Australian & New Zealand Journal of Statistics

Aust. N. Z. J. Stat. **55**(3), 2014, 321–342 doi: 10.1111/anzs.12043

ESTIMATION OF A SEMIPARAMETRIC RECURSIVE BIVARIATE PROBIT MODEL WITH NONPARAMETRIC MIXING

GIAMPIERO MARRA[1,*], GEORGIOS PAPAGEORGIOU[2] AND ROSALBA RADICE[3]

University College London, Imperial College London and University of London

Summary

We consider an extension of the recursive bivariate probit model for estimating the effect of a binary variable on a binary outcome in the presence of unobserved confounders, nonlinear covariate effects and overdispersion. Specifically, the model consists of a system of two binary outcomes with a binary endogenous regressor which includes smooth functions of covariates, hence allowing for flexible functional dependence of the responses on the continuous regressors, and arbitrary random intercepts to deal with overdispersion arising from correlated observations on clusters or from the omission of non-confounding covariates. We fit the model by maximizing a penalized likelihood using an Expectation-Maximisation algorithm. The issues of automatic multiple smoothing parameter selection and inference are also addressed. The empirical properties of the proposed algorithm are examined in a simulation study. The method is then illustrated using data from a survey on health, aging and wealth.

Key words: nonparametric maximum likelihood estimation; penalised regression spline; recursive bivariate probit model; unobserved confounding.

1. Introduction

Quantifying the effect of a predictor of interest (also referred to as treatment) on a particular response variable is a challenging task in observational studies. This is because it is often the case that confounders which are associated with both treatment and response are either unknown or not readily quantifiable (this problem is known in econometrics as endogeneity of the variable of interest). Moreover, covariate-response relationships can exhibit nonlinear patterns and observations may be overdispersed. In such a context, the use of standard estimators neglecting the aforementioned issues yields inconsistent estimates. In this article, we consider the case in which the researcher is interested in estimating the effect of a binary endogenous variable on a binary outcome in the presence of unobserved confounders, nonlinear covariate-response relationships and overdispersion

*Author to whom correspondance should be addressed.
[1]Department of Statistical Science, University College London, Gower Street, London, WC1E 6BT, UK e-mail: giampiero@stats.ucl.ac.uk.
[2]Department of Epidemiology and Biostatistics, Imperial College London, Norfolk Place, London, W2 1PG, UK.
[3]Department of Economics, Mathematics and Statistics, Birkbeck, University of London, Malet Street, London, WC1E 7HX, UK.
Acknowledgements. We would like to thank two reviewers for the detailed comments, which helped us improve the manuscript and clarify the main messages.
[The copyright line for this article was changed on 21 July 2014 after original online publication.]

resulting from either correlations among observations on the same clusters or from the omission of non-confounding covariates.

Instrumental variable techniques are widely used for isolating the effect of a given predictor in the presence of unobserved confounding (e.g. Wooldridge 2010; Marra & Radice 2011b and references therein), and are increasingly used in epidemiological and medical studies (e.g. Goldman *et al.* 2001 and references therein). In the context of binary responses, it is well known, from both theoretical and empirical results, that bivariate likelihood estimation methods are superior to conventional two-stage instrumental variable procedures (e.g. Bhattacharya *et al.* 2006; Wooldridge 2010). First introduced by Heckman (1978), the recursive bivariate probit model represents an effective way to estimate the effect a binary regressor has on a binary outcome in the presence of unobservables. The semiparametric version of Heckman's model is an important extension since undetected nonlinearity can have severe consequences on the estimation of covariate effects (e.g. Marra & Radice 2011a). Chib & Greenberg (2007) proposed two Bayesian fitting procedures for the class of instrumental variable models including the semiparametric recursive bivariate probit model. However, as the authors point out, very large sample sizes are required to obtain reasonable estimates of the binary treatment effect, hence undermining the utility of the method for practical modeling. Marra & Radice (2011a) considered the same model and introduced a penalized likelihood based procedure which permits reliable estimation of the model coefficients at reasonably small sample sizes.

The neglect of the possible presence of overdispersion may have a detrimental impact on the estimation of the effect of an endogenous variable. This issue is dealt with by generalising the method of Marra & Radice (2011a) to include random effects, which are generated by unknown densities. The usual parametric approach, which assumes that random effects are generated by a bivariate normal density (Greene 2012), is avoided here as restrictive. Consequences of parametric assumptions have been studied extensively within the class of generalised linear mixed models (GLMMs). Several authors have shown that misspecification of the random effects distribution can affect negatively the estimation of regression parameters; see for instance Neuhau *et al.* (1992), Heagerty & Kurland (2001), Chen *et al.* (2002), and Agresti *et al.* (2004). In addition, the assumed distribution is a very important factor for the prediction of the random effects themselves. In fact, the shape of the distribution of the empirical Bayes estimates tends to have features that are similar to the assumed random effects distribution, even if in reality assumed and true distributions are not close together (Verbeke & Lesaffre 1996; Papageorgiou & Hinde 2012). With a nonparametric approach such pitfalls are avoided. The results of Laird (1978) and Lindsay (1983) have shown that the nonparametric maximum likelihood estimate of a mixing distribution is a discrete distribution. General fitting algorithms have been provided by Laird (1978), Lindsay (1983), Follmann & Lambert (1989) and Lesperance & Kalbfleisch (1992).

The proposed model is fitted by maximizing a penalised likelihood using an Expectation-Maximisation algorithm, where the issues of automatic multiple smoothing parameter selection and inference are also addressed. The empirical properties of the proposed algorithm are examined in a simulation study. The method is then illustrated using data from a survey on health, aging and wealth. Specifically, the aim is to estimate the effect of private health insurance on private medical care utilization. In such data, endogeneity is likely to arise because insurance coverage is not randomly assigned but rather is the result

of supply and demand. Moreover, estimation of the effect of private health insurance on private medical care utilization may be adversely affected by overdispersion resulting from the heterogeneity present in the observations due to unobserved covariates related to either the response or the treatment variable. Buchmueller *et al.* (2005) provide an excellent review of these issues, which, if neglected, can lead to a biased estimate of the relationship of interest.

2. Model specification

The recursive bivariate probit model consists of a reduced form or treatment equation for the potentially endogenous binary variable and a second structural form or outcome equation for the binary response variable. The mixed effects semiparametric version of this model takes the form

$$\begin{aligned} y_{1ij}^* &= u_{1i} + \mathbf{x}_{1ij}^\top \boldsymbol{\theta}_1 + \sum_{k_1=1}^{K_1} s_{1k_1}(z_{1k_1ij}) + \varepsilon_{1ij} \\ y_{2ij}^* &= u_{2i} + \vartheta y_{1ij} + \mathbf{x}_{2ij}^\top \boldsymbol{\theta}_2 + \sum_{k_2=1}^{K_2} s_{2k_2}(z_{2k_2ij}) + \varepsilon_{2ij} \end{aligned} \quad , i = 1, \ldots, m,\ j = 1, \ldots, n_i, \qquad (1)$$

where m denotes the number of clusters, n_i is the number of observations within the ith cluster, and y_{vij}^* is a latent continuous variable which determines its observable counterpart y_{vij} through the rule $\mathbf{1}(y_{vij}^* > 0)$, for $v = 1, 2$, where $\mathbf{1}(\cdot)$ is the indicator function; ϑ is the coefficient of the endogenous binary variable y_{1ij}; vector $\mathbf{x}_{1ij}$ contains P_1 parametric model components (such as dummy and categorical observed confounders, but not intercepts as we do not impose a zero mean on the random effects), with corresponding parameter vector $\boldsymbol{\theta}_1$. The s_{1k_1} are unknown smooth functions of the K_1 continuous observed confounders z_{1k_1ij}. Varying coefficients models can be obtained by multiplying a smooth term by some predictor (Hastie & Tibshirani 1993). Smooth functions of two covariates such as $s_{11,12}(z_{11ij}, z_{12ij})$ can also be implemented (e.g. Wood 2006, pp. 154–167). Similarly, $\mathbf{x}_{2ij}$ is a vector of dimension P_2 with associated parameter vector $\boldsymbol{\theta}_2$, the s_{2k_2} are unknown smooth terms of the K_2 continuous observed confounders z_{2k_2ij}. For identification purposes, the smooth functions are subject to the centering constraint $\sum_{ij} s_k(z_{kij}) = 0$ for all terms (Wood 2006 pp. 167–168). The pair of random effects (u_{1i}, u_{2i}) is cluster specific, hence it induces correlation among multiple observations on the same cluster or can be used to handle overdispersion in case of independent observations, i.e. $n_i = 1$ for all i. For instance, a large value of u_{1i} will tend to make y_{1ij}^* large for all n_i observations within the ith cluster. Similar comments hold for u_{2i}. As in Chib & Greenberg (2007) and Marra & Radice (2011a), we make the assumption that unobserved confounders have a linear impact on the response. That is, the error terms $(\varepsilon_{1ij}, \varepsilon_{2ij})$ are assumed to follow the bivariate distribution

$$\begin{pmatrix} \varepsilon_{1ij} \\ \varepsilon_{2ij} \end{pmatrix} \overset{iid}{\sim} \mathrm{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho_\varepsilon \\ \rho_\varepsilon & 1 \end{bmatrix} \right),$$

where ρ_ε is a correlation coefficient and the error variances are set to 1 as the model parameters can only be identified up to a scale coefficient (e.g. Greene 2012). Parameter ρ_ε accounts for the correlation between the responses not accounted for by the pair

(u_{1i}, u_{2i}). As in Greene (2012), u_{1i} and u_{2i} can be correlated. Further, it is assumed that the error terms and random effects are independent.

The recursive structure of (1) follows from the condition of logical consistency. It states that only one observed endogenous variable is allowed on the right-hand side of system (1). This is because the probabilities for the different possible value combinations of the two binary variables have to sum to one (e.g. Maddala 1983, p. 118). To identify the model parameters, it is typically assumed that the exclusion restriction on the exogenous variables holds (e.g. Maddala 1983, p. 122). That is, the exogenous covariates in the first equation of (1) should contain at least one regressor not included in the second equation. Such covariates are regarded as instrumental variables which induce variation in the treatment, do not directly affect the outcome, and are independent of the error terms given the covariates (e.g. Chib & Greenberg 2007). However, under correct model specification, this restriction may not be strictly necessary as pointed out by Wilde (2000) and Marra & Radice (2011a).

The smooth functions are represented using regression splines. The key idea is to approximate a generic function $s_k(z_{kij})$ by a linear combination of known spline basis functions, $b_{kq}(z_{kij})$, and regression parameters, β_{kq},

$$s_k(z_{kij}) = \sum_{q=1}^{Q_k} \beta_{kq} b_{kq}(z_{kij}) = \mathbf{B}_k(z_{kij})^\top \boldsymbol{\beta}_k,$$

where Q_k is the number of bases (hence regression coefficients) used to represent s_k, $\mathbf{B}_k(z_{kij})$ is a vector containing Q_k basis functions evaluated at observation z_{kij}, i.e. $\mathbf{B}_k(z_{kij}) = \{b_{k1}(z_{kij}), b_{k2}(z_{kij}), \ldots, b_{kQ_k}(z_{kij})\}^\top$, and $\boldsymbol{\beta}_k$ is the corresponding parameter vector. Basis functions should be chosen to have convenient mathematical properties and good numerical stability. Many choices are possible within the framework adopted in this article. These include B-splines, cubic and thin plate regression splines (see, e.g. Ruppert *et al.* 2003; Wood 2006 for a more detailed introduction); we opt for the latter. Based on the above regression spline representation, model (1) is written as

$$\begin{aligned} y^*_{1ij} &= u_{1i} + \mathbf{x}^\top_{1ij}\boldsymbol{\theta}_1 + \mathbf{B}^\top_{1ij}\boldsymbol{\beta}_1 + \varepsilon_{1ij} = \eta_{1ij} + \varepsilon_{1ij} \\ y^*_{2ij} &= u_{2i} + \vartheta y_{1ij} + \mathbf{x}^\top_{2ij}\boldsymbol{\theta}_2 + \mathbf{B}^\top_{2ij}\boldsymbol{\beta}_2 + \varepsilon_{2ij} = \eta_{2ij} + \varepsilon_{2ij}, \end{aligned}$$

where $\mathbf{B}^\top_{vij} = \{\mathbf{B}_{v1}(z_{v1ij})^\top, \ldots, \mathbf{B}_{vK_v}(z_{vK_vij})^\top\}$, $\boldsymbol{\beta}^\top_v = (\boldsymbol{\beta}^\top_{v1}, \ldots, \boldsymbol{\beta}^\top_{vK_v})$, for $v = 1, 2$, and the linear predictors, η_{vij}, have the obvious definitions.

In the current context, the effect of y_{1ij} is of primary interest. This is typically calculated using the average treatment effect (ATE). Given estimates for the random effects, parametric and smooth function components, the ATE can be estimated as follows

$$\frac{1}{\sum_{i=1}^m n_i} \sum_{i=1}^m \sum_{j=1}^{n_i} \frac{\Phi_2\left(\hat{\eta}_{2ij}^{(y_{1ij}=1)}, \hat{\eta}_{1ij}; \rho_\varepsilon\right)}{\Phi(\hat{\eta}_{1ij})} - \frac{\Phi_2\left(\hat{\eta}_{2ij}^{(y_{1ij}=0)}, -\hat{\eta}_{1ij}; -\rho_\varepsilon\right)}{1 - \Phi(\hat{\eta}_{1ij})},$$

where Φ and Φ_2 are the distribution functions of a standardized univariate normal and a standardized bivariate normal with correlation ρ_ε, and $\hat{\eta}_{2ij}^{(y_{1ij}=r)}$ indicates the linear predictor evaluated at r equal to 1 or 0. Coefficient ρ_ε is also of interest as it can be used to ascertain the presence of unobserved confounding (endogeneity). It can be interpreted as the

correlation between the unobserved confounders in the two equations (e.g. Monfardini & Radice 2008). If $\rho_\varepsilon = 0$ then ε_{1ij} and ε_{2ij} are uncorrelated and hence there is not a problem of endogeneity. Because model (1) can capture, and hence separate, two different sources of variability (represented by ε_{vij} and u_{vi}), estimation of ρ_ε will be done more reliably by model (1) than by a model which does not account for overdispersion (e.g. Greene 2012).

3. Methods

3.1. Estimation approach

Recall that the error terms $(\varepsilon_{1ij}, \varepsilon_{2ij})$ are assumed to follow a bivariate normal distribution. Define the parameter vector $\boldsymbol{\delta} = (\boldsymbol{\theta}_1^\top, \boldsymbol{\theta}_2^\top, \boldsymbol{\beta}_1^\top, \boldsymbol{\beta}_2^\top, \vartheta, \rho_\varepsilon, \boldsymbol{\nu}^\top)^\top$, and pairs of random effects $\boldsymbol{u}_i = (u_{1i}, u_{2i})^\top$. Vector $\boldsymbol{\nu}$ contains the parameters pertaining to the random effects distribution (see next section). In the current context, the data identify four possible events, $(y_{1ij} = e_1, y_{2ij} = e_2)$ with $e_v \in \{0, 1\}$ for $v = 1, 2$, with the following conditional probabilities

$$p^u_{11ij} \equiv \Pr(y_{1ij} = 1, y_{2ij} = 1|\boldsymbol{\delta}; \boldsymbol{u}_i) = \Phi_2(\eta_{1ij}, \eta_{2ij}; \rho_\varepsilon), \tag{2}$$

$$p^u_{10ij} \equiv \Pr(y_{1ij} = 1, y_{2ij} = 0|\boldsymbol{\delta}; \boldsymbol{u}_i) = \Phi(\eta_{1ij}) - p^u_{11ij}, \tag{3}$$

$$p^u_{01ij} \equiv \Pr(y_{1ij} = 0, y_{2ij} = 1|\boldsymbol{\delta}; \boldsymbol{u}_i) = \Phi(\eta_{2ij}) - p^u_{11ij}, \tag{4}$$

$$p^u_{00ij} \equiv \Pr(y_{1ij} = 0, y_{2ij} = 0|\boldsymbol{\delta}; \boldsymbol{u}_i) = 1 - p^u_{11ij} - p^u_{10ij} - p^u_{01ij}. \tag{5}$$

The penalised log-likelihood function of the observed data $\boldsymbol{y} = \{\boldsymbol{y}_i, i = 1, \ldots, m\}$, where $\boldsymbol{y}_i = \{\boldsymbol{y}_{ij} = (y_{1ij}, y_{2ij})^\top : j = 1, \ldots, n_i\}$, is

$$\ell_p(\boldsymbol{\delta}|\boldsymbol{y}) = \ell(\boldsymbol{\delta}|\boldsymbol{y}) - \frac{1}{2}\boldsymbol{\beta}^\top \mathbf{S}_{\boldsymbol{\lambda}}\boldsymbol{\beta}, \tag{6}$$

where $\ell(\boldsymbol{\delta}|\boldsymbol{y}) = \log\{\prod_{i=1}^m f(\boldsymbol{y}_i|\boldsymbol{\delta})\}$, $\boldsymbol{\beta}^\top = (\boldsymbol{\beta}_1^\top, \boldsymbol{\beta}_2^\top)$, $\mathbf{S}_{\boldsymbol{\lambda}} = \sum_{v=1}^2 \sum_{k_v=1}^{K_v} \lambda_{vk_v}\mathbf{S}_{vk_v}$, and the $\mathbf{S}_{vk_v}$ are positive semi-definite known square matrices measuring the (second-order, here) roughness of the smooth terms in the model, that is $\boldsymbol{\beta}^\top\left(\sum_{v=1}^2 \sum_{k_v=1}^{K_v} \lambda_{vk_v}\mathbf{S}_{vk_v}\right)\boldsymbol{\beta} = \sum_{v=1}^2 \sum_{k_v=1}^{K_v} \lambda_{vk_v} \int s''_{vk_v}(z_{vk_v})^2 dz_{vk_v}$. The λ_{vk_v} are smoothing parameters controlling the trade-off between fit and smoothness.

Note that because of the presence of smooth components in the model, unpenalised estimation would yield exceedingly wiggly curve estimates which can have a detrimental impact on the estimation of the ATE (Marra & Radice 2011a). This is why the log-likelihood is augmented by a penalty term. In addition, because ρ_ε is bounded in $[-1, 1]$, we use the common transform for correlation $\rho_\varepsilon^+ = \tanh^{-1}(\rho_\varepsilon) = 0.5\log\{(1 + \rho_\varepsilon)/(1 - \rho_\varepsilon)\}$, so that $[-1, 1]$ is mapped to the real line.

3.1.1. EM penalised log-likelihood maximisation

We make no assumptions about the form of the density that gives rise to the model's random effects $\boldsymbol{u}_i$. The nonparametric maximum likelihood estimate of a mixing distribution

is discrete (Laird 1978; Lindsay 1983) and thus the density of $\boldsymbol{u}_i$ can be represented by F bivariate mass points, $\boldsymbol{m}_1 = (m_{11}, m_{12}), \ldots, \boldsymbol{m}_F = (m_{F1}, m_{F2})$, with corresponding probabilities, $\pi_1, \ldots, \pi_F$, where $\sum_{l=1}^{F} \pi_l = 1$. Hence the parameter vector $\boldsymbol{v}$, first introduced in Section 3.1, consists of $\boldsymbol{v} = (\boldsymbol{m}_1^\top, \ldots, \boldsymbol{m}_F^\top, \pi_1, \ldots, \pi_{F-1})^\top$. We will treat F as a tuning constant.

An EM algorithm (Dempster *et al.* 1977) is employed for maximising (6). We consider $(\boldsymbol{y}, \boldsymbol{u}) = \{(\boldsymbol{y}_i, \boldsymbol{u}_i) : i = 1, \ldots, m\}$ to be the complete data and indirectly maximise $\ell_p(\boldsymbol{\delta}|\boldsymbol{y})$ by iteratively maximising the expectation of the penalized log-likelihood of the complete data, where the expectation is taken with respect to the conditional distribution of the missing given the observed data

$$\begin{aligned}
\mathcal{Q}_p(\boldsymbol{\delta}|\boldsymbol{\delta}^{(a)}) &= \mathcal{Q}(\boldsymbol{\delta}|\boldsymbol{\delta}^{(a)}) - \frac{1}{2}\boldsymbol{\beta}^\top \mathbf{S}_\lambda \boldsymbol{\beta} = \mathrm{E}\{\log f(\boldsymbol{y}, \boldsymbol{u})|\boldsymbol{y}, \boldsymbol{\delta}^{(a)}\} - \frac{1}{2}\boldsymbol{\beta}^\top \mathbf{S}_\lambda \boldsymbol{\beta} \\
&= \sum_{i=1}^{m} [\mathrm{E}\{\log f(\boldsymbol{y}_i|\boldsymbol{u}_i)|\boldsymbol{y}_i, \boldsymbol{\delta}^{(a)}\} + \mathrm{E}\{\log \Pr(\boldsymbol{u}_i)|\boldsymbol{y}_i, \boldsymbol{\delta}^{(a)}\}] - \frac{1}{2}\boldsymbol{\beta}^\top \mathbf{S}_\lambda \boldsymbol{\beta} \\
&= \sum_{i=1}^{m} \sum_{j=1}^{n_i} \mathrm{E}\{\log f(\boldsymbol{y}_{ij}|\boldsymbol{u}_i)|\boldsymbol{y}_i, \boldsymbol{\delta}^{(a)}\} + \sum_{i=1}^{m} \mathrm{E}\{\log \Pr(\boldsymbol{u}_i)|\boldsymbol{y}_i, \boldsymbol{\delta}^{(a)}\} - \frac{1}{2}\boldsymbol{\beta}^\top \mathbf{S}_\lambda \boldsymbol{\beta},
\end{aligned}$$

where $\boldsymbol{\delta}^{(a)}$ is the current value of the parameter vector. Let $\boldsymbol{\delta}^{(a+1)} = \boldsymbol{\phi}(\boldsymbol{\delta}^{(a)})$ be the parameter vector that maximises $\mathcal{Q}_p(\boldsymbol{\delta}|\boldsymbol{\delta}^{(a)})$. Under regularity conditions, at convergence $\hat{\boldsymbol{\delta}} = \boldsymbol{\phi}(\hat{\boldsymbol{\delta}})$ maximises both the complete and the observed data log-likelihoods.

Conditionally on the data and current parameter estimates, the distribution of $\boldsymbol{u}_i$ is discrete with points $\boldsymbol{m}_l^{(a)}, l = 1, \ldots, F$, and probability masses given by

$$w_{il}^{(a)} = \Pr(\boldsymbol{u}_i = \boldsymbol{m}_l^{(a)}|\boldsymbol{y}_i, \boldsymbol{\delta}^{(a)}) = \frac{f(\boldsymbol{y}_i|\boldsymbol{\delta}^{(a)}; \boldsymbol{u}_i = \boldsymbol{m}_l^{(a)})\pi_l^{(a)}}{\sum_{l=1}^{F} f(\boldsymbol{y}_i|\boldsymbol{\delta}^{(a)}; \boldsymbol{u}_i = \boldsymbol{m}_l^{(a)})\pi_l^{(a)}}.$$

Given the $w_{il}^{(a)}$, we have the following expression for the penalised complete data log-likelihood

$$\begin{aligned}
\mathcal{Q}_p(\boldsymbol{\delta}|\boldsymbol{\delta}^{(a)}) = &\sum_{i=1}^{m} \sum_{j=1}^{n_i} \sum_{l=1}^{F} w_{il}^{(a)} \log \Pr(y_{1ij}, y_{2ij}|\boldsymbol{\delta}^{(a)}; \boldsymbol{u}_i = \boldsymbol{m}_l^{(a)}) \\
&+ \sum_{i=1}^{m} \sum_{l=1}^{F} w_{il}^{(a)} \log \pi_l^{(a)} - \frac{1}{2}\boldsymbol{\beta}^\top \mathbf{S}_\lambda \boldsymbol{\beta}, \qquad (7)
\end{aligned}$$

where the $\Pr(\cdot|\cdot)$ are given in (2)–(5).

Note that in (7) the parameter vector $\boldsymbol{\delta}$ separates into two independent subvectors, namely the vector $\boldsymbol{\delta}_1 \equiv (\boldsymbol{\theta}_1^\top, \boldsymbol{\theta}_2^\top, \boldsymbol{\beta}_1^\top, \boldsymbol{\beta}_2^\top, \vartheta, \rho_\varepsilon^+, \boldsymbol{m}_1^\top, \ldots, \boldsymbol{m}_F^\top)^\top$ that appears in the triple sum and penalty term and the vector $\boldsymbol{\delta}_2 \equiv (\pi_1, \ldots, \pi_{F-1})^\top$ that appears in the double sum. Consequently, maximisation of $\mathcal{Q}_p$ is achieved in two steps. Firstly, the triple sum has summands that, ignoring fixed $w_{il}^{(a)}$, are exactly the same as those that would have been obtained by assuming a model without random effects. It follows that the triple sum with penalty term can be maximized using the algorithm presented in Marra & Radice (2011a). Secondly, the double sum is used to update the masses of the random effects distribution resulting in closed form formulas $\pi_l^{(a+1)} \propto \sum_{i=1}^{m} w_{il}^{(a)}$.

3.1.2. Smoothing parameter selection

If the model has more than two or three smooth terms, then it becomes crucial to estimate the smoothing parameters using an automatic, quick and reliable procedure. There are several techniques for automatic multiple smoothing parameter selection for univariate models (see Ruppert *et al.* 2003; Wood 2006 for a detailed overview). These include the performance-oriented iteration method first introduced by Gu (1992) which consists of applying the generalized cross validation or unbiased risk estimator (UBRE, Craven & Wahba 1979) to each working linear model of the penalized iteratively re-weighted least squares scheme used to fit the model. In what follows, we employ an adaptation of Gu's approach. Also, we suppress the superscript (a) to avoid clutter.

Given values for λ_{vk_v} and $\boldsymbol{\delta}_2$, an estimate for $\boldsymbol{\delta}_1$ can be obtained by minimisation of

$$\|\mathbf{z}^* - \mathbf{X}^*\boldsymbol{\delta}_1\|^2 + \boldsymbol{\delta}_1^\top \mathbf{S}_{\boldsymbol{\lambda}}^*\boldsymbol{\delta}_1 \quad \text{w.r.t.}\ , \quad \boldsymbol{\delta}_1, \tag{8}$$

where

$$\mathbf{z}^* = \sum_{l=1}^{F} \boldsymbol{w}_l\sqrt{\mathbf{W}}_l\{\mathbf{X}_l\boldsymbol{\delta}_1 + \mathbf{W}_l^{-1}\mathbf{d}_l\} \text{ and } \mathbf{X}^* = \sum_{l=1}^{F} \boldsymbol{w}_l\sqrt{\mathbf{W}}_l\mathbf{X}_l,$$

$\mathbf{S}_{\boldsymbol{\lambda}}^*$ is an overall blockdiagonal penalty matrix made up of the $\lambda_{vk_v}\mathbf{S}_{vk_v}$ and zero vectors corresponding to the model parameters which are not penalised, and $\boldsymbol{w}_l$ is a vector containing the masses as defined in the previous section. Assuming, for simplicity and without loss of generality, that $n_i = 1$ and $Q_{vk_v} = Q$ for each vk_v so that the total number of observations is m, $\mathbf{d}_{il}$ is a 3-dimensional vector given by $\{\partial\mathcal{Q}_{il}/\partial\eta_{1il}, \partial\mathcal{Q}_{il}/\partial\eta_{2il}, \partial\mathcal{Q}_{il}/\partial\eta_{3il}\}^\top$, $\eta_{3il} = \rho_\varepsilon^+$, $\mathcal{Q}_{il} = w_{il}\log\Pr(y_{1i}, y_{2i}|\boldsymbol{\delta}; \boldsymbol{u}_i = \boldsymbol{m}_l)$, $\mathbf{W}_{il}$ is a 3×3 matrix with (r,h)th element

$$(\mathbf{W}_{il})_{rh} = -\mathrm{E}\left[\frac{\partial^2\mathcal{Q}_{il}}{\partial\eta_{ril}\partial\eta_{hil}}\right], r, h = 1, \ldots, 3, \tag{9}$$

$\mathbf{X}_{il}$ is a $3 \times \{(F + P_1 + K_1 \times Q) + (F + P_2 + K_2 \times Q) + 1\}$ block diagonal matrix, i.e. $\mathbf{X}_{il} = \mathrm{diag}\{(\mathbf{c}_{1il}^\top, \mathbf{x}_{1i}^\top, \mathbf{B}_{1i}^\top), (\mathbf{c}_{2il}^\top, \mathbf{x}_{2i}^\top, \mathbf{B}_{2i}^\top), 1\}$, where each of the vectors $\mathbf{c}_{1il}$ and $\mathbf{c}_{2il}$ contain F zero elements but the lth which is set to 1, and the definitions of the linear predictors in (9) follow from the definition of $\mathbf{X}_{il}$. The square root and inverse of $\mathbf{W}_l$ are obtained via eigendecomposition.

The smoothing parameter vector $\boldsymbol{\lambda}$ is selected so that the estimated smooth terms are as close as possible to the true functions (Craven & Wahba 1979). Given an estimate for $\boldsymbol{\delta}_1$, multiple smoothing parameter estimation for problem (8) can be achieved by minimization of the approximate UBRE score

$$\mathcal{V}_u^w(\boldsymbol{\lambda}) = \frac{1}{n_*}\|\mathbf{z}^* - \mathbf{X}^*\boldsymbol{\delta}_1\|^2 - 1 + \frac{2}{n_*}\mathrm{tr}(\mathbf{A}_{\boldsymbol{\lambda}}), \tag{10}$$

where the working linear model quantities are calculated using the parameter estimates from the optimisation step mentioned in Section 3.1.1, $n_* = 3m$, $\mathbf{A}_{\boldsymbol{\lambda}} = \mathbf{X}^*(\mathbf{X}^{*\mathrm{T}}\mathbf{X}^* + \mathbf{S}_{\boldsymbol{\lambda}}^*)^{-1}\mathbf{X}^*$ is the hat matrix, and $\mathrm{tr}(\mathbf{A}_{\boldsymbol{\lambda}})$ the estimated degrees of freedom of the penalised model. For each working linear model of iteration (8), $\mathcal{V}_u^w(\boldsymbol{\lambda})$ is minimized with respect to $\boldsymbol{\lambda}$. In practice, this is implemented employing the approach by Wood

(2004), which is based on the Newton-Raphson method. In evaluating score (10) and their derivatives, efficiency and stability are achieved using a combination of pivoted QR and singular value decompositions (see Wood 2004 for full details). Note that because each of the $\mathbf{W}_l$ is a non-diagonal matrix of dimension $n_* \times n_*$, computation can quickly become prohibitive, hence its sparse structure is exploited in implementation.

3.2. Inference

Inference in penalised models is complicated by the presence of smoothing penalties which undermines the usefulness of classic frequentist results for practical modelling. Solutions to this problem have been introduced in the literature (see, e.g., Gu 2002; Wood 2006 for an overview). Here we show how to construct pointwise confidence intervals for the terms of a mixed effects semi-parametric bivariate model by adapting the well known Bayesian confidence intervals, originally proposed by Wahba (1983) and Silverman (1985). An appealing feature of these intervals is that they have close to nominal 'across-the-function' *frequentist* coverage probabilities (Marra & Wood 2012). This is because the Wahba/Silverman type intervals include both a bias and variance component. Moreover, their empirical performance has little sensitivity to the neglect of smoothing parameter uncertainty. For a generic term $s_k(z_{ki})$ intervals can be constructed by seeking constants C_{ki} and A, such that

$$\text{ACP} = \frac{1}{n}\text{E}\left\{\sum_i \mathbf{1}(|\hat{s}_k(z_{ki}) - s_k(z_{ki})| \leq q_{\alpha/2}A/\sqrt{C_{ki}})\right\} = 1 - \alpha, \tag{11}$$

where 'ACP' denotes 'Average Coverage Probability', α is a constant between 0 and 1, and $q_{\alpha/2}$ is the $\alpha/2$ critical point from a standard normal distribution. Defining $b_k(z_k) = \text{E}\{\hat{s}_k(z_k)\} - s_k(z_k)$ and $v_k(z_k) = \hat{s}_k(z_k) - \text{E}\{\hat{s}_k(z_k)\}$, so that $\hat{s}_k - s_k = b_k + v_k$, and I to be a random variable uniformly distributed on $\{1, 2, \ldots, n\}$, we have that $\text{ACP} = \Pr(|B_k + V_k| \leq q_{\alpha/2}A)$, where $B_k = \sqrt{C_{kI}}b(z_{kI})$ and $V_k = \sqrt{C_{kI}}v(z_{kI})$. It is then necessary to find the distribution of $B_k + V_k$ and values for C_{ki} and A so that requirement (11) is met. As shown in Marra & Wood (2012), in the context of non-Gaussian response models involving several smooth components, such a requirement is approximately met when confidence intervals for the $\hat{s}_k(z_{ki})$ are constructed using the distribution

$$\boldsymbol{\delta}|\boldsymbol{y} \dot{\sim} \text{N}(\hat{\boldsymbol{\delta}}, \mathbf{V}_{\boldsymbol{\delta}}), \tag{12}$$

where, in our context, $\boldsymbol{y}$ refers to the binary response vectors, $\hat{\boldsymbol{\delta}}$ is an estimate of $\boldsymbol{\delta}$, and $\mathbf{V}_{\boldsymbol{\delta}} = (\mathcal{I} + \mathbf{S}^*_{\boldsymbol{\lambda}})^{-1}$ where $\mathcal{I}$ is the information matrix. Specifically,

$$\hat{s}_k(z_{ki}) \dot{\sim} \text{N}(s_k(z_{ki}), \mathbf{B}_k(z_{ki})^\top \mathbf{V}_{\boldsymbol{\delta}_k} \mathbf{B}_k(z_{ki})),$$

where $\mathbf{V}_{\boldsymbol{\delta}_k}$ and $\mathbf{B}_k(z_{ki})$ are the submatrix of $\mathbf{V}_{\boldsymbol{\delta}}$ and the basis functions corresponding to the regression spline parameters associated with $s_k(z_{ki})$. In addition, intervals for non-linear functions of the model parameters, such as the ATE, can be conveniently obtained by simulation from (12).

In practice, $\mathcal{I}$ can be replaced by its observed version $\mathcal{J}$. In the present context, however, $\mathcal{J}$ cannot be obtained as a byproduct of the estimation procedure and of the second order derivatives of $\mathcal{Q}$, in (7), used therein. Second derivatives of the log-likelihood

of the 'complete' data $(\boldsymbol{y}, \boldsymbol{u})$ would overestimate the information about the model parameters in the sample. Ultimately, this is attributed to treating the weights w_{il} that appear in $\mathcal{Q}$ as fixed. We therefore find the observed information using the method of Louis (1982), by which the observed information matrix is expressed as

$$\mathcal{J} = -\frac{\partial^2 \mathcal{Q}}{\partial \boldsymbol{\delta} \partial \boldsymbol{\delta}^\top} - \sum_{i=1}^{m} \mathrm{Var}(\boldsymbol{\psi}_i),$$

which makes clear how the second derivative of the complete data likelihood $\mathcal{Q}$ is adjusted for the unobserved data $\boldsymbol{u}$. Details about the approach, including the definition of $\boldsymbol{\psi}_i$, are provided in Appendix A. It is important to stress that there is no contradiction in fitting the model using the method of Section 3.1 and then constructing intervals following a Bayesian result, and such an approach has been employed many times in the literature (e.g. Gu 2002; Wood 2006 and references therein).

3.3. Algorithm

As indicated previously, we treat F, the number of mass points of the nonparametric mixing distribution, as a tuning parameter. It is common practice to find its value as the one that minimizes Akaike's information criterion. This, for a given value of F, takes the following form: $\mathrm{AIC}(F) = -2\ell(\hat{\boldsymbol{\delta}}|\boldsymbol{y}) + 2\mathrm{dim}(\hat{\boldsymbol{\delta}}|F)$, where $\mathrm{dim}(\hat{\boldsymbol{\delta}}|F)$ denotes the effective dimension of $\hat{\boldsymbol{\delta}}$ for a fixed value of F, and the log-likelihood is obtained from Equation (6), in which we express $f(\boldsymbol{y}_i|\boldsymbol{\delta}) = \sum_{l=1}^{F} f(\boldsymbol{y}_i|\boldsymbol{\delta}; \boldsymbol{u}_i = \boldsymbol{m}_l)\Pr(\boldsymbol{u}_i = \boldsymbol{m}_l|\boldsymbol{\delta})$, as also shown in Appendix A.

Having fixed a value of F, we need to choose starting values, $\boldsymbol{\delta}^0$, for the model parameters. Starting values for $(\boldsymbol{\theta}_1^\top, \boldsymbol{\theta}_2^\top, \boldsymbol{\beta}_1^\top, \boldsymbol{\beta}_2^\top, \vartheta, \rho_\varepsilon^+)^\top$ are chosen as the maximum likelihood estimates of the corresponding fixed effects model, i.e. the model assuming $F = 1$, fitted using the method of Marra & Radice (2011a). Starting values for the masses $\pi_k, k = 1, \ldots, F$, are all set to $1/F$, while starting values for the mass points, $(\boldsymbol{m}_1^\top, \ldots, \boldsymbol{m}_F^\top)^\top$, as set to a multiple (here, square root of two) of the Gauss-Hermite quadrature nodes.

Given F and $\boldsymbol{\delta}^0$, parameter estimates are found using an iterative algorithm. Iteration $(a + 1)$ consists of finding the maximizer of $\mathcal{Q}_p(\boldsymbol{\delta}|\boldsymbol{\delta}^{(a)})$ using the algorithm described after (7). For a given estimate of $\boldsymbol{\delta}$, smoothing parameter selection is achieved by minimisation of (10), as described in Section 3.1.2. The two main steps, one for $\boldsymbol{\delta}$ the other for $\boldsymbol{\lambda}$, are iterated until convergence. The rule that we follow for stopping the iterative algorithm is that the maximum absolute change in the parameter estimates from successive iterations is less than $\epsilon = 10^{-6}$.

At convergence, we calculate log-likelihood and $\mathrm{AIC}(F)$ to guide model choice, standard errors of the estimates by inverting the observed information obtained as described in the previous section, and random effects predictions using (14) as these are needed for estimating the ATE.

4. Simulation study

To gain insight into the empirical effectiveness of the proposed method, a Monte Carlo simulation study was conducted. All computations were performed in the

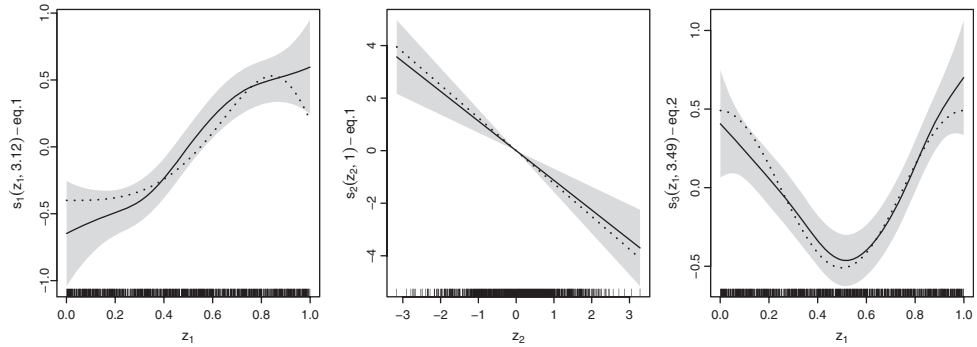

Figure 1. Test functions used in the simulation study (dotted lines) and a realization of estimated smooths (black lines) with corresponding 95% pointwise confidence intervals (shaded regions). The smooth estimates were obtained applying the proposed method on data from a typical sample with $n_i = 1$ and $m = 2000$, and bivariate gamma random variates for the random components. The estimates are on the scale of the respective linear predictors. Due to identifiability constraints, the curves centered around zero. The numbers in brackets in the y-axis captions are the estimated degrees of freedom of the smooth curves, while the rug plot, at the bottom of each graph, shows the observed covariate values.

`R` environment (R Development Core Team 2013) using the package `SemiParBIVProbit` (Marra & Radice 2013).

4.1. Design and model fitting details

The sampling experiments were based on the model

$$\begin{aligned} y^*_{1ij} &= u_{1i} - 0.75x_{1ij} + s_1(z_{1ij}) + s_2(z_{2ij}) + \varepsilon_{1ij} \\ y^*_{2ij} &= u_{2i} - 1.5y_{1ij} + x_{1ij} + s_3(z_{1ij}) + \varepsilon_{2ij}, \end{aligned} \tag{13}$$

where the binary outcomes y_{1ij} and y_{2ij} were determined as described in Section 2. The test functions used were $s_1(z_{1ij}) = 0.6\{z^3_{1ij} + \sin(\pi z^3_{1ij})\}$, $s_2(z_{2ij}) = -1.25z_{2ij}$ and $s_3(z_{1ij}) = 0.5\cos(2\pi z_{1ij})$ (see Fig. 1). Covariates x_{1ij}, z_{1ij} and z_{2ij} were generated as binary, uniform and normal correlated predictors, respectively. This was achieved by drawing standardised multivariate normal random variables with correlation 0.5 (using `rmvnorm()` in the package `mvtnorm`) and then transforming the first two of them with `round()` and `pnorm()` (e.g. Marra & Radice 2011a). Bivariate normal errors with zero means, standard deviations equal to one, and correlations $\rho_\varepsilon = \pm(0.1, 0.5, 0.9)$ were considered. Sample sizes were set to 2000 and 6000 in the following two ways. In the first case, n_i was set to a randomly chosen number between 9 and 11 and m was set to 200 and 600. In the second case, $n_i = 1$ and m was set to 2000 and 6000. The pairs of random effects (u_{1i}, u_{2i}) were generated according to three scenarios: bivariate normal variates with mean vector (0,2), standard deviations $\sigma_1 = \sigma_2 = 0.5$ and correlation $\rho_u = 0.5$; mixture of two equally weighted bivariate normals with mean vectors $(-2, -2)$ and $(2, 2)$, and with the remaining parameters as above; bivariate gamma variates with shape and scale parameters equal to 0.5. This last was achieved via a normal copula with $\rho_u = 0.5$ using `mvdc()` in the package `copula`. Each scenario was replicated 250 times and the quantities of interest, estimated ATE and ρ_ε (see final paragraph of Section 2), recorded.

The smooth components of continuous covariates in the model were represented using penalized thin plate regression splines with basis dimensions equal to 10 and penalties based on second-order derivatives (Wood 2006, pp. 154–160). The spline basis representation used here is a low rank eigen-approximation version of the full rank version introduced by Duchon (1977). It represents a general solution to the problem of estimating efficiently, and without having to choose knot locations, a smooth function of multiple predictors from noisy observations of the function. Smoothing parameters were chosen by approximate UBRE as described in Section 3.1.2. The tuning constant F was identified to be 3; further increasing the value of this parameter did not change the results reported in the next section.

True values for the ATEs, under the scenarios detailed above, were obtained via simulation. Specifically, 10000 replicate datasets were generated according to model (13) and ATEs calculated based on the true linear predictors. The simulated average true ATEs for the normal, mixture of normals and gamma cases are -0.43, -0.15 and -0.45, respectively.

Estimates of the ATE were obtained using the proposed mixed model and, for the sake of comparison, the semiparametric bivariate model of Marra & Radice (2011a) which neglects the presence of random effects (henceforth, these two models will be referred to as mixed SRBP and SRBP, respectively). The calculation of the ATE for mixed SRBP requires an estimate of the random effects distribution. This was obtained, using empirical Bayes, as weighted averages of the estimated mass points, $\hat{\boldsymbol{m}}_1, \ldots, \hat{\boldsymbol{m}}_F$, with respective weights $\Pr(\hat{\boldsymbol{m}}_1|\boldsymbol{y}_i), \ldots, \Pr(\hat{\boldsymbol{m}}_F|\boldsymbol{y}_i)$. That is, for each i,

$$\hat{\boldsymbol{u}}_i^{\mathrm{EB}} = \sum_{l=1}^{F} \hat{\boldsymbol{m}}_l \Pr(\hat{\boldsymbol{m}}_l|\boldsymbol{y}_i) = \sum_{l=1}^{F} \hat{\boldsymbol{m}}_l \frac{\Pr(\boldsymbol{y}_i|\hat{\boldsymbol{m}}_l)\hat{\pi}_l}{\sum_{l'=1}^{F} \Pr(\boldsymbol{y}_i|\hat{\boldsymbol{m}}_{l'})\hat{\pi}_{l'}}. \tag{14}$$

4.2. Results

Tables 1 and 2 display the percentage biases and the root mean squared errors (RMSEs) of the estimated ATEs and ρ_ε's obtained using SRBP and mixed SRBP, when n_i is a randomly chosen number between 9 and 11 and m is set to 200 and 600, and random effects are generated using bivariate normal (N), mixture of normals (MN) and gamma (G) distributions. Tables 4 and 5, reported in Appendix B, provide the same information but for the case in which n_i is set 1 and m set to 2000 and 6000.

The main results can be summarized as follows:

- Table 1 shows that, under the N and G scenarios, mixed SRBP is only slightly better than SRBP, in terms of accuracy and precision of the estimated ATEs. This suggests that, under the N and G cases, the model neglecting cluster specific random effects can still yield good estimates of the average treatment effect. A likely explanation is that the parameter that links the two equations of the bivariate model (i.e. ρ_ε) captures correlations due to both unobserved confounders and cluster or 'litter' effect. However, this is not true when the bivariate random effects distribution is not unimodal, the case in which mixed SRBP considerably outperforms SRBP. These conclusions are in agreement with previously reported findings on the impact of misspecification of the random effects distribution on parameter estimation within the class of GLMMs; see Heagerty (1999), Chen *et al.* (2002) and Agresti *et al.* (2004).

TABLE 1

Percentage biases and root mean squared errors of the estimated average treatment effects (ATEs) obtained using the semiparametric recursive bivariate probit model without and with random effects (SRBP and mixed SRBP, respectively).

		$n \approx 2000$						$n \approx 6000$					
		N		MN		G		N		MN		G	
ρ_ϵ	Method	Bias (%)	RMSE	Bias (%)	RMSE	Bias (%)	RMSE	Bias (%)	RMSE	Bias (%)	RMSE	Bias (%)	RMSE
0.1	SRBP	−2.65	0.039	−64.68	0.107	−1.65	0.040	−2.58	0.025	−66.57	0.103	−2.01	0.026
	mixed SRBP	−0.47	0.037	0.13	0.021	−1.58	0.039	−0.21	0.023	1.84	0.012	−0.53	0.026
0.5	SRBP	−2.74	0.036	−49.23	0.083	−1.49	0.036	−2.38	0.023	−49.76	0.078	−1.62	0.024
	mixed SRBP	−0.43	0.033	−0.58	0.025	−0.04	0.035	0.03	0.021	2.13	0.011	−0.05	0.021
0.9	SRBP	−1.78	0.029	−30.49	0.055	−1.30	0.028	−2.22	0.019	−32.76	0.052	−1.44	0.020
	mixed SRBP	−0.01	0.027	−4.41	0.029	0.29	0.028	−0.36	0.016	−3.83	0.020	0.18	0.019
−0.1	SRBP	−2.77	0.041	−72.95	0.121	−1.46	0.039	−2.68	0.026	−75.50	0.117	−1.79	0.024
	mixed SRBP	−0.34	0.042	−1.26	0.027	−2.58	0.047	−0.55	0.025	1.39	0.013	−0.10	0.026
−0.5	SRBP	−2.64	0.041	−90.54	0.147	−1.31	0.037	−2.67	0.026	−94.11	0.145	−1.45	0.023
	mixed SRBP	−0.63	0.038	−4.15	0.039	−0.61	0.037	−0.77	0.024	0.41	0.015	−0.52	0.023
−0.9	SRBP	−3.33	0.040	−115.13	0.183	−1.94	0.037	−3.18	0.027	−115.15	0.177	−0.50	0.022
	mixed SRBP	−0.53	0.034	3.67	0.028	−1.31	0.035	0.01	0.021	1.00	0.023	−0.41	0.022

Notes: Letters N, MN and G stand for bivariate normal, mixture of normals and gamma variates for the random effects. True ATE values are −0.43, −0.15 and −0.45 for the N, MN and G cases. n_i was set to a randomly chosen number between 9 and 11, and m was set to 200 and 600; these produced sample sizes approximately equal to 2000 and 6000. Results are based on 200 replications. See Section 4.1 for further details.

TABLE 2

Percentage biases and root mean squared errors of the estimated correlations between the model errors (ρ_ϵ) obtained using the semiparametric recursive bivariate probit model without and with random effects (SRBP and mixed SRBP, respectively).

		$n \approx 2000$						$n \approx 6000$					
		N		MN		G		N		MN		G	
ρ_ϵ	Method	Bias (%)	RMSE	Bias (%)	RMSE	Bias (%)	RMSE	Bias (%)	RMSE	Bias (%)	RMSE	Bias (%)	RMSE
0.1	SRBP	84.79	0.125	753.90	0.755	39.37	0.102	83.23	0.097	750.04	0.750	−38.40	0.099
	mixed SRBP	−7.93	0.104	9.41	0.201	−5.34	0.242	−8.18	0.065	0.78	0.110	−4.89	0.111
0.5	SRBP	1.18	0.075	81.38	0.408	−0.71	0.078	1.08	0.045	81.03	0.405	−0.75	0.070
	mixed SRBP	−1.37	0.084	−5.20	0.180	−0.82	0.067	−1.73	0.056	−7.44	0.090	−0.63	0.058
0.9	SRBP	−7.98	0.086	6.88	0.064	−5.23	0.065	−8.49	0.080	6.63	0.060	−5.01	0.062
	mixed SRBP	−1.77	0.050	−3.12	0.052	−0.28	0.042	−2.66	0.033	−2.63	0.044	−0.31	0.039
−0.1	SRBP	−123.51	0.155	−926.96	0.928	−63.31	0.111	−122.16	0.133	−922.60	0.923	−61.69	0.081
	mixed SRBP	10.10	0.121	−13.36	0.246	−1.00	0.104	18.08	0.065	−15.55	0.125	8.09	0.062
−0.5	SRBP	−40.77	0.219	−254.41	1.274	−22.88	0.135	−40.28	0.207	−253.49	1.268	−22.23	0.120
	mixed SRBP	0.82	0.079	−2.31	0.294	0.18	0.073	2.00	0.047	−8.74	0.157	−0.01	0.044
−0.9	SRBP	−30.81	0.283	−178.98	1.613	−18.77	0.180	−30.81	0.280	−179.10	1.612	−19.73	0.182
	mixed SRBP	−2.52	0.046	−7.50	0.231	−0.52	0.038	−2.79	0.033	−8.13	0.145	−1.24	0.023

Notes: Letters N, MN and G stand for bivariate normal, mixture of normals and gamma variates for the random effects. See the caption of Table 1 for further details.

- Table 2 shows that, under all random effects distribution scenarios, mixed SRBP performs considerably better than SRBP, in terms of accuracy and precision of the estimated ρ_εs. The unsatisfactory performance of SRBP can be attributed to the fact that a model neglecting the presence of overdispersion will not be able to disentangle different sources of variability (in this case, one due to endogeneity and the other due to overdispersion). This finding is important because the parameter linking the two model equations is useful to ascertaining the presence of endogeneity, and the estimates produced using SRBP can clearly lead to erroneous conclusions.
- The findings for the more computationally challenging scenarios, in which $n_i = 1$ and $m = 2000, 6000$, are essentially the same as those reported above, except that, as expected, the estimates obtained using mixed SRBP are more variable. These results can be found in Tables 4 and 5 given in Appendix B.

Figure 1 provides an example of estimated smooths with corresponding 95% Bayesian pointwise confidence intervals obtained using the mixed SRBP model. The function estimates recover the true functions reasonably well. This is a good result given the complexity of the model.

5. Empirical illustration

The modeling framework described in this article is illustrated using data from an Italian population based survey. The aim of this study is to estimate the causal effect of private health insurance on private medical care utilization in the presence of unobserved confounding and overdispersion. The problem of unobserved confounding arises in such data because insurance coverage is not randomly assigned as in a controlled trial but rather is the result of supply and demand, including individual preferences and health status. As a consequence, differences in outcomes for insured and uninsured individuals might be due not only to the effect of health insurance but also to the effect of unobserved characteristics that are associated with insurance coverage and medical care utilization. If we do not account for the endogeneity of coverage insurance then the estimated effect will be biased, hence leading to distorted assessments of health policy implications. Overdispersion, which in this study can result from unobserved predictors of either private health insurance or private medical care utilization, can also bias the effect of interest. Buchmueller *et al.* (2005) provide an excellent review of these issues. The direction of the bias due to unobserved confounding is unclear *a priori*. Specifically, standard economic models of insurance markets point to the problem of adverse selection: individuals with a greater demand for medical care, because of poor health for instance, are expected to have a greater demand for insurance. In this case, adverse selection would impart a positive bias on the estimate of the insurance effect on medical care utilization. On the other hand, there could be a problem of moral hazard; once insured, individuals consume more care than optimal. Here, moral hazard would contribute to bias in the opposite direction.

5.1. Data

We used data from the Survey on Health, Aging and Wealth (SHAW; Brugiavini *et al.* 2002) which was conducted by the leading Italian polling agency DOXA in 2001. The SHAW sample consists of 1068 households whose head is over 50 years old and

mainly provides information about individual health status, utilization of health services, types of insurance coverage, as well as socio-economic features. The response is utilization of private health care (`util`): an indicator variable that takes value 1 if the subject has private examinations and 0 otherwise. The treatment variable is private health insurance (`ins`): a dummy variable with value 1 if the respondent has private insurance coverage and 0 otherwise. The observed confounders are the continuous covariates age (`age`), income (`inc`), body mass index (`bmi`), the binary variables indicating whether the individual is a male (`male`), is unmarried or widower (`single`), is unemployed (`unemp`), suffers from chronic conditions (`cond`), has a condition that limits activities of daily life (`lim`), suffers from hearing and/or eyesight troubles (`heey`), has ever smoked (`smoke`), and a factor indicating self-reported health status (poor, good and excellent self-perceived health: `poor`, `good` and `exc`, respectively).

5.2. Health care modeling

The methodology presented here is suitable to tackle both endogeneity and overdispersion; the bivariate model allows us to account for unobserved confounding and for the source of variation due to the heterogeneity in the households. Following previous work on the subject (e.g. Holly *et al.* 1998; Fabbri & Monfardini 2003; Marra & Radice 2011*b*), we specified a mixed SRBP model with main terms only. Specifically, the equations for `ins` and `util` are:

$$\begin{aligned}
\texttt{ins}_i^* &= u_{1i} + \theta_{11}\texttt{male}_i + \theta_{12}\texttt{single}_i + \theta_{13}\texttt{unemp}_i + \theta_{14}\texttt{cond}_i + \theta_{15}\texttt{lim}_i + \theta_{16}\texttt{heey}_i \\
&\quad + \theta_{17}\texttt{poor}_i + \theta_{18}\texttt{exc}_i + \theta_{19}\texttt{smoke} + s_{1\texttt{age}}(\texttt{age}_i) + s_{1\texttt{inc}}(\texttt{inc}_i) + s_{1\texttt{bmi}}(\texttt{bmi}_i) + \varepsilon_{1i}, \\
\texttt{util}_i^* &= u_{2i} + \vartheta\texttt{ins} + \theta_{21}\texttt{male}_i + \theta_{22}\texttt{single}_i + \theta_{23}\texttt{unemp}_i + \theta_{24}\texttt{cond}_i + \theta_{25}\texttt{lim}_i \\
&\quad + \theta_{26}\texttt{heey}_i + \theta_{27}\texttt{poor}_i + \theta_{28}\texttt{exc}_i + \theta_{29}\texttt{smoke} + s_{2\texttt{age}}(\texttt{age}_i) + s_{2\texttt{inc}}(\texttt{inc}_i) \\
&\quad + s_{2\texttt{bmi}}(\texttt{bmi}_i) + \varepsilon_{2i}.
\end{aligned}$$

The parameters in the model have the obvious definitions and thin plate regression splines of the continuous covariates with the same settings as those used for the simulation study were employed. The optimal value for tuning parameter F was identified to be 2, that is the random effects distribution is represented by a two point discrete distribution. The non-linear specification for `age`, `inc` and `bmi` arises from the fact that these covariates embody productivity and life-cycle effects that are likely to affect `ins` and `util` non-linearly. In fact, Holly *et al.* (1998) and Fabbri & Monfardini (2003) considered a model for health care utilization that contains linear and quadratic terms in `age`, `inc` and `bmi`, whereas Marra & Radice (2011b) specified a model containing smooth functions of them. For comparison purposes, we also employ the SRBP model and a classic univariate probit model using the same functional form specification. Mixed SRBP can account simultaneously for unobserved confounding and overdispersion, SRBP accounts for unobserved confounding only whereas the probit model cannot account for either of these issues.

Results are displayed in Table 3. Bayesian confidence intervals for the ATE and correlation coefficient were obtained using 1000 coefficient vectors simulated from the posterior distribution of the estimated model parameters (see Section 3.2).

For the mixed SRBP model, the estimated bivariate mass points are $\boldsymbol{m}_1 = (1.50, -0.24)$ and $\boldsymbol{m}_2 = (-8.27, 0.20)$, with probabilities 0.71 and 0.29, suggesting the absence of relevant predictors of private health insurance. The estimates of ρ_ϵ are both negative and statistically

TABLE 3

Estimates of the ATE and ρ_ε in the health care study obtained using the univariate probit model, and semiparametric recursive bivariate probit without and with nonparametric random effects (SRBP and mixed SRBP, respectively).

	$\widehat{\text{ATE}}$(95%CIs)	$\hat{\rho}_\varepsilon$(95%CIs)
Probit	0.07 (−0.05,0.19)	–
SRBP	0.25 (0.13,0.36)	−0.26 (−0.44,−0.07)
mixed SRBP	0.38 (0.18,0.58)	−0.46 (−0.68,−0.24)

Notes: Bayesian confidence intervals for ATE and ρ_ε were calculated using 1000 coefficient vectors simulated from the posterior distribution of the estimated model parameters.

significant, suggesting the presence of endogeneity. Specifically, the point estimate obtained with mixed SRBP is larger than that of SRBP, although their intervals overlap. This confirms the finding by Holly *et al.* (1998) which is consistent with the interpretation that unobserved confounders are present and have an opposite significant effect on `ins` and `util`.

Moving on to the estimated ATE, the result obtained with the univariate probit model suggests that the effect of private health care insurance is not significant. However, this estimate may be biased due to the unmodelled effects. If we look at the results obtained with the SRBP models, that is models which account for unobserved confounding, private health care insurance has a significant positive impact on the probability of using private health care services. Specifically, the mixed SRBP estimate suggests that the probability of using private medical services increases by 0.38 points for an individual with private health coverage as compared to an individual without private insurance. The point estimate obtained with mixed SRBP is larger than that obtained using SRBP, although their intervals overlap. Results for the other parametric coefficients (not reported here) are in agreement with those found in the literature. The change in the correlation coefficient and ATE of mixed SRBP suggests that decomposing the disturbance in the model into a part attributed to endogeneity and another attributed to overdispersion might have led to a more accurate estimate of the effect of interest. Figure 2 shows the impacts of `age`, `inc` and `bmi` for the treatment and outcome equations obtained using the mixed SRBP model. These results support the presence of nonlinear effects in the outcome equation.

In summary, if we employ a univariate probit model to estimate the effect of private health insurance, the impact appears not to be statistically significant. However, this result is likely to be biased by the presence of unobserved confounding and overdispersion. The estimates obtained with the SRBP models, which account for unobserved confounding, are likely to be more realistic, with mixed SRBP also accounting for overdispersion.

6. Discussion

In this paper, we introduce an algorithm for the simultaneous estimation of the equations of a semiparametric recursive bivariate probit model with nonparametric mixing. Estimation is carried out by maximising a penalised likelihood function using an Expectation-Maximisation algorithm. We also address the issues of automatic multiple smoothing parameter selection and inference. Results from our simulation study suggest that the approach is effective for estimating the effect of an endogenous binary predictor on a binary outcome. Interestingly, the model neglecting overdispersion yields average

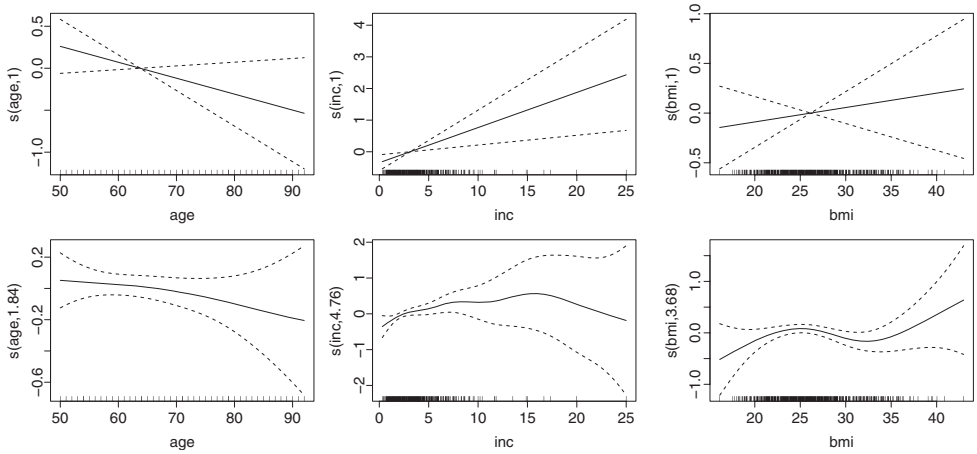

Figure 2. Function estimates in the health care study, on the scale of the respective linear predictors, obtained using a mixed SRBP model. Dashed lines represent 95% Bayesian confidence intervals. The plots in the two panels show the estimated smooth terms of `age`, `inc` and `bmi` for the treatment and outcome equations, respectively. The numbers in brackets in the y-axis captions are the estimated degrees of freedom of the smooth curves.

treatment effect estimates exhibiting substantial bias only for the case of bimodal random effects densities. However, this is not true for the estimation of the parameter linking the two model equations (which is important for ascertaining the presence of endogeneity), where substantial bias is observed in all simulation settings. The methodology was illustrated using data from a survey on private medical care utilization. For this application, differences in the point estimates of the average treatment effects were found between the models with and without random effects, and a classic univariate probit model.

Maximum likelihood estimators are typically sensitive to model error misspecifications. This creates a need for considering different joint distributions of the model errors. A copula approach can be used to that end (e.g. Nelsen 2006). As for the nonparametric approach to the estimation of the random effects distribution, although it yields reasonably efficient parameter estimates, it has several drawbacks. For instance, the resulting discrete estimate of the distribution is not satisfactory as it is more likely to be continuous than discrete. A more relevant drawback is the amount of information required to obtain an accurate estimate of the nonparametric mixing distribution (Carroll & Hall 1988), which can ultimately affect the precision of the effect of interest. We plan on extending the approach presented here in order to include random effects generated by flexible densities that avoid the restrictive assumption of normality but also allow for smooth estimates of the random effects densities. Such densities can, for instance, be represented by mixtures of Gaussians.

Appendix A: Observed information matrix

We briefly describe the method we used to obtain the observed information matrix. First, the log-likelihood, $\ell(\boldsymbol{\delta}|\boldsymbol{y}) = \sum_{i=1}^{m} \log f(\boldsymbol{y}_i|\boldsymbol{\delta})$, of the hierarchical model is written

in terms of both the observed data and random effects, as $\sum_{i=1}^{m} \log\{\sum_{l=1}^{F} f(\boldsymbol{y}_i|\boldsymbol{\delta}; \boldsymbol{u}_i = \boldsymbol{m}_l) \Pr(\boldsymbol{u}_i = \boldsymbol{m}_l|\boldsymbol{\delta})\}$, which for the sake of notational convenience is expressed as $\ell(\boldsymbol{\delta}) = \sum_{i=1}^{m} \log\{\sum_{l=1}^{F} f(\boldsymbol{y}_i, \boldsymbol{u}_l|\boldsymbol{\delta})\}$. From this, we obtain the score function, $\boldsymbol{\Psi}(\boldsymbol{y}, \boldsymbol{\delta}) \equiv \partial\ell(\boldsymbol{\delta})/\partial\boldsymbol{\delta}$, as

$$\begin{aligned}\boldsymbol{\Psi} &= \sum_{i=1}^{m}\sum_{l=1}^{F} \frac{f(\boldsymbol{y}_i, \boldsymbol{u}_l|\boldsymbol{\delta})}{\{\sum_{l=1}^{F} f(\boldsymbol{y}_i, \boldsymbol{u}_l|\boldsymbol{\delta})\}} \frac{\partial \log f(\boldsymbol{y}_i, \boldsymbol{u}_l|\boldsymbol{\delta})}{\partial\boldsymbol{\delta}} \\ &= \sum_{i=1}^{m}\sum_{l=1}^{F} w_{il} \frac{\partial \log f(\boldsymbol{y}_i, \boldsymbol{u}_l|\boldsymbol{\delta})}{\partial\boldsymbol{\delta}} = \sum_{i=1}^{m}\sum_{l=1}^{F} w_{il}\boldsymbol{\psi}_{il},\end{aligned}$$

where w_{il} and $\boldsymbol{\psi}_{il}$ have the obvious definitions. Note that the above score function and the one obtained by differentiating (7) are exactly the same (except for the penalty term). Now, the observed information matrix $\mathcal{J} \equiv -\partial^2\ell(\boldsymbol{\delta})/\partial\boldsymbol{\delta}\partial\boldsymbol{\delta}^\top = -\partial\boldsymbol{\Psi}/\partial\boldsymbol{\delta}^\top$, is obtained as

$$\begin{aligned}\mathcal{J} = &- \sum_{i=1}^{m}\sum_{l=1}^{F} w_{il} \frac{\partial^2 \log f(\boldsymbol{y}_i, \boldsymbol{u}_l|\boldsymbol{\delta})}{\partial\boldsymbol{\delta}\partial\boldsymbol{\delta}^\top} \\ &- \sum_{i=1}^{m}\sum_{l=1}^{F} w_{il} \frac{\partial \log f(\boldsymbol{y}_i, \boldsymbol{u}_l|\boldsymbol{\delta})}{\partial\boldsymbol{\delta}} \frac{\partial \log f(\boldsymbol{y}_i, \boldsymbol{u}_l|\boldsymbol{\delta})}{\partial\boldsymbol{\delta}^\top} \\ &+ \sum_{i=1}^{m} \left\{\sum_{l=1}^{F} w_{il} \frac{\partial \log f(\boldsymbol{y}_i, \boldsymbol{u}_l|\boldsymbol{\delta})}{\partial\boldsymbol{\delta}}\right\} \left\{\sum_{l=1}^{F} w_{il} \frac{\partial \log f(\boldsymbol{y}_i, \boldsymbol{u}_l|\boldsymbol{\delta})}{\partial\boldsymbol{\delta}}\right\}^\top \\ = &- \frac{\partial^2 \mathcal{Q}}{\partial\boldsymbol{\delta}\partial\boldsymbol{\delta}^\top} - \sum_{i=1}^{m} \mathrm{Var}(\boldsymbol{\psi}_i),\end{aligned}$$

with expressions involving model parameters being evaluated at parameter estimates obtained at convergence of the fitting algorithm.

Appendix B: Additional simulation results

TABLE 4

Percentage biases and root mean squared errors of the estimated average treatment effects (ATEs) obtained using the semiparametric recursive bivariate probit model without and with random effects (SRBP and mixed SRBP, respectively).

		$n = 2000$						$n = 6000$					
		N		MN		G		N		MN		G	
ρ_ϵ	Method	Bias (%)	RMSE	Bias (%)	RMSE	Bias (%)	RMSE	Bias (%)	RMSE	Bias (%)	RMSE	Bias (%)	RMSE
0.1	SRBP	−2.89	0.040	−62.67	0.104	−3.90	0.041	−2.69	0.028	−64.71	0.099	−3.94	0.033
	mixed SRBP	1.76	0.080	7.89	0.117	1.96	0.080	0.87	0.078	5.35	0.109	0.99	0.076
0.5	SRBP	−2.32	0.035	−46.41	0.081	−3.51	0.037	−2.29	0.028	−46.92	0.072	−3.49	0.029
	mixed SRBP	2.02	0.075	2.31	0.093	1.54	0.061	1.56	0.076	1.97	0.088	1.23	0.056
0.9	SRBP	−2.18	0.027	−30.44	0.057	−4.07	0.033	−2.21	0.019	−31.55	0.043	−5.11	0.028
	mixed SRBP	1.47	0.046	3.20	0.056	0.31	0.045	1.22	0.037	2.82	0.052	0.44	0.039
−0.1	SRBP	−3.08	0.042	−71.31	0.118	−3.88	0.042	−3.12	0.037	−73.85	0.094	−3.68	0.037
	mixed SRBP	2.81	0.090	3.70	0.121	2.10	0.080	2.53	0.085	3.17	0.108	2.00	0.081
−0.5	SRBP	−3.14	0.042	−91.01	0.147	−4.33	0.045	−3.47	0.036	−98.57	0.136	−4.95	0.036
	mixed SRBP	2.88	0.089	4.21	0.140	1.46	0.077	1.96	0.081	3.75	0.128	1.41	0.069
−0.9	SRBP	−4.22	0.043	−110.13	0.177	−3.32	0.040	−5.56	0.038	−99.78	0.152	−4.41	0.034
	mixed SRBP	2.81	0.088	−3.18	0.152	−0.65	0.064	2.08	0.079	−2.25	0.143	−0.71	0.062

Notes: Letters N, MN and G stand for bivariate normal, mixture of normals and gamma variates for the random effects. True ATE values are −0.43, −0.15 and −0.45 for the N, MN and G cases. n_i was set 1 and m was set to 2000 and 6000; these produced sample sizes equal to 2000 and 6000. Results are based on 200 replications. See Section 4.1 for further details.

TABLE 5

Percentage biases and root mean squared errors of the estimated correlations between the model errors (ρ_ε) obtained using the semiparametric recursive bivariate probit model without and with random effects (SRBP and mixed SRBP, respectively).

		$n = 2000$						$n = 6000$					
		N		MN		G		N		MN		G	
ρ_ε	Method	Bias (%)	RMSE	Bias (%)	RMSE	Bias (%)	RMSE	Bias (%)	RMSE	Bias (%)	RMSE	Bias (%)	RMSE
0.1	SRBP	78.82	0.117	755.05	0.756	34.60	0.094	73.45	0.087	747.05	0.686	35.60	0.082
	mixed SRBP	−7.94	0.364	8.01	0.421	−8.41	0.329	−6.13	0.299	6.57	0.412	−5.41	0.314
0.5	SRBP	1.43	0.074	81.95	0.411	−0.44	0.072	2.02	0.068	79.83	0.385	−2.75	0.068
	mixed SRBP	0.95	0.279	8.97	0.308	−0.02	0.316	0.83	0.267	6.83	0.299	−0.45	0.286
0.9	SRBP	−8.07	0.084	6.92	0.064	−5.74	0.068	−8.45	0.069	6.57	0.059	−6.14	0.054
	mixed SRBP	−2.23	0.104	−0.62	0.167	−1.11	0.108	−2.13	0.094	−0.74	0.151	−0.58	0.087
−0.1	SRBP	−116.27	0.145	−927.66	0.929	−57.40	0.108	−108.54	0.129	−947.78	0.867	−54.73	0.096
	mixed SRBP	9.15	0.396	−11.28	0.433	13.06	0.388	7.45	0.361	−9.57	0.445	11.31	0.347
−0.5	SRBP	−39.06	0.210	−254.34	1.273	−20.06	0.129	−41.27	0.187	−275.84	1.135	−19.65	0.109
	mixed SRBP	3.97	0.289	−12.95	0.439	9.75	0.244	2.67	0.263	−10.65	0.411	9.24	0.219
−0.9	SRBP	−29.87	0.276	−179.84	1.620	−18.69	0.177	−31.25	0.263	−187.65	1.434	−16.65	0.169
	mixed SRBP	−11.17	0.189	−14.35	0.408	−1.28	0.065	−9.38	0.177	−10.24	0.396	−1.54	0.066

Notes: Letters N, MN and G stand for bivariate normal, mixture of normals and gamma variates for the random effects. See the caption of Table 4 for further details.

References

AGRESTI, A., CAFFO, B. & OHMAN-STRICKLAND, P. (2004). Examples in which misspecification of a random effects distribution reduces efficiency, and possible remedies. *Comput. Statist. Data Anal.* **47**, 639–653.

BHATTACHARYA, J., GOLDMAN, D. & MCCAFFREY, D. (2006). Estimating probit models with self-selected treatments. *Statist. Med.* **25**, 389–413.

BRUGIAVINI, A., JAPPELLI, T. & WEBER, G. (2002). The survey on health, aging and wealth. CSEF Working Papers 86, Centre for Studies in Economics and Finance (CSEF), University of Naples, Italy. Available from URL: http://ideas.repec.org/p/sef/csefwp/86.html [Last accessed 24 August 2013].

BUCHMUELLER, T., GRUMBACH, K., KRONICK, R. & KAHN, J. (2005). The effect of health insurance on medical care utilization and implications for insurance expansion: a review of the literature. *Med. Care Res. Rev.* **62**, 3–30.

CARROLL, R.J. & HALL, P. (1988). Optimal rates of convergence for deconvolving a density. *J. Amer. Statist. Assoc.* **83**, 1184–1186.

CHEN, J., ZHANG, D. & DAVIDIAN, M. (2002). A Monte Carlo EM algorithm for generalized linear mixed models with flexible random effects distribution. *Biostatistics* **3**, 347–360.

CHIB, S. & GREENBERG, E. (2007). Semiparametric modeling and estimation of instrumental variable models. *J. Comput. Graph. Statist.* **16**, 86–114.

CRAVEN, P. & WAHBA, G. (1979). Smoothing noisy data with spline functions. *Numer. Math.* **31**, 377–403.

DEMPSTER, A.P., LAIRD, N.M. & RUBIN, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. R. Statist. Soc. Ser. B Statist. Methodol.* **39**, 1–22.

DUCHON, J. (1977). Splines minimizing rotation-invariant semi-norms in solobev spaces. In *'Construction Theory of Functions of Several Variables*, eds. W. Schemp & K. Zeller, pp. 85–100. Springer: Springer.

FABBRI, D. & MONFARDINI, C. (2003). Public vs. private health care services demand in italy. *Giornale degli Economisti* **62**, 93–123.

FOLLMANN, D.A. & LAMBERT, D. (1989). Generalizing logistic regression by nonparametric mixing. *J. Amer. Statist. Assoc.* **84**, 295–300.

GOLDMAN, D., BHATTACHARYA, J., MCCAFFREY, D., DUAN, N., LEIBOWITZ, A., JOYCE, G. & MORTON, S. (2001). Effect of insurance on mortality in an hiv-positive population in care. *J. Amer. Statist. Assoc.* **96**, 883–894.

GREENE, W.H. (2012). *Econometric Analysis*. New York: Prentice Hall.

GU, C. (1992). Cross validating non-gaussian data. *J. Comput. Graph. Statist.* **1**, 169–179.

GU, C. (2002). *Smoothing Spline ANOVA Models*. London: Springer-Verlag.

HASTIE, T. & TIBSHIRANI, R. (1993). Varying-coefficient models. *J. R. Statist. Soc. Ser. B Statist. Methodol.* **55**, 757–796.

HEAGERTY, P.J. (1999). Marginally specified logistic-normal models for longitudinal binary data. *Biometrics* **55**, 688–698.

HEAGERTY, P.J. & KURLAND, B.F. (2001). Misspecified maximum likelihood estimates and generalised linear mixed models. *Biometrika* **88**, 973–985.

HECKMAN, J. (1978). Dummy endogenous variables in a simultaneous equation system. *Econometrica* **46**, 931–959.

HOLLY, A., GARDIOL, L., DOMENIGHETTI, G. & BISIG, B. (1998). An econometric model of health care utilization and health insurance in switzerland. *Eur. Econ. Rev.* **42**, 513–522.

LAIRD, N. (1978). Nonparametric maximum likelihood estimation of a mixing distribution. *J. Amer. Statist. Assoc.* **73**, 805–811.

LESPERANCE, M.L. & KALBFLEISCH, J.D. (1992). An algorithm for computing the nonparametric MLE of a mixing distribution. *J. Amer. Statist. Assoc.* **87**, 120–126.

LINDSAY, B.G. (1983). The geometry of mixture likelihoods, Part II: the exponential family. *Annals Statist.* **11**, 783–792.

LOUIS, T.A. (1982). Finding the observed information matrix when using the EM algorithm. *J. R. Statist. Soc. Ser. B Statist. Methodol.* **44**, 226–233.

MADDALA, G.S. (1983). *Limited Dependent and Qualitative Variables in Econometrics*. Cambridge: Cambridge University Press.

MARRA, G. & RADICE, R. (2011a). Estimation of a semiparametric recursive bivariate probit model in the presence of endogeneity. *Canad. J. Statist.* **39**, 259–279.

MARRA, G. & RADICE, R. (2011b). A flexible instrumental variable approach. *Statist. Model.* **11**, 581–279.

MARRA, G. & RADICE, R. (2013). SemiParBIVProbit: Semiparametric Bivariate Probit Modelling. R package version 3.2-8. Available from URL: http://CRAN.R-project.org/package=SemiParBIVProbit [Last accessed 24 August 2013.]

MARRA, G. & WOOD, S.N. (2012). Coverage properties of confidence intervals for generalized additive model components. *Scand. J. Statist.* **39**, 53–74.

MONFARDINI, C. & RADICE, R. (2008). Testing exogeneity in the bivariate probit model: a monte carlo study. *Oxford B. Econ. Statist.* **70**, 271–282.

NELSEN, R. (2006). *An Introduction to Copulas*. New York: Springer.

NEUHAUS, J.M., HAUCK, W.W. & KALBFLEISCH, J.D. (1992). The effects of mixture distribution misspecification when fitting mixed-effects logistic models. *Biometrika* **79**, 755–762.

PAPAGEORGIOU, G. & HINDE, J. (2012). Multivariate generalized linear mixed models with semi-nonparametric and smooth nonparametric random effects densities. *Statist. Comput.* **22**, 79–92.

R Development Core Team. (2013). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0. Available from URL: http://www.R-project.org [Last accessed 24 August 2013.]

RUPPERT, D., WAND, M.P. & CARROLL, R.J. (2003). *Semiparametric Regression*. New York: Cambridge University Press.

SILVERMAN, B. (1985). Some aspects of the spline smoothing approach to non-parametric regression curve fitting. *J. R. Statist. Soc. Ser. B Statist. Methodol.* **47**, 1–52.

VERBEKE, G. & LESAFFRE, E. (1996). A linear mixed-effects model with heterogeneity in the random-effects population. *J. Amer. Statist. Assoc.* **91**, 217–221.

WAHBA, G. (1983). Bayesian 'confidence intervals' for the cross-validated smoothing spline. *J. R. Statist. Soc. Ser. B Statist. Methodol.* **45**, 133–150.

WILDE, J. (2000). Identification of multiple equation probit models with endogenous dummy regressors. *Econom. Lett.* **69**, 309–312.

WOOD, S.N. (2004). Stable and efficient multiple smoothing parameter estimation for generalized additive models. *J. Amer. Statist. Assoc.* **99**, 673–686.

WOOD, S.N. (2006). *Generalized Additive Models: An Introduction with R*. Chapman & Hall/CRC, London.

WOOLDRIDGE, J.M. (2010). *Econometric Analysis of Cross Section and Panel Data*. Cambridge: MIT Press.