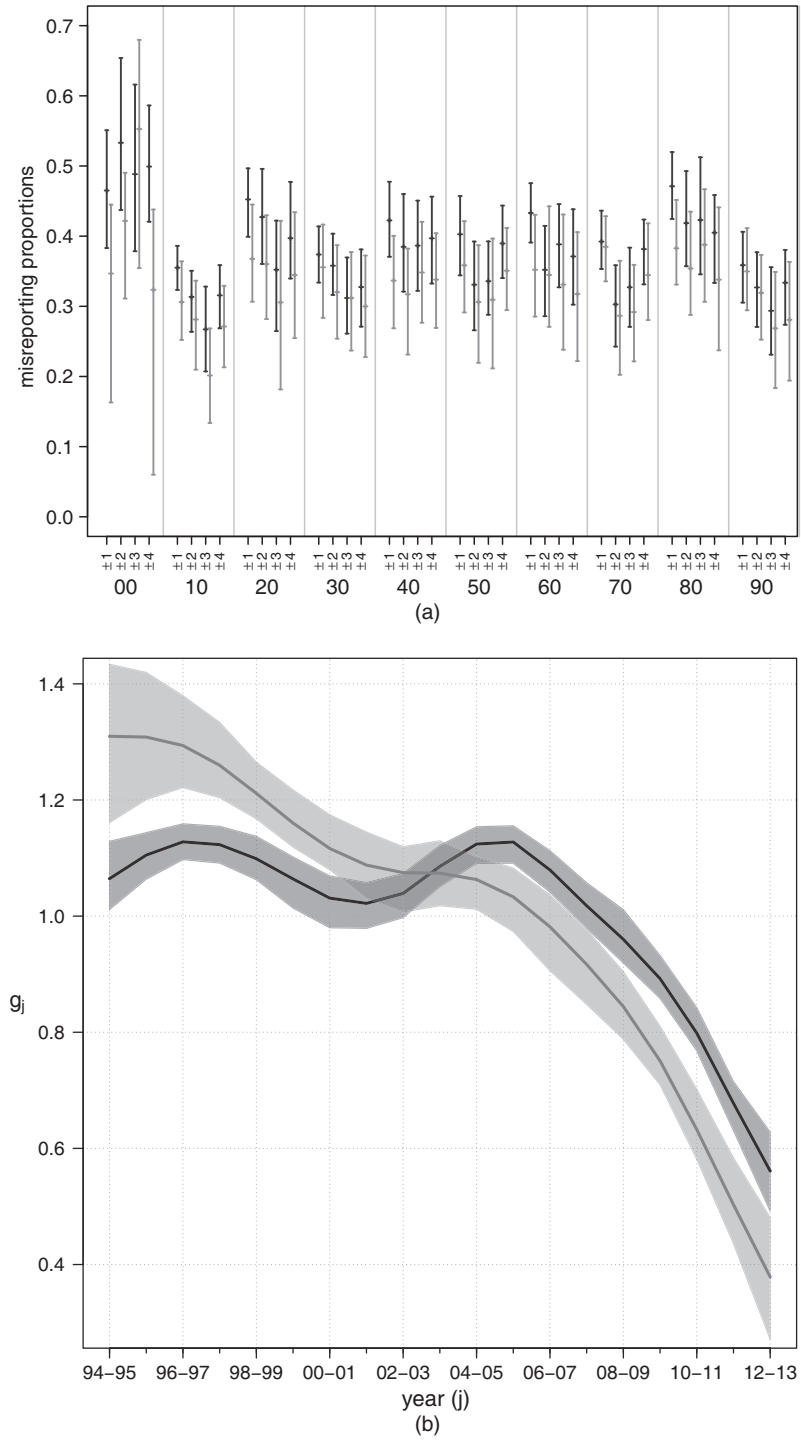

Fig. 9. Birth weight data: (a) fitted misreporting proportions (I, $\hat{\mathbf{p}}$ + 95% confidence interval, all data; I, $\hat{\mathbf{p}}$ + 95% confidence interval, below 2500 g); (b) estimated trend vector for the preference pattern (———, $\hat{\mathbf{g}}$ + 95% confidence interval (■), all data; ———, $\hat{\mathbf{g}}$ + 95% confidence interval (■), below 2500 g)

more accurate because of the awareness that these babies are more sensitive to overtreatment or undertreatment in drug prescription.

The strength of this pattern changed over time as described by **g** in Fig. 9(b). There has been a substantial improvement in the accuracy of birth weight measurements over the period 1994–2013. The improvement was particularly strong for the last years, after about 2005, and the improvement starts earlier for the infants with low birth weights.

6. Discussion

We have presented a model for digit preference that varies over time, age or another variable. The model assumes a smooth underlying two-dimensional density, generating latent data which are partially misclassified because of the digit preference habit of the person(s) collecting the data. It differs from Camarda *et al.* (2008) because it provides a description of changes in digit preference which is critical in assessing learning effects or organizational measures: the misclassification pattern is assumed to be structurally stable, but its strength may change.

To acquire confidence in the model we applied it to simulated data, obtaining reassuring results. We also used it to study two real data sets. One is on self-reported weights and the question was whether digit preference becomes more or less strong at higher ages. Clinical trials and therapeutic decisions in medical practices are obtained by self-reported weights (Crawford *et al.*, 2002) and self-reported anthropometric data are commonly used to estimate the prevalence of obesity in population and community-based studies (Lu *et al.*, 2016; Bolton-Smith *et al.*, 2000; Bowring *et al.*, 2012; Dekkers *et al.*, 2008). An understanding of the changes over ages in the inaccuracy of those measures can help to target age groups for improving and correcting final outcomes of research studies.

The second example concerns birth weights of newborns and its relevance is evident. Decisions on the treatment of very small babies are partially based on birth weight and accuracy is important. Emmerson and Roberts (2013) reported on an improvement campaign in a Manchester (UK) hospital to reduce digit preference. We applied the model to their data. It offers tools for monitoring progress, especially the trend of the strength of digit preference. It can show how strong and fast progress is, and when it levels off it can warn that no further gains are to be expected.

The results of our analysis show that digit preference became less strong over the years, but that it was not eliminated. Now that scales are digital, we might wonder why this is so. A possible explanation is that, even though a scale is digital, it takes time for the number on the display to settle to a stable value. Or it might fluctuate slightly, because of movements of the baby. In both cases a nurse might choose to round to an easy number.

Our model corrects for misclassification and delivers a more reliable estimate of the underlying density than can be obtained from the raw data. Summaries like quantiles can be estimated with higher precision and changes over time in the population under study can be detected earlier.

We believe that our model can be of value in many places. It can be used to determine the quality of many registrations and how that changed over time. When campaigns are started to improve procedures, progress can be monitored.

One reviewer posed an intriguing question: does digit preference in weight, and possibly in height, influence the body mass index (BMI)? The surprising answer is ‘no’. To see this, consider, for example, all people with a reported weight of 150 lb. To compute their BMI, 150 is divided by the square of their height. However, these heights are not all equal, so many different values are obtained, forming a distribution. The concentrated probability mass at 150 lb is smeared

out over a relatively large range of BMIs. The sum of all such smears gives the distribution of all BMIs. We have checked this for the NHANES data and indeed the BMI shows no trace of digit preference.

The estimated misreporting pattern in Fig. 9 shows one surprising detail: higher values of the estimated misreporting proportions are found for observations that are four digits away from the target than for a distance of three digits. This is likely to be due to the fact that we assumed that multiples of 5 do not receive extra observations. We attempted to generalize our model in this respect. We allowed the same category to transfer some of its counts to either the closest digit ending with 0 or ending with 5, so, for example, latent counts in 2343 could have been misreported as either 2340 or 2345. Although this extension worked well on simulated data, the first results on the real data were not satisfactory. One reason is that the birth weight data are too sparse and do not contain sufficiently large counts in all categories to inform the model about this dual option; the final outcomes always favoured ending with 0.

A second peculiarity became clear at closer inspection of Fig. 1: whereas the trend in digit preference towards categories ending with 0 declines, it looks as if the proportions of counts in categories ending with 5 increase slightly. Thus a common modulating trend vector, which the current version of the model assumes and which worked well for weight multiples of 10, would not apply when we include the multiples of 5. A possible explanation of the unexpected difference in digit preference trends might be that a campaign for less rounding towards 0 is only partially successful in the sense that many healthcare workers already feel satisfied when rounding to 5. We plan a generalization of our model which will allow different trends for the misreporting patterns towards weights ending with 0 and with 5.

A further, but even more challenging, model extension would be to modulate a single preference pattern for 00, 10, …, 90 both across weight (i.e. multiples of 100) and time. Such an approach would allow a study of misreporting patterns over time and across the weight range without splitting the weight axis into subgroups, as we did in Section 5.2. We shall pursue this idea in future research.

Acknowledgements

The authors thank Dr Stephen A. Roberts who kindly provided the birth weight data from Emmerson and Roberts (2013). The comments of two reviewers and the Associate Editor on an earlier version helped to improve the paper considerably.

Appendix A

A reviewer remarked that it is not necessary to use a complicated model, like ours, to obtain a smooth density estimate from the self-reported weight data that were presented in Section 5.1. This is true when the amount of smoothness is subjectively chosen. Fig. 10(a) shows a histogram with bins of width 1 and a smooth series, computed with *P*-splines (Eilers and Marx, 1996). The data are for females between 40 and 50 years of age. The choice of the *B*-spline basis is special, because a knot was placed at the centre of each bin. Therefore the basis reduces to the identity matrix, its size being equal to the number of bins (so that $\gamma = \exp(I\alpha) = \exp(\alpha)$; see the first paragraph on the fifth page). This choice avoids discussions about the size of the basis (see Eilers *et al.* (2015) for further details). The value of the smoothing parameter is quite high: $\lambda = 10^6$. It was chosen subjectively, to obtain a pleasing result. A similar result can be obtained with almost any smoother, with proper subjective tuning.

An objective method for choosing λ is more attractive and will be preferred in most cases. Figs 10(b) and 10(c) show what happens for these data: the profile of the AIC (Fig. 10(b)) has a minimum close to $\lambda = 0.04$, which is a very low value. The resulting 'smooth' fit is shown in Fig. 10(c): it follows the observed counts very closely, and the AIC seemingly fails.

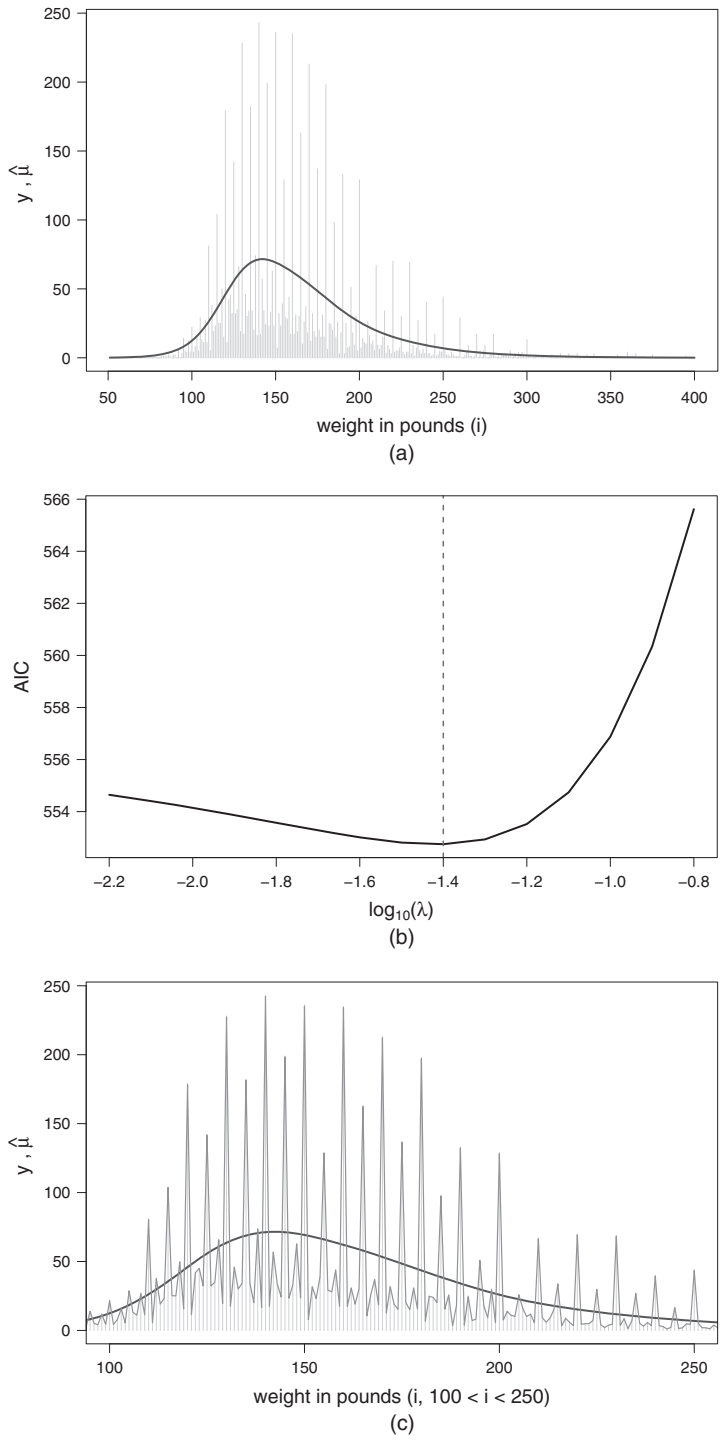

Fig. 10. (a) Illustration of subjective (———, $\hat{\mu}$, $\lambda = 10^6$) and (c) objective (only shown between 100 and 250 lb; ———, $\hat{\mu}$, $\lambda = 10^6$; ———, $\hat{\mu}$, AIC-based $\lambda = 10^{-1.4}$) smoothing of the self-reported weight data (▮): (b) AIC profile ($\hat{\lambda} = 0.04$) (data for females between age 40 and 50 years)

The explanation is simple: the model for the observed counts, say $\mathbf{y}$, is $y_i \sim \text{Pois}(\mu_i)$, with expected values in $\boldsymbol{\mu}$. If $\boldsymbol{\mu}$ is smooth, the AIC will indicate sufficiently strong smoothing to recover it as a smooth curve. But, if it is not smooth, the AIC will not allow a large λ. In our models for digit preference we put a non-smooth transfer mechanism on top of a smooth underlying density (the values in $\boldsymbol{\gamma}$). By proper modelling of the transfers we obtain a smooth result automatically.

References

Bolton-Smith, C., Woodward, M., Tunstall-Pedoe, H. and Morrison, C. (2000) Accuracy of the estimated prevalence of obesity from self reported height and weight in an adult Scottish population. *J. Epidem. Commty Hlth*, **54**, 143–148.

Bowring, A. L., Peeters, A., Freak-Poli, R., Lim, M. S., Gouillou, M. and Hellard, M. (2012) Measuring the accuracy of self-reported height and weight in a community-based sample of young people. *BMC Med. Res. Methodol.*, **12**, 1–8.

Camarda, C. G., Eilers, P. H. C. and Gampe, J. (2008) Modelling general patterns of digit preference. *Statist. Modllng*, **8**, 385–401.

Crawford, S. L., Johannes, C. B. and Stellato, R. K. (2002) Assessment of digit preference in self-reported year at menopause: choice of an appropriate reference distribution. *Am. J. Epidem.*, **156**, 676–683.

Currie, I. D., Durban, M. and Eilers, P. H. C. (2004) Smoothing and forecasting mortality rates. *Statist. Modllng*, **4**, 279–298.

Currie, I. D., Durban, M. and Eilers, P. H. C. (2006) Generalized linear array models with applications to multidimensional smoothing. *J. R. Statist. Soc.* B, **68**, 259–280.

Dekkers, J. C., van Wier, M. F., Hendriksen, I. J., Twisk, J. W. and van Mechelen, W. (2008) Accuracy of self-reported body weight, height and waist circumference in a Dutch overweight working population. *BMC Med. Res. Methodol.*, **8**, 1–13.

Efron, B. and Tibshirani, R. J. (1998) *An Introduction to the Bootstrap*. New York: Chapman and Hall.

Eilers, P. H. C. (2007) Ill-posed problems with counts, the composite link model and penalized likelihood. *Statist. Modllng*, **7**, 239–254.

Eilers, P. H. C. and Marx, B. D. (1996) Flexible smoothing with *B*-splines and penalties (with discussion). *Statist. Sci.*, **11**, 89–102.

Eilers, P. H. C., Marx, B. D. and Durbán, M. (2015) Twenty years of *P*-splines. *Statist. Ops Res. Trans.*, **39**, 149–186.

Emmerson, A. J. and Roberts, S. A. (2013) Rounding of birth weights in a neonatal intensive care unit over 20 years: an analysis of a large cohort study. *BMJ Open*, **3**, article e003650.

Lu, S., Su, J., Xiang, Q., Zhou, J. and Wu, M. (2016) Accuracy of self-reported height, weight, and waist circumference in a general adult Chinese population. *Popln Hlth Metr.*, **14**, 1–9.

National Health and Nutrition Examination Survey (2013) National Health and Nutrition Examination Survey (NHANES). US National Center for Health Statistics, Atlanta. (Available from `http://www.cdc.gov/nchs/nhanes.htm`.)

R Core Team (2016) *R: a Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Schlossmacher, E. (1973) An iterative technique for absolute deviations curve fitting. *J. Am. Statist. Ass.*, **68**, 857–865.

Thompson, R. and Baker, R. J. (1981) Composite link functions in generalized linear models. *Appl. Statist.*, **30**, 125–131.

Tibshirani, R. (1996) Regression shrinkage and selection via the lasso. *J. R. Statist. Soc.* B, **58**, 267–288.