Statistica Sinica **20** (2010), 513-536

A SEMIPARAMETRIC APPROACH FOR A MULTIVARIATE SAMPLE SELECTION MODEL

Marie Chavent, Benoît Liquet and Jérôme Saracco

University of Bordeaux

Abstract: Most of the common estimation methods for sample selection models rely heavily on parametric and normality assumptions. We consider in this paper a multivariate semiparametric sample selection model and develop a geometric approach to the estimation of the slope vectors in the outcome equation and in the selection equation. Contrary to most existing methods, we deal symmetrically with both slope vectors. Moreover, the estimation method is link-free and distribution-free. It works in two main steps: a multivariate sliced inverse regression step, and a canonical analysis step. We establish $\sqrt{n}$-consistency and asymptotic normality of the estimates. We describe how to estimate the observation and selection link functions. The theory is illustrated with a simulation study.

Key words and phrases: Canonical analysis, eigen-decomposition, multivariate SIR, semiparametric regression models, sliced inverse regression (SIR).

1. Introduction

Sample selection models (SSM) are described by two equations. A selection equation specifies the state "observed / non-observed (missing)" of the dependent variable y as a function of explanatory variables x. An outcome equation specifies the value of the dependent variable y as another function of explanatory variables x. Numerous papers dealing with univariate SSM have been published. The adjective "univariate" refers to $y \in \mathbb{R}$. In this paper, we focus on multivariate SSM, that is, when $y \in \mathbb{R}^q,\ q > 1$.

Let us first briefly review univariate SSM. Heckman (1979) introduced what is now regarded as the prototype selection model. Amemiya (1985) refers to this model as the type II Tobit model:

$$\begin{aligned}
&(E1)\ : y_1^* = \theta_1 + x'\beta_1 + \varepsilon_1\\
&(E2)\ : y_2^* = \theta_2 + x'\beta_2 + \varepsilon_2\\
&(E3)\ : y_2 = \mathbb{I}[y_2^* > 0]\\
&(E4)\ : y_1 = y_1^* y_2\\
&(E5)\ : (\varepsilon_1, \varepsilon_2)'|x \sim \mathcal{N}(0, \Gamma), \qquad \Gamma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix},
\end{aligned}$$

where the notation $\mathbb{I}$ designates the indicator function. The observed variables are $y_1 \in \mathbb{R}$, $y_2 \in \{0,1\}$ and $x \in \mathbb{R}^p$. Note that in this model, the explanatory variable x does not include the y variable, contrary to Maddala (1983) who considered a more general simultaneous equation modelling framework where the outcome y can appear on both right and left hand sides of $(E1)$ and $(E2)$. Note also that, in $(E4)$, missing values are denoted by zero, leading to possible confusions with zero as an actual observed value for y_1. Equation $(E3)$ is the selection equation and $(E2)$ is the potential outcome equation. The maximum likelihood method is generally used to estimate such models. The score function is highly non-linear. The convergence of the algorithm heavily depends on the choice of good initial values, and the asymptotic properties of the estimate are very sensitive to model specification. This has been discussed by Goldberger (1983), among others. Alternative methods have been designed. Heckman (1979) proposed a two-step method estimating first the selection equation, and then using the result to estimate the outcome equation in a second stage. Many authors have considered parametric estimation methods. For a survey of these aspects, see Amemiya (1985), Maddala (1983, 1993), or Blundell and Smith (1993).

Semiparametric estimation methods have been developed to bypass the sensitivity to specification assumptions. They handle more general models, especially for error specification. Melenberg and van Soest (1993) give an overview of the semiparametric estimation methods for SSM. Most semiparametric estimation techniques of SSM also proceed in two stages. The first gives a consistent estimate of the slope of the selection equation. The second stage works with the non-missing y only: (i) building a biased estimate of the slope of the outcome equation, and (ii) correcting for this bias with the help of the slope estimated in the first step. Duan and Li (1987), Newey (1991) Ahn and Powell (1993), and Lee (1994) follow such a scheme.

In this paper, we examine multivariate sample selection models (MSSM) which are a generalization of the type II Tobit model when the dependent variable y is a vector of $\mathbb{R}^q$. This kind of model can also be seen as a generalization of classical multivariate Tobit model: $y = \max(y^*, 0)$ where $y^* = Cx + \varepsilon$, $\varepsilon \sim \mathcal{N}(0, \Gamma)$, and C is a $q \times p$ matrix of coefficients (see for instance Eiswerth and Shonkwiler (2006) for a brief presentation and an ecological application of this model).

We focus on a semiparametric MSSM by introducing unknown link functions in the selection and outcome equations in order to get a more flexible model. Moreover, we do not assume that the distribution of the error term is a multivariate normal distribution. Like Duan and Li (1987) in the univariate case, we propose a link-free and distribution-free estimation method. Contrary to most existing methods, we deal symmetrically with both slopes (of the selection and outcome equations).

In Section 2, we give a description of the semiparametric MSSM. In Section 3, we employ the geometric approach to the estimation of the slopes of the outcome and selection equations from a population point of view, and we give the corresponding sample version in order to obtain the slope estimators. The estimation method works in two steps (which have nothing to do with the two classical stages of the approaches mentioned above). The first one performs a multivariate sliced inverse regression (MSIR) analysis. The second step converts the MSIR indices to estimators of the slopes by means of two canonical analyses. The corresponding numerical algorithm is fast (since the method is based on only a few matrix operations and eigen-decompositions, without need for any time-consuming iterative computations) and does not require starting values. Asymptotic properties of the slope estimators are derived in Section 4. Simulation results are reported in Section 5. Finally, concluding remarks are given in Section 6.

2. A Semiparametric Multivariate Sample Selection Model

We consider the following semiparametric multivariate sample selection model: for $j = 1, \ldots, q$,

$$y^{(j)} = \begin{cases} g_1^{(j)}\left(\tilde{x}_1'\tilde{\gamma}_1, \varepsilon_1^{(j)}\right) & \text{if } g_2^{(j)}\left(\tilde{x}_2'\tilde{\gamma}_2, \varepsilon_2^{(j)}\right) > 0 \\ MV & \text{otherwise.} \end{cases} \tag{2.1}$$

- The symbol MV symbolically indicates a missing (non-observed) value for $y^{(j)}$ in order to avoid any confusion with zero as an observed value.
- The dependent variable $y = (y^{(1)}, \ldots, y^{(q)}) \in \mathbb{R}^q$ (when each $y^{(j)}$ is observed) is a q-dimensional random vector. In the following, we will see that there is no need to require all values for the $y^{(j)}$'s to be real.
- The functions $g_1^{(j)}$ and $g_2^{(j)}$ are unknown link functions. For the j-th component $y^{(j)}$ of y, $g_1^{(j)}$ is called the observation link function and $g_2^{(j)}$ the selection link function.
- The variables $\tilde{x}_1 \in \mathbb{R}^{p_1}$ and $\tilde{x}_2 \in \mathbb{R}^{p_2}$ are subvectors of a random vector $x \in \mathbb{R}^p$, assumed to have an elliptically symmetric distribution with parameters $\mu = E(x)$ and $\mathrm{Var}(x) = \Sigma$. Let A_k, $k = 1, 2$, be a $p \times p_k$ matrix that selects the components of $\tilde{x}_k$ in x, that is: $\tilde{x}_k = A_k'x$. This matrix has exactly one "1" in each column and at most one "1" in each row, and the other elements are "0". From the definition of A_k, this matrix is a full column rank matrix such that $A_k'A_k = I_{p_k}$. The matrices A_1 and A_2 are assumed to be known a priori; they are not chosen arbitrarily by the user, they need to be assumed based on existing theory on the exclusion of specific variables. It follows that $\tilde{x}_1$ and

$\tilde{x}_2$ are elliptically distributed with parameters $\mu_k = E(\tilde{x}_k) = A_k'\mu,\ k = 1, 2$, and $\Sigma_k = \mathrm{Var}(\tilde{x}_k) = A_k'\Sigma A_k,\ k = 1, 2$.

- Let $\varepsilon^{(j)} = (\varepsilon_1^{(j)}, \varepsilon_2^{(j)})'$, and $\varepsilon = \left(\varepsilon^{(1)\prime}, \ldots, \varepsilon^{(q)\prime}\right)'$. The error term ε is a random vector independent of x, with an unknown distribution.
- The parameters $\tilde{\gamma}_1$ and $\tilde{\gamma}_2$ are the $p_1 \times 1$ and $p_2 \times 1$ real unknown slope parameters. Introduce $\gamma_k = A_k\tilde{\gamma}_k \in \mathbb{R}^p,\ k = 1, 2$, in order to expand $\tilde{\gamma}_k$ to a $p \times 1$ vector with zeros corresponding to the non-selected components.

Under the generality of the unknown link functions in this model, the intercepts, the vector lengths and vector signs of $\tilde{\gamma}_1$ and $\tilde{\gamma}_2$ are not identifiable. Without additional assumptions, only the directions of the observation and selection slope vectors are identifiable. Then, our main purpose is to estimate the directions of the vectors $\tilde{\gamma}_1$ and $\tilde{\gamma}_2$. The nonparametric estimation of $g_1^{(j)}$ and $g_2^{(j)}$ will also be discussed.

We consider model (2.1) as a particular case of a more general multivariate two-index semiparametric regression model of the form

$$y = f(x'\gamma_1, x'\gamma_2, \varepsilon). \tag{2.2}$$

Model (2.2) was introduced by Li (1991) when $y \in \mathbb{R}$. Li (1991) introduced sliced inverse regression in order to estimate the subspace of $\mathbb{R}^p$, spanned by the γ_k's, which is called the e.d.r. (effective dimension reduction) space. In (2.2), since the link function f is assumed to be arbitrary and unknown, the γ_k's are not individually identifiable, while the e.d.r. space is identifiable. Some extensions of the SIR approach to multivariate y have been studied by Aragon (1997), Li, Aragon, Shedden and Thomas Agnan (2003), Saracco (2005), and Barreda, Gannoun and Saracco (2007). It is interesting to note that SIR and Pooled Marginal SIR (a multivariate SIR approach that is used in the next section) do not require a metric structure for the outcome variable(s). Thus, MV values for the $y^{(j)}$'s are easily managed.

In our context, we have to take into account extra information about the e.d.r. space, namely, structural zeros in the slopes γ_1 and γ_2, with a link function f depending on the unknown functions $g_1^{(j)}$ and $g_2^{(j)}$, for $j = 1, \ldots, q$.

We now exhibit in Theorem 1 a geometrical property of this model on which the proposed approach is based. Let $E = \mathrm{Span}(\gamma_1, \gamma_2)$ of $\mathbb{R}^p$. Without additional conditions, we have $\dim(E) \leq 2$. If γ_1 and γ_2 are linearly independent, then $\dim(E) = 2$, and $\{\gamma_1, \gamma_2\}$ is a basis of the e.d.r. space. In order to ensure that we are working on a two-index model (that is $\dim(E) = 2$), let us assign the following *identifiability conditions*.

(i) Each vector $\tilde{x}_k,\ k = 1, 2$, has at least an x-component not present in the other $\tilde{x}_k,\ k = 2, 1$; such a component could be considered k-specific.

(ii) At least one component of γ_k among the k-specific components is non-null, $k = 1, 2$.

Note that these identifiability conditions are stronger than the usual identifiability condition, which is that $\tilde{x}_2$ contains an x-component that is not in $\tilde{x}_1$. The underlying reason for the stronger condition is that the proposed method deals symmetrically with the selection and outcome slope vectors. Knowingly, we do not make use of an important piece of information, namely that the selection probabilities depend only on one of the two index variables, $\tilde{x}_2'\tilde{\gamma}_2$.

We now bring these conditions into a geometrical perspective. Let $E_k = \mathrm{Span}(A_k)$.

Theorem 1. *Under the assumptions of model* (2.1) *and the identifiability conditions, for* $k = 1, 2$, $E \cap E_k = \mathrm{Span}(\gamma_k)$.

Proof. From the definition of A_k, we have $\dim(E_k) = p_k$. The identifiability conditions give (i) $E_1 \not\subset E_2$ and $E_2 \not\subset E_1$, and (ii) $E \cap E_1 \neq E$ and $E \cap E_2 \neq E$. Since $\dim(E) = 2$, we have $\dim(E \cap E_k) \leq 2$. From the definition of E and E_k, $\gamma_k \in E \cap E_k$ and then $\dim(E \cap E_k) \geq 1$. From the identifiability conditions we get, for $k^\star \neq k$, $\gamma_{k^\star} \in E$ and $\gamma_{k^\star} \notin E_k$, thus $\gamma_{k^\star} \notin E \cap E_j$ and $\dim(E \cap E_k) < 2$. Finally, $\dim(E \cap E_k) = 1$ and $E \cap E_k \subset \mathbb{R}^p$ is spanned by γ_k.

We specify in the next section how to determine a basis of E and to deduce a basis $E \cap E_j$ from a population point of view. Then we describe how to estimate the directions of γ_1 and γ_2.

Remark 1. The full model defined in (2.1) can be interpreted as an item non-response model, that is the response status for each outcome measure (or survey item) is governed by a specific selection equation. We can also introduce a simplified model in terms of the type of missing data encountered. Thus we assume that the same selection equation is used for all outcomes: each selection link function $g_2^{(j)}(.)$ is equal to the same link function $g_2(.)$. With unique error term ε_2, the model can be written as

$$y = \begin{cases} g_1\left(\tilde{x}_1'\tilde{\gamma}_1, \varepsilon_1\right) & \text{if } g_2\left(\tilde{x}_2'\tilde{\gamma}_2, \varepsilon_2\right) > 0 \\ MV & \text{otherwise,} \end{cases}$$

where the observation link function $g_1(.)$ takes its values (when they are observed) in $\mathbb{R}^q$ and the error term ε_1 is a q-dimensional random vector. This model can be interpreted as a case non-response model, when the response status for multiple outcomes is clustered at the individual level: an individual either responds to all outcome measures (case response) or does not respond to any outcome measure (case non-response).

Remark 2. The proposed approach can cope with the generalized two-limit selection model of the form: for $j = 1, \ldots, q$,

$$y^{(j)} = \begin{cases} L_1^{*(j)} & \text{if } g_2^{(j)}(\tilde{x}_2'\tilde{\gamma}_2, \varepsilon_2) \leq L_1^{(j)} \\ g_1^{(j)}(\tilde{x}_1'\tilde{\gamma}_1, \varepsilon_1) & \text{if } L_1^{(j)} < g_2^{(j)}(\tilde{x}_2'\tilde{\gamma}_2, \varepsilon_2) < L_2^{(j)} \\ L_2^{*(j)} & \text{if } g_2^{(j)}(\tilde{x}_2'\tilde{\gamma}_2, \varepsilon_2) \geq L_2^{(j)}, \end{cases} \tag{2.3}$$

where $L_1^{*(j)}$ and $L_2^{*(j)}$ are qualitative measures of specific situations, $L_1^{(j)}$ and $L_2^{(j)}$ are thresholds of the selection equation. This model is a multivariate extension of the two-limit Tobit model (see e.g. Maddala (1993)). In addition to the two-limit selection model, it might also be useful to consider more general selection models with multiple non-response categories, such as refusals, don't know, etc., with a distinct selection equation for each category.

Examples for potential application of the proposed model. Semiparametric MSSM has many possible applications in economics. For example, it can be used to study the determinants of innovation behaviour or financial choices. The Community Innovation Survey collects data on the innovative characteristics of EU firms. The data include measures of innovation and related expenditures (Intramural R&D, extramural R&D, Acquisition of machinery, equipment and software, and other external knowledge). MSSM could be useful to exploit this information. The selection equation could give the state "observed / non-observed" of the dependent variable y (having innovation activities) and the outcome equation would give the value of dependent variables (the amount of expenditure for each of the four innovation activities) when innovation activities are observed.

Semiparametric MSSM could also be used in a clinical study when the researcher considers relative potency. For instance, consider a clinical study of two related drugs A and B that belong to the same class (such as two statins), with the primary goal to determine the relative potency for the two drugs. In this kind of application, it is reasonable to assume that the relative potency is determined biologically by the intrinsic nature of the two drugs, therefore the same relative potency (that is the same $\tilde{\gamma}_j$ coefficients) holds for various components of the multivariate outcome measure.

3. Population and Sample Approaches

Our approach splits into two principal steps. In the first step, the idea is to use multivariate sliced inverse regression in order to get a Σ-orthogonal basis of the e.d.r. space $E = \text{Span}(\gamma_1, \gamma_2)$. In the second step, since the linear subspaces E_1 and E_2 are known (because the matrices A_1 and A_2 are assumed to be known

a priori), canonical analyses of the couples (E, E_1) and (E, E_2) can provide bases of $E \cap E_1 = \text{Span}(\gamma_1)$ and $E \cap E_2 = \text{Span}(\gamma_2)$.

3.1. Population version

Step 1: Pooled marginal sliced inverse regression. For model (2.2), Saracco (2005) has shown that pooled marginal sliced inverse regression based on the SIR_α approach, named PMS_α hereafter, provides a basis, denoted $B = [v_1, v_2]$, of the e.d.r. space E, that is $\text{Span}(B) = E$. The major novelty is to consider a transformation (slicing) $T_j(.)$ of $y^{(j)}$ with a specific slice for the missing value (MV) of $y^{(j)}$. The vectors b_k are the eigenvectors corresponding to the two largest eigenvalues of a Σ-symmetric matrix.

More precisely, the idea of this method is to consider the q univariate SIR_α methods of each component $y^{(j)}$ of y on x (based on a specific slicing T_j), and to combine the corresponding M_α matrices (denoted by $M_{\alpha_j}^{(j)}$) in the following pooling:

$$M_{\alpha,P} = \sum_{j=1}^{q} w_j M_{\alpha_j}^{(j)}, \tag{3.1}$$

for positive weights w_j and parameters $\alpha_j \in [0, 1]$. In the $M_{\alpha,P}$ matrix, the α index stands for the vector $(\alpha_1, \ldots, \alpha_q)$ and the P index stands for "pooled". Each transformation T_j categorizes each response $y^{(j)}$ into a new response with $H_j + 1$ levels. We assume that the support of each $y^{(j)}$ is partitioned into H_j fixed slices $s_1^{(j)}, \ldots, s_h^{(j)}, \ldots, s_{H_j}^{(j)}$, plus one slice $s_0^{(j)}$ for the missing value of $y^{(j)}$. For $j = 1 \ldots, q$, the matrices $M_{\alpha_j}^{(j)}$ are $M_{\alpha_j}^{(j)} = (1 - \alpha_j) M_I^{(j)} \Sigma^{-1} M_I^{(j)} + \alpha_j M_{II}^{(j)}$, with

$$\begin{aligned}
M_I^{(j)} &= \text{Var}(E(x|T_j(y^{(j)}))) \\
&= \textstyle\sum_{h=0}^{H_j} p_h^{(j)} (m_h^{(j)} - \mu)(m_h^{(j)} - \mu)', \\
M_{II}^{(j)} &= E\left\{ \left(\text{Var}(x|T_j(y^{(j)})) - E(\text{Var}(x|T_j(y^{(j)}))) \right) \Sigma^{-1} \right. \\
&\qquad \left. \left(\text{Var}(x|T_j(y^{(j)})) - E(\text{Var}(x|T_j(y^{(j)}))) \right)' \right\} \\
&= \textstyle\sum_{h=0}^{H_j} p_h^{(j)} \left(V_h^{(j)} - \overline{V}^{(j)} \right) \Sigma^{-1} \left(V_h^{(j)} - \overline{V}^{(j)} \right),
\end{aligned}$$

where $p_h^{(j)} = P(y^{(j)} \in s_h^{(j)})$, $m_h^{(j)} = E(x|y^{(j)} \in s_h^{(j)})$, $\text{Var}_h^{(j)} = \text{Var}(x|y^{(j)} \in s_h^{(j)})$ and $\overline{V}^{(j)} = \displaystyle\sum_{h=0}^{H_j} p_h^{(j)} V_h^{(j)}$. The matrix $M_I^{(j)}$ is the usual matrix used in the classical SIR approach, often named SIR-I because it relies on a property of the first inverse conditional moment of x given y, while $M_{II}^{(j)}$ correspond with the SIR-II approach using information from the inverse conditional variance of x given y.

When $\alpha_j = 0$ (resp. $\alpha_j = 1$), the method used with $M_{\alpha_j}^{(j)}$ is equivalent to the SIR-I (resp. SIR-II) approach for the j-th component of y.

For (2.2), crucial conditions for the theoretical success of the SIR_α and PMS_α methods are a linearity condition

$$E(v'x|\gamma_1'x, \gamma_2'x) \text{ is linear for any } v, \tag{3.2}$$

and a constant variance condition

$$\text{Var}(x|\gamma_1'x, \gamma_2'x) \text{ is non-random.} \tag{3.3}$$

Note that (3.2) is satisfied when x has an elliptically symmetric distribution, and (3.3) is satisfied when x follows a multivariate normal distribution (which is elliptical). Moreover, some mild departure from the elliptical symmetry does not affect the application of SIR or MSIR, see for instance Li (1991, 1997). Note also that low-dimensional projections from high-dimensional data are known to improve the elliptical symmetry of data distribution, see for details Diaconis and Freedman (1984) or Hall and Li (1993). Finally, an insightful discussion about the SIR methodology and applications can be found in Chen and Li (1998), and most of these comments are still valid for the MSIR approach.

Under (3.2) and (3.3), the eigenvectors v_1, v_2 associated with the largest two eigenvalues of $\Sigma^{-1}M_{\alpha,P}$ are e.d.r. directions and span the e.d.r. space.

Step 2a: Two canonical analysis. Consider the subspaces E_k and E of $\mathbb{R}^p$ equipped with the inner product Σ. Canonical analysis is a useful tool to find a Σ-orthogonal basis of $E_k \cap E$. This basis is formed by the eigenvectors corresponding to the eigenvalue 1 of $P_{E_k}P_E$, where P_{E_k} and P_E are, respectively, the Σ-orthogonal projectors onto E_k and E.

Specifically, we have $P_E = B(B'\Sigma B)^{-1}B'\Sigma = BB'\Sigma$ and $P_{E_k} = A_k(A_k'\Sigma A_k)^{-1}A_k'\Sigma$. It is equivalent and simpler to diagonalize $P_{E_k}P_EP_{E_k}$ which is a Σ-symmetric matrix. Call b_k the unique eigenvector corresponding to the eigenvalue 1 of $P_{E_k}P_EP_{E_k}$. From Theorem 1, the eigenvector b_k is colinear to γ_k and is Σ-normalized: $b_k'\Sigma b_k = 1$.

Step 2b: Retrieval of the direction of $\tilde{\gamma}_k$. We can derive a vector, $\tilde{b}_k$, colinear to $\tilde{\gamma}_k$: $\tilde{b}_k = A_k'b_k$. This vector $\tilde{b}_k$ is Σ_k-normalized: $\tilde{b}_k'\Sigma_k\tilde{b}_k = 1$.

3.2. Estimation of the directions

Directions are obtained from computations based only on covariance matrices. Substituting estimates in place of these matrices yields estimated directions. Let $\{(y_i, x_i),\ i = 1, \ldots, n\}$ be a sample from the reference model (2.1). Let $\hat{\Sigma}$ be the empirical covariance matrix of the x_i's.

Step 1: Estimating a basis of the e.d.r. space E by the PMS$_\alpha$ method. We have to estimate the matrix $M_{\alpha,P}$. To do this, using the H_j+1 slices of each component $y^{(j)}$, it is straightforward to estimate the matrices $M_I^{(j)}$ and $M_{II}^{(j)}$ by substituting empirical versions of the moments for their theoretical counterparts, and therefore to obtain the estimated matrices $\hat{M}_{\hat{\alpha}_j}^{(j)}$. Note that, for the choice of the slices of T_j, $s_0^{(j)}$ contains the cases corresponding to the missing value (MV) of $y^{(j)}$. The other slices, $s_h^{(j)}$, $h = 1, \ldots, H_j$, are made by splitting the range of the non-missing values of the jth component of y into slices of nearly equal weight. The choice of number H_j of slices is less crucial than the choice of the smoothing parameter in nonparametric regression: in practice, we propose to choose H_j such that $2 < H_j < [n_j^*/2]$, where n_j^* is the number of non-missing $y_i^{(j)}$ in the sample and $[a]$ denotes the integer part of a. For the choice of the weights w_j, we use equal weights $w_j = 1/q$ for $j = 1, \ldots, q$ if we have no a priori information on the importance of each component $y^{(j)}$ of y. The parameters α_j are individually chosen for each matrix $M_{\alpha_j}^{(j)}$, and we propose to use the method based on the test approach of Saracco (2001) that does not require the estimation of the link functions. Therefore we obtain the estimated matrix

$$\hat{M}_{\hat{\alpha},P} = \frac{1}{q} \sum_{j=1}^{q} \hat{M}_{\hat{\alpha}_j}^{(j)}. \tag{3.4}$$

The two estimated e.d.r. directions, $\hat{v}_1$ and $\hat{v}_2$, are then the eigenvectors corresponding to the two largest eigenvalues of $\hat{\Sigma}^{-1}\hat{M}_{\hat{\alpha},P}$. These vectors form a $\hat{\Sigma}$-orthonormal system. Let $\hat{E} = \mathrm{Span}(\hat{B})$ where $\hat{B} = [\hat{v}_1, \hat{v}_2]$.

Step 2a: Estimating the direction of γ_k, $k = 1, 2$. We obtain these directions by canonical analyses of $(\hat{E}, E_1)$ and $(\hat{E}, E_2)$: the estimate of the direction of γ_k is the eigenvector $\hat{b}_k$ corresponding to the major eigenvalue of the $\hat{\Sigma}$-symmetric matrix $\hat{P}_{E_k}\hat{P}_{\hat{E}}\hat{P}_{E_k}$, where $\hat{P}_{\hat{E}} = \hat{B}(\hat{B}'\hat{\Sigma}\hat{B})^{-1}\hat{B}'\hat{\Sigma} = \hat{B}\hat{B}'\hat{\Sigma}$ and $\hat{P}_{E_k} = A_k(A_k'\hat{\Sigma}A_k)^{-1}A_k\hat{\Sigma}$.

Step 2b: Estimating the direction of $\tilde{\gamma}_k$, $k = 1, 2$. The estimates of the direction of $\tilde{\gamma}_k$ are then given by $\hat{\tilde{b}}_k = A_k'\hat{b}_k$.

Remarks.

- In order to obtain an estimate of the entire vector $\tilde{\gamma}_k$ (and not only of its direction), we can normalize this vector in the Σ_k metric, and impose the sign of a non-null component of $\tilde{\gamma}_k$.
- For the two-limit model (2.3), there must be one slice for each kind of missing y value. The other slices are built, splitting the other cases in the usual way.

We study the asymptotic properties of the estimators $\hat{\tilde{b}}_1$ and $\hat{\tilde{b}}_2$ in the next section. First, however, we discuss a topic of practical concern connected with the estimation process: the estimation of the link functions of the model (2.1).

Rough approximation of the link functions $g_1^{(j)}$ and estimation of the state of y probabilities. Let us simplify the reference model by assuming an additive error component: for $j = 1, \ldots, q$,

$$y^{(j)} = \begin{cases} g_1^{(j)}(\tilde{x}_1'\tilde{\gamma}_1) + \varepsilon_1^{(j)} & \text{if } \; g_2^{(j)}(\tilde{x}_2'\tilde{\gamma}_2) + \varepsilon_2^{(j)} \; > \; 0 \\ 0 & \text{otherwise,} \end{cases}$$

with $E(\varepsilon_1^{(j)}) = E(\varepsilon_2^{(j)}) = 0$. A rough approximation of the jth-observation link function, $g_1^{(j)}$, may be obtained nonparametrically by kernel or spline methods. Eubank (1988) and Haerdle (1990) give an operational description of these tools. We may, for instance, build a naive Nadaraya-Watson kernel estimate from the subsample of cases where $y^{(j)}$ is non missing by regressing $y^{(j)}$ on $\tilde{x}_1'\hat{\tilde{b}}_1$. This estimator is generally a biased estimator of $g_1^{(j)}$ since $E(\varepsilon_1^{(j)} | \tilde{x}_1'\tilde{\gamma}_1, g_2(\tilde{x}_2'\tilde{\gamma}_2) + \varepsilon_2^{(j)} > 0)$ is non-null.

Here is the case of the selection link functions, $g_2^{(j)}$. What is interesting is to estimate the probability of the state of $y^{(j)}$. In order to describe the state, introduce the qualitative variable $t^{(j)}$ for the one-limit selection model (2.1) (resp. for the two-limits selection model (2.3)) as

$$t^{(j)} = \begin{cases} 1 \text{ if } y^{(j)} \text{ is observed} \\ 0 \text{ otherwise} \end{cases} \qquad \left(\text{resp.} \quad t^{(j)} = \begin{cases} 0 \text{ if } y^{(j)} = L_1^* \\ 1 \text{ if } y^{(j)} \text{ is observed} \\ 2 \text{ if } y^{(j)} = L_2^* \end{cases} \right).$$

From each sample $\{(t_i^{(j)}, r_i),\ i = 1, \ldots, n\}$ where $r_i = \tilde{x}_{2i}'\hat{\tilde{b}}_2$, we can obtain a naïve Nadaraya-Watson estimate of the probability $P(t^{(j)} = t | r = \tilde{x}_2'\tilde{b}_2)$ as

$$\hat{p}_n^{(j)}(t|r) = \sum_{i=1}^{n} \frac{K\left((r - r_i)/\nu_n\right)}{\sum_{l=1}^{n} K\left((r - r_l)/\nu_n\right)} \mathbb{I}[t_i^{(j)} = t], \tag{3.5}$$

where K is a kernel function and ν_n is the bandwidth chosen by cross validation, for example.

4. Asymptotic Theory

In the sequel, the notation $X_n \longrightarrow_d X$ means that X_n converges in distribution to X as $n \to \infty$. Let $D_1 \otimes D_2$ denote the Kronecker product of the matrices D_1 and D_2 (see Tyler (1981) for some useful properties of the

Kronecker product). From now on, for each $s \times s$ matrix $D = \left(d^{(jk)}\right)$, let $\mathrm{vec}(D) = \left(d^{(11)}, \ldots, d^{(s1)}, d^{(21)}, d^{(22)}, \ldots, d^{(ss)}\right)'$ be the s^2-dimensional column vector of all elements of D.

The necessary assumptions are gathered together below for easy reference.

(A1) $\{(y_i, x_i),\ i = 1, \ldots, n\}$ is a sample of independent observations from model (2.1).

(A2) The supports of each component $y^{(j)}$ (when observed) of y are partitioned into H_j fixed slices $s_1^{(j)}, \ldots, s_h^{(j)}, \ldots, s_{H_j}^{(j)}$ such that $p_h^{(j)} \neq 0$, with a special slice $s_0^{(j)}$ for the missing $y^{(j)}$.

(A3) The covariance matrix Σ is positive definite.

(A4) The two largest eigenvalues of $\Sigma^{-1} M_{\alpha,P}$ satisfy $\lambda_1 \geq \lambda_2 > \lambda_3 \geq 0$.

4.1. Convergence in probability of the estimated directions

Theorem 2. *Under conditions given in* (3.2) *and* (3.3)*, and under* (A1), (A2) *and* (A3)*, we have* $\hat{\tilde{b}}_k = \tilde{b}_k + O_p(n^{-1/2})$*, with the vector* $\tilde{b}_k$ *colinear to* $\tilde{\gamma}_k$*, for* $k = 1, 2,$.

Proof. Classical asymptotic theory gives us $\hat{\Sigma} = \Sigma + O_p(n^{-1/2})$. By the asymptotic theory of PMS_α (see Saracco (2005)), we get $\hat{B} = B + O_p(n^{-1/2})$. Thus,

$$\hat{P}_{\hat{E}} = P_E + O_p(n^{-1/2}). \tag{4.1}$$

From the identifiability conditions, $\mathrm{rank}(A_k' \Sigma A_k) = p_k$. Since $A_k' \hat{\Sigma} A_k = A_k' \Sigma A_k + O_p(n^{-1/2})$, we get $(A_k' \hat{\Sigma} A_k)^{-1} = (A_k' \Sigma A_k)^{-1} + O_p(n^{-1/2})$ and

$$\hat{P}_{E_k} = P_{E_k} + O_p(n^{-1/2}),\ j = 1, 2. \tag{4.2}$$

Combining (4.1) with (4.2) yields $\hat{P}_{E_k} \hat{P}_{\hat{E}} \hat{P}_{E_k} = P_{E_k} P_E P_{E_k} + O_p(n^{-1/2})$, $k = 1, 2$. Consequently, the eigenvector of $\hat{P}_{E_k} \hat{P}_{\hat{E}} \hat{P}_{E_k}$ corresponding to the major eigenvalue converges at the same rate to the corresponding eigenvector for $P_{E_k} P_E P_{E_k}$: $\hat{b}_k = b_k + O_p(n^{-1/2})$, $k = 1, 2$. Finally, since $\hat{\tilde{b}}_k = A_k' \hat{b}_k$ and $\tilde{b}_k = A_k' b_k$, we conclude that $\hat{\tilde{b}}_k = \tilde{b}_k + O_p(n^{-1/2})$, $k = 1, 2$. From Theorem 1, we have $\tilde{b}_k$ colinear to $\tilde{\gamma}_k$.

4.2. Asymptotic distribution of $\hat{\tilde{b}}_k$, $k = 1, 2$

Theorem 3. *Under conditions* (3.2) *and* (3.3)*, and under* (A1), (A2), (A3) *and* (A4)*, we have, for* $k = 1, 2$, $\sqrt{n}(\hat{\tilde{b}}_k - \tilde{b}_k) \longrightarrow_d \mathcal{N}(0, A_k' G_k C^* G_k' A_k)$*, where the expression of* G_k *is given in* (4.3)*, and* C^* *can be found in* Saracco (2005).

Proof. The proof is divided into three steps.

STEP 1: ASYMPTOTIC DISTRIBUTION OF THE CANONICAL ANALYSIS MATRIX. Consider the decomposition

$$\begin{aligned}\sqrt{n}(\hat{P}_{E_k}\hat{P}_{\hat{E}}\hat{P}_{E_k} - P_{E_k}P_EP_{E_k}) \\ = \sqrt{n}(\hat{P}_{E_k} - P_{E_k})\hat{P}_{\hat{E}}(\hat{P}_{E_k} - P_{E_k}) + \sqrt{n}(\hat{P}_{E_k} - P_{E_k})\hat{P}_{\hat{E}}P_{E_k} \\ +\sqrt{n}P_{E_k}\hat{P}_{\hat{E}}(\hat{P}_{E_k} - P_{E_k}) + \sqrt{n}(P_{E_k}\hat{P}_{\hat{E}}P_{E_k} - P_{E_k}P_EP_{E_k}).\end{aligned}$$

The first term of the right hand side is $O_p(n^{-1/2})$. Thus, $\sqrt{n}\big[\mathrm{vec}(\hat{P}_{E_k}\hat{P}_{\hat{E}}\hat{P}_{E_k}) - \mathrm{vec}(P_{E_k}P_EP_{E_k})\big]$ has the same asymptotic distribution as the last three terms of the decomposition. These terms can be written as

$$\begin{aligned}&([P'_{E_k}\hat{P}'_{\hat{E}} \otimes I_p] + [I_p \otimes P_{E_k}\hat{P}_{\hat{E}}])\sqrt{n}[\mathrm{vec}(\hat{P}_{E_k}) - \mathrm{vec}(P_{E_k})] \\ &+[P'_{E_k} \otimes P_{E_k}]\sqrt{n}[\mathrm{vec}(\hat{P}_{\hat{E}}) - \mathrm{vec}(P_E)].\end{aligned}$$

We prove in the Appendix that $\sqrt{n}\big[\mathrm{vec}(\hat{P}_{E_k}) - \mathrm{vec}(P_{E_k})\big]$ has the same asymptotic distribution as $N_k\sqrt{n}[\mathrm{vec}(\hat{\Sigma}) - \mathrm{vec}(\Sigma)]$ where N_k is defined in (A.1). Moreover, from Saracco (2005), since $\hat{\Sigma}^{-1}\hat{M}_{\alpha,P}$ converges in probability to $\Sigma^{-1}M_{\alpha,P}$ we have, with a probability converging to 1, for n sufficiently large, $||\hat{\Sigma}^{-1}\hat{M}_{\alpha,P} - \Sigma^{-1}M_{\alpha,P}|| \leq \lambda_2/2$, where λ_2 is the second major eigenvalue of $\Sigma^{-1}M_{\alpha,P}$. Then we can apply the Lemma 4.1 of Tyler (1981) and obtain the asymptotic distribution of the eigenprojector on the estimated e.d.r. space: $\sqrt{n}\big[\mathrm{vec}(\hat{P}_{\hat{E}}) - \mathrm{vec}(P_E)\big]$ has the same asymptotic distribution as $C_w\sqrt{n}\big[\mathrm{vec}(\hat{\Sigma}^{-1}\hat{M}_{\alpha,P}) - \mathrm{vec}(\Sigma^{-1}M_{\alpha,P})\big]$, where $C_w = -\sum_{\lambda\in w}[(M_{\alpha,P}\Sigma^{-1} - \lambda I_p)^+ \otimes P_\lambda + P'_\lambda \otimes (\Sigma^{-1}M_{\alpha,P} - \lambda I_p)^+]$, with $w = \{\lambda_1, \lambda_2\}$.

Finally, the asymptotic distribution of $\sqrt{n}(\hat{P}_{E_k}\hat{P}_{\hat{E}}\hat{P}_{E_k} - P_{E_k}P_EP_{E_k})$ is then the same as

$$\hat{A}^0\sqrt{n}\left(\begin{bmatrix}\mathrm{vec}(\hat{\Sigma}^{-1}\hat{M}_{\alpha,P}) \\ \mathrm{vec}(\hat{\Sigma})\end{bmatrix} - \begin{bmatrix}\mathrm{vec}(\Sigma^{-1}M_{\alpha,P}) \\ \mathrm{vec}(\Sigma)\end{bmatrix}\right),$$

where $\hat{A}^0 = \begin{bmatrix} A_1^0 \mid \hat{A}_2^0 \end{bmatrix}$, with $A_1^0 = (P'_{E_k} \otimes P_{E_k})C_w$ and $\hat{A}_2^0 = \big([P_{E_k}\hat{P}'_{\hat{E}} \otimes I_p] + [I_p \otimes P_{E_k}\hat{P}_{\hat{E}}]\big)N_k$. Moreover, it is easy to show that $\hat{A}^0 \longrightarrow_P A^0$ where $A^0 = \begin{bmatrix} A_1^0 \mid A_2^0 \end{bmatrix}$ with $A_2^0 = ([P_{E_k}P'_E \otimes I_p] + [I_p \otimes P_{E_k}P_E])\,N_k$.

STEP 2: ASYMPTOTIC DISTRIBUTION OF THE MAJOR EIGENVECTOR. Remembering that $\hat{b}_k$ (resp. b_k) is the eigenvector corresponding to the major eigenvalue of $\hat{P}_{E_k}\hat{P}_{\hat{E}}\hat{P}_{E_k}$ (resp. $P_{E_k}P_EP_{E_k}$), we apply Lemma 2 of Saracco (1997). First, we need to specify the asymptotic distribution of

$$\sqrt{n}\left(\begin{bmatrix}\mathrm{vec}(\hat{P}_{E_k}\hat{P}_{\hat{E}}\hat{P}_{E_k}) \\ \mathrm{vec}(\hat{\Sigma})\end{bmatrix} - \begin{bmatrix}\mathrm{vec}(P_{E_k}P_EP_{E_k}) \\ \mathrm{vec}(\Sigma)\end{bmatrix}\right).$$

From Step 1, this vector has the same asymptotic distribution as

$$\hat{B}^0\sqrt{n}\left(\begin{bmatrix}\mathrm{vec}(\hat{\Sigma}^{-1}\hat{M}_{\alpha,P})\\ \mathrm{vec}(\hat{\Sigma})\end{bmatrix}-\begin{bmatrix}\mathrm{vec}(\Sigma^{-1}M_{\alpha,P})\\ \mathrm{vec}(\Sigma)\end{bmatrix}\right),$$

where $\hat{B}^0=\begin{bmatrix}A_1^0 & \hat{A}_2^0\\ 0_{p^2,p^2} & I_{p^2}\end{bmatrix}$. Since $\hat{A}_2^0 \longrightarrow_P A_2^0$, we get $\hat{B}^0 \longrightarrow_P B^0$ where $B^0=\begin{bmatrix}A_1^0 & A_2^0\\ 0_{p^2,p^2} & I_{p^2}\end{bmatrix}$.

Moreover, from an application of the Delta method, Saracco (2005) shows that

$$\sqrt{n}\left(\begin{bmatrix}\mathrm{vec}(\hat{\Sigma}^{-1}\hat{M}_{\alpha,P})\\ \mathrm{vec}(\hat{\Sigma})\end{bmatrix}-\begin{bmatrix}\mathrm{vec}(\Sigma^{-1}M_{\alpha,P})\\ \mathrm{vec}(\Sigma)\end{bmatrix}\right)\longrightarrow_d \Phi^*=\begin{bmatrix}\mathrm{vec}(\Phi)\\ \mathrm{vec}(\Phi_\Sigma)\end{bmatrix}\sim\mathcal{N}(0,C^*),$$

The expression for C^* can be found in Saracco (2005). Thus we obtain:

$$\sqrt{n}\left(\begin{bmatrix}\mathrm{vec}(\hat{P}_{E_k}\hat{P}_{\hat{E}}\hat{P}_{E_k})\\ \mathrm{vec}(\hat{\Sigma})\end{bmatrix}-\begin{bmatrix}\mathrm{vec}(P_{E_k}P_EP_{E_k})\\ \mathrm{vec}(\Sigma)\end{bmatrix}\right)\longrightarrow_d B^0\Phi^*,$$

where $B^0\Phi^*\sim\mathcal{N}(0,B^0C^*B^{0\prime})$. We can now apply Lemma 2 of Saracco (1997), and get $\sqrt{n}(\hat{b}_k-b_k)\longrightarrow_d R_k$, where $R_k=[b_k'\otimes(P_{E_k}P_EP_{E_k}-I_p)^+]B^0\begin{bmatrix}\mathrm{vec}(\Phi)\\ \mathrm{vec}(\Phi_\Sigma)\end{bmatrix}-(1/2)(b_k'\Phi_\Sigma b_k)b_k$. Tedious but simple computations give us for R_k a multivariate normal distribution with mean zero and covariance matrix $G_kC^*G_k'$, where the matrix G_k is

$$\begin{bmatrix}\{b_k'\otimes(P_{E_k}P_EP_{E_k}-I_p)^+\}(P_{E_k}'\otimes P_{E_k})C_w & \\ \{b_k'\otimes(P_{E_k}P_EP_{E_k}-I_p)^+\}((P_{E_k}P_E)'\otimes I_p+I_p\otimes P_{E_k}P_E)N_k-\frac{1}{2}b_k(b_k'\otimes b_k')\end{bmatrix}. \tag{4.3}$$

Step 3: Asymptotic distribution of $\hat{\tilde{b}}_k$.

Finally, since $\hat{\tilde{b}}_k=A_k'\hat{b}_k$ and $\tilde{b}_k=A_k'b_k$, we get $\sqrt{n}(\hat{\tilde{b}}_k-\tilde{b}_k)\longrightarrow_d \tilde{R}_j=A_k'R_k$, where $\tilde{R}_k\sim\mathcal{N}(0,A_k'G_kC^*G_k'A_k)$.

Remark. From a theoretical point of view, the asymptotic covariances of the two estimators can be estimated by replacing the theoretical terms by their empirical $\sqrt{n}$-consistent counterparts. The corresponding estimated asymptotic matrices converge to the true ones at rate $\sqrt{n}$. From a computational point of view, it is tedious to obtain these estimators of the asymptotic covariances. Nevertheless, we can easily compute bootstrap estimators that are very close to the true matrices (obtained by Monte-Carlo method). We illustrate this point in Section 5.1 on a simulated example.

5. Simulation Results

In order to evaluate the numerical performance of the proposed method, a simulation study was carried out. Following Duan and Li (1991), we measure the quality of the estimate $\hat{\tilde{b}}_k$ of the direction of $\tilde{\gamma}_k$ by

$$\cos^2\left(\hat{\tilde{b}}_k, \tilde{\gamma}_k\right) = \frac{\left(\hat{\tilde{b}}_k' \Sigma_k \tilde{\gamma}_k\right)^2}{(\hat{\tilde{b}}_k' \Sigma_k \hat{\tilde{b}}_k)(\tilde{\gamma}_k' \Sigma_k \tilde{\gamma}_k)},$$

where $\Sigma_k = A_k' \Sigma A_k$. The closer the squared cosine is to one, the better the estimation.

We generated simulated data from the semiparametric multivariate ($q = 2$) model (2.1) with

$$\begin{cases} g_1^{(1)}(\tilde{x}_1'\tilde{\gamma}_1, \varepsilon_1^{(1)}) = \exp(\tilde{x}_1'\tilde{\gamma}_1) + \varepsilon_1^{(1)} \\ g_2^{(1)}(\tilde{x}_2'\tilde{\gamma}_2, \varepsilon_2^{(1)}) = \tilde{x}_2'\tilde{\gamma}_2 + \varepsilon_2^{(1)} \\ g_1^{(2)}(\tilde{x}_1'\tilde{\gamma}_1, \varepsilon_1^{(2)}) = (\tilde{x}_1'\tilde{\gamma}_1)^3 + 3(\tilde{x}_1'\tilde{\gamma}_1) + \varepsilon_1^{(2)} \\ g_2^{(2)}(\tilde{x}_2'\tilde{\gamma}_2, \varepsilon_2^{(2)}) = (\tilde{x}_2'\tilde{\gamma}_2)^2 + \varepsilon_2^{(2)}, \end{cases} \tag{5.1}$$

where x follows a p-dimensional standardized normal distribution, and $\tilde{x}_1$ (resp. $\tilde{x}_2$) is the $(p-1)$-dimensional vector corresponding to the first (resp. last) $(p-1)$ coordinates of x. The error term $\varepsilon = (\epsilon_1^{(1)}, \epsilon_2^{(1)}, \epsilon_1^{(2)}, \epsilon_2^{(2)})'$ was normally distributed: $\epsilon \sim \mathcal{N}_4(\mu_\epsilon, \Sigma_\epsilon)$. Two designs of the covariance of ϵ were

$$\Sigma_\epsilon^I = \begin{pmatrix} 1 & \rho & 0 & 0 \\ \rho & 1 & 0 & 0 \\ 0 & 0 & 1 & \rho \\ 0 & 0 & \rho & 1 \end{pmatrix} \quad \text{and} \quad \Sigma_\epsilon^{II} = \begin{pmatrix} 1 & \rho & \rho & \rho \\ \rho & 1 & \rho & \rho \\ \rho & \rho & 1 & \rho \\ \rho & \rho & \rho & 1 \end{pmatrix},$$

with different values of ρ (0.1, 0.5 and 0.9). In the matrix Σ_ϵ^I the error terms associated with the components $y^{(1)}$ and $y^{(2)}$ were assumed to be independent, which is not the case with the covariance matrix Σ_ϵ^{II}. Note that we never considered the most favourable case with an independent error term between the observed equation and the selection equation. To control the number of non-observed values for the $y^{(j)}$ component, we used two different values of μ_ϵ in order to obtain around 25% (resp. 50%) of non-observed values for $y^{(1)}$ and $y^{(2)}$, we chose $\mu_\epsilon = (0, 1.5, 0, -0.5)$ (resp. $\mu_\epsilon = (0, 0, 0, -2)$). For the slope parameters, we took $\tilde{\gamma}_1 = (1, 1, -1, -1, 0, \ldots, 0)'$ and $\tilde{\gamma}_2 = (0, \ldots, 0, 1, -1, 1, -1)'$.

To study the performance of the proposed method, we considered different sample sizes (n = 100, 200 and 300), various dimensions of the explanatory

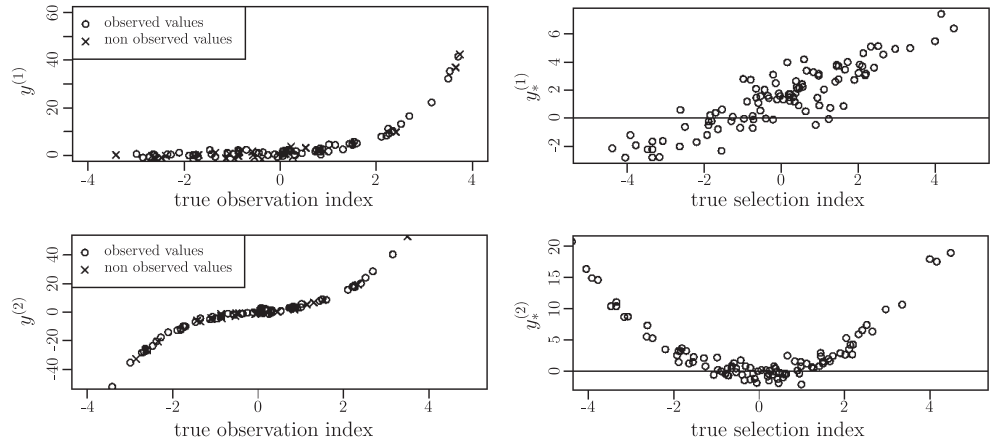

Figure 5.1. Plots of $y^{(j)}$ versus the true "observation" index $\tilde{x}_1'\tilde{\gamma}_1$ (on the left) and plots of the latent variables $y_*^{(j)}$ versus the true "selection" index $\tilde{x}_2'\tilde{\gamma}_2$ (on the right).

variable ($p = 5$, 10), the two different choices of covariance matrix (Σ_ϵ^I and Σ_ϵ^{II}), and two levels $\mathcal{L}$ of non-observed values for $y^{(j)}$ (25% and 50%). The number of slices in the PMS$_\alpha$ method, H_j, was specified to be $H_j = \max(\sqrt{n_j^*}, p)$, where n_j^* was the number of observed $y_i^{(j)}$'s in the sample.

Simulations were performed with R. All of the source codes are available from the authors by e-mail.

5.1. Simulated example

In this subsection, we consider the simulated sample with $n = 100$ for $p = 5$, $\Sigma_\epsilon = \Sigma_\epsilon^{II}$, $\rho = 0.5$ and $\mathcal{L} = 25\%$. On the left hand side of Figure 5.1 are plots of the response variables $y^{(1)}$ and $y^{(2)}$ versus the true "observation" index $\tilde{x}_1'\tilde{\gamma}_1$. Let us introduce the two latent variables $y_*^{(1)} = g_2^{(1)}(\tilde{x}_2'\tilde{\gamma}_2, \varepsilon_2^{(1)})$ and $y_*^{(2)} = g_2^{(2)}(\tilde{x}_2'\tilde{\gamma}_2, \varepsilon_2^{(2)})$; on the right hand side of Figure 5.1, we plot them versus the true "selection" index $\tilde{x}_2'\tilde{\gamma}_2$. The horizontal line allows us to determine for which observations the $y_i^{(j)}$'s values will be non-observed in the left hand side graphics.

The directions of $\tilde{\gamma}_1$ and $\tilde{\gamma}_2$ were estimated to get $\hat{\tilde{b}}_1 = (-0.483, -0.565, 0.447, 0.497)'$ and $\hat{\tilde{b}}_2 = (-0.613, 0.539, -0.350, 0.459)'$. The corresponding squared cosines were, respectively, equal to 0.993 and 0.962. Note that $\hat{\tilde{b}}_1$ (resp. $\hat{\tilde{b}}_2$) gave nearly the same direction as $\tilde{\gamma}_1$ (resp. $\tilde{\gamma}_2$). Moreover, we computed the quality of the estimation $\hat{E}$ of the e.d.r. space E using $\text{Trace}(P_E P_{\hat{E}})/2$, equal to 0.886 for this simulated sample. Even if this subspace was relatively poorly estimated compared with the quality of each estimated direction, the second step (which takes into account additional information) ensures that we recovered the

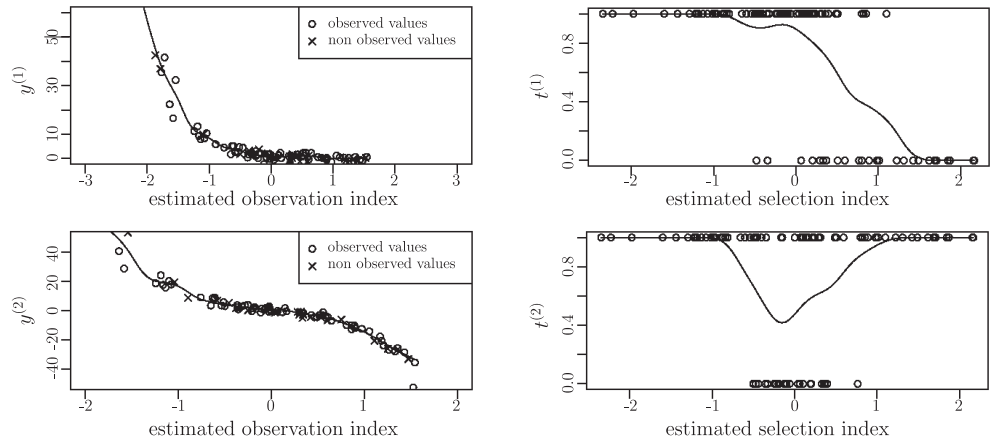

Figure 5.2. Kernel estimate of the observation link functions (left hand side) and Nadaraya-Watson estimate of the probability of $t^{(j)} = 1$ (that is $y^{(j)}$ observed).

good directions of the observation and selection slope vectors. We estimated the asymptotic covariance matrices, denoted by $\hat{V}(\hat{\tilde{b}}_1)$ and $\hat{V}(\hat{\tilde{b}}_2)$, with the bootstrap method (with 500 replications):

$$\hat{V}(\hat{\tilde{b}}_1) = 10^{-3} \begin{pmatrix} 4.17 & -1.93 & 1.74 & -0.90 \\ & 4.00 & 0.24 & 4.16 \\ & & 2.84 & -0.36 \\ & & & 17.1 \end{pmatrix},$$

$$\hat{V}(\hat{\tilde{b}}_2) = 10^{-2} \begin{pmatrix} 4.54 & -2.38 & -2.83 & 0.97 \\ & 4.36 & 2.66 & -1.52 \\ & & 4.22 & -0.58 \\ & & & 2.64 \end{pmatrix}.$$

These matrices are very close to the "true" asymptotic covariance matrices,$V(\hat{\tilde{b}}_1)$ and $V(\hat{\tilde{b}}_2)$ (not given here), calculated via the Monte Carlo approach. Note that the variance terms in $\hat{V}(\hat{\tilde{b}}_2)$ are greater than those obtained in $\hat{V}(\hat{\tilde{b}}_1)$, because of the low level $\mathcal{L}$ (=25%) of non-observed values for $y^{(j)}$.

In Figure 5.2, on the left hand side are plots of the response variable $y^{(j)}$ versus the estimated "observation" index $\tilde{x}_1'\hat{\tilde{b}}_1$. Note that, since we have $\hat{\tilde{b}}_1 \simeq -\tilde{\gamma}_1/||\tilde{\gamma}_1||$ (resp. $\hat{\tilde{b}}_2 \simeq -\tilde{\gamma}_2/||\tilde{\gamma}_2||$), the scatterplots of Figures 5.1 and 5.2 (left hand side) do not have the same orientation. We add on these plots the Nadaraya-Watson estimate of the observation link functions. On the right hand side, we plot the $t^{(j)}$'s values versus the estimated "selection" index $\tilde{x}_2'\hat{\tilde{b}}_2$, and we also plot

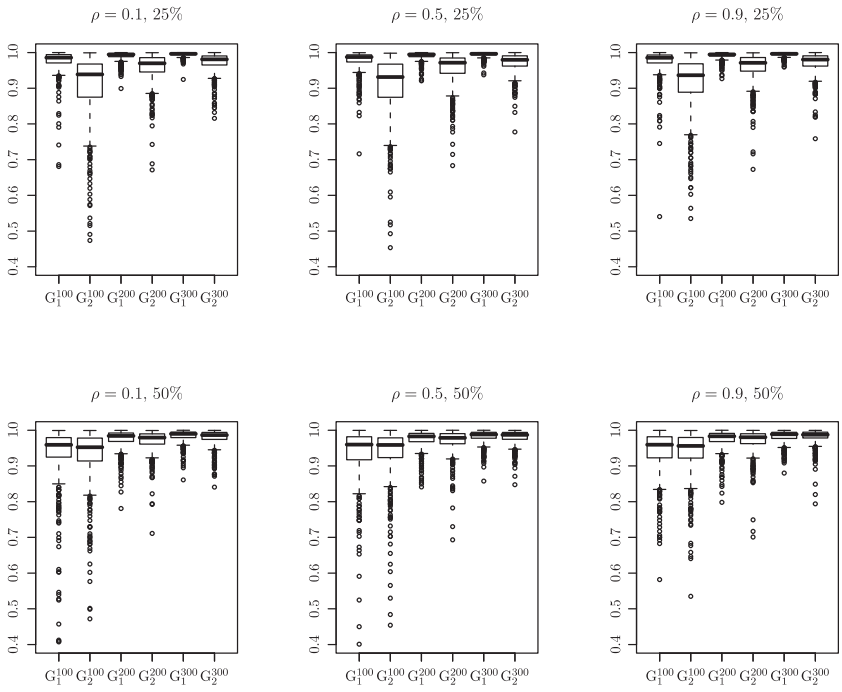

Figure 5.3. Boxplots of the squared cosines when $\Sigma_\epsilon = \Sigma_\epsilon^I$ and $p = 5$.

the Nadaraya-Watson estimate of the probability to observe $y^{(j)}$, based on (3.5).

5.2. Results of the simulation study

In our study, we considered combinations of the level $\mathcal{L}$ of non-observed values for $y^{(j)}$ (25% or 50%), the form of the error covariance matrix Σ_ϵ (Σ_ϵ^I or Σ_ϵ^{II} with $\rho = 0.1$, 0.5 or 0.9, and the dimension p of the covariable ($p = 5$ or 10). We also took into account the sample sizes $n = 100$, 200 or 300.

For each combination, $N = 500$ samples were generated. For each sample $l = 1, \ldots, N$, the directions of the slope vectors $\tilde{\gamma}_1$ and $\tilde{\gamma}_2$ were estimated and we got $\hat{\tilde{b}}_1^l$ and $\hat{\tilde{b}}_2^l$. Then, we evaluated the corresponding values of the quality measure: $c_k^l = \cos^2(\hat{\tilde{b}}_k^l, \tilde{\gamma}_k)$ for $k = 1, 2$, and $l = 1, \ldots, N$.

We show the results via the boxplots of squared cosines for different combinations. When $p = 5$ and $\Sigma_\epsilon = \Sigma_\epsilon^I$ (resp. $\Sigma_\epsilon = \Sigma_\epsilon^{II}$), Figure 5.3 (resp. Figure 5.4) gives the boxplots for $\tilde{\gamma}_1$ and $\tilde{\gamma}_2$, denoted by G1 and G2 in the graphics, for the different values of ρ, $\mathcal{L}$ and n. Figure 5.5 shows the boxplots when $p = 10$, $\Sigma_\epsilon = \Sigma_\epsilon^{II}$, and $n = 300$, for various ρ; note that the vertical scale in this figure goes from 0.75 to 1 (contrary to the previous one that goes from 0.4 to 1).

From Figures 5.3, 5.4 and 5.5, we can see that the results with these simulated data were very good. More precisely, one can observe that:

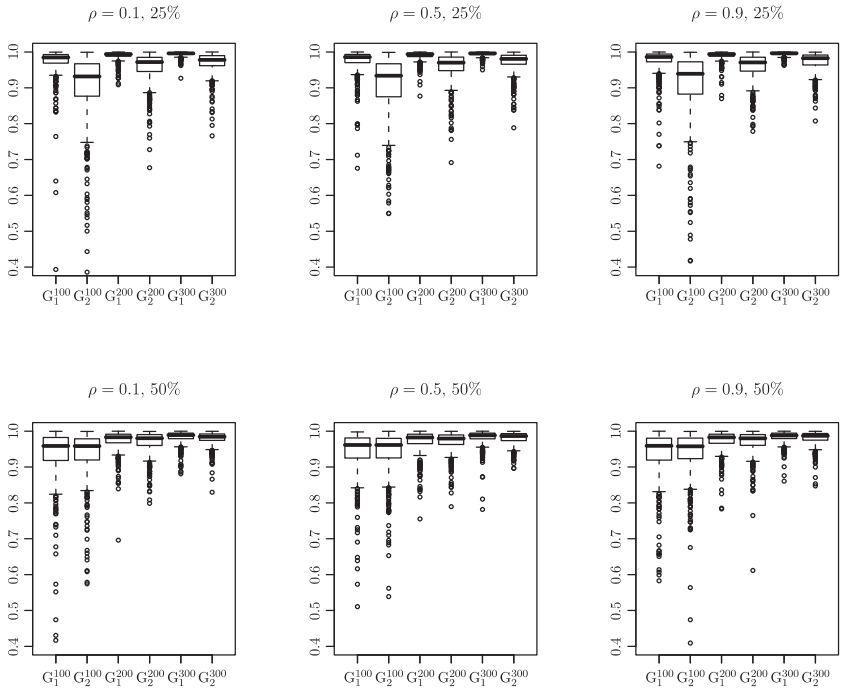

Figure 5.4. Boxplots of the squared cosines when $\Sigma_\epsilon = \Sigma_\epsilon^2$ and $p = 5$.

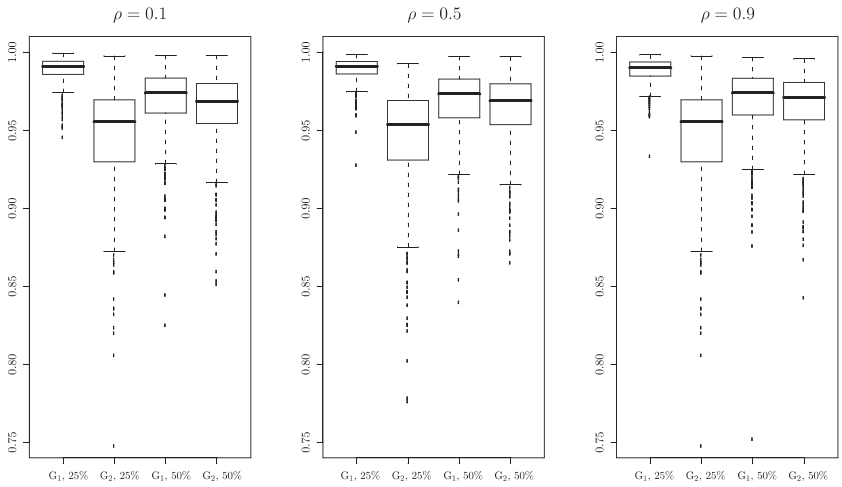

Figure 5.5. Boxplots of the squared cosines when $\Sigma_\epsilon = \Sigma_\epsilon^2$, $n = 300$ and $p = 10$.

- The estimations of the $\tilde{\gamma}_1$ and $\tilde{\gamma}_2$'s directions were good since almost all boxplots of the squared cosines were concentrated in the interval $[0.9, 1]$.
- The form of the covariance matrix of the error term ε and the value of the

parameter ρ did not seem to have any influence on the quality of the estimates.

- The level $\mathcal{L}$ of the non-observed values for the $y^{(j)}$'s had only a slight influence on the quality of the estimation of the selection slope vectors $\tilde{\gamma}_2$, especially in terms of spread of the squared cosine values. When this level was low ($\mathcal{L} = 25\%$), there was less information on the selection part of the model so the quality of the $\tilde{\gamma}_2$ estimates was slightly lower than when this level was larger ($\mathcal{L} = 50\%$). On the other hand, not surprisingly, there was an opposite behavior for the estimates of the observation slope parameter $\tilde{\gamma}_1$, since there is less information on the observation part of the model when $\mathcal{L}$ is large.
- The sample size n had a quite predictable influence on the quality of the estimates: the larger the sample size, the greater the squared cosines. When $n = 200$ or 300, the quality of the two estimated directions was very good.
- Dimension p of the explanatory variable x did not seem to have any effect on the quality of the estimates.

5.3. Simulation with a non-normal distributed covariable x

In order to investigate the robustness of the method when x does not follow a multivariate normal distribution, we generated each component of x from various distributions (far from the normal distribution): discrete rectangular distribution on $\{1, \ldots, 4\}$, continuous rectangular distribution on $[0, \sqrt{12}]$, binomial distribution $\mathcal{B}(4, 0.2)$. We did not change either the form or the other parameters ($p = 5$, $\Sigma_\epsilon = \Sigma_\epsilon^{II}$, with various values of ρ) of the simulated model described at the beginning of Section 5. In order to control the level $\mathcal{L}$, we used different values for μ_ϵ to obtain around 25% (resp. 50%) of non-observed values for the $y^{(1)}$'s and the $y^{(2)}$'s. Moreover we took $\tilde{\gamma}_1 = (1, 1, -1, -1)'/2$ and $\tilde{\gamma}_2 = (1, -1, 1, -1)'/2$ for the observation and selection slope parameters.

In each case, $N = 500$ samples of size $n = 200$ were generated, and for each simulated sample, the directions of $\tilde{\gamma}_1$ and $\tilde{\gamma}_2$ were estimated with the proposed method. Then the corresponding squared cosines were calculated. Figure 5.6 reports the results of this simulation study via the boxplots of these squared cosines for the discrete rectangular distribution. One can see that the estimations of the directions of the slopes for the selection equations and the outcome equations were quite good, even for a discrete x that does not follow an elliptically symmetric distribution. Note that, as in the multivariate normal case (in the previous subsection), we can observe the same influence of the level $\mathcal{L}$ on the quality of the estimates and no influence of parameter ρ. Very similar results (not detailed here) were observed for the two other distributions.

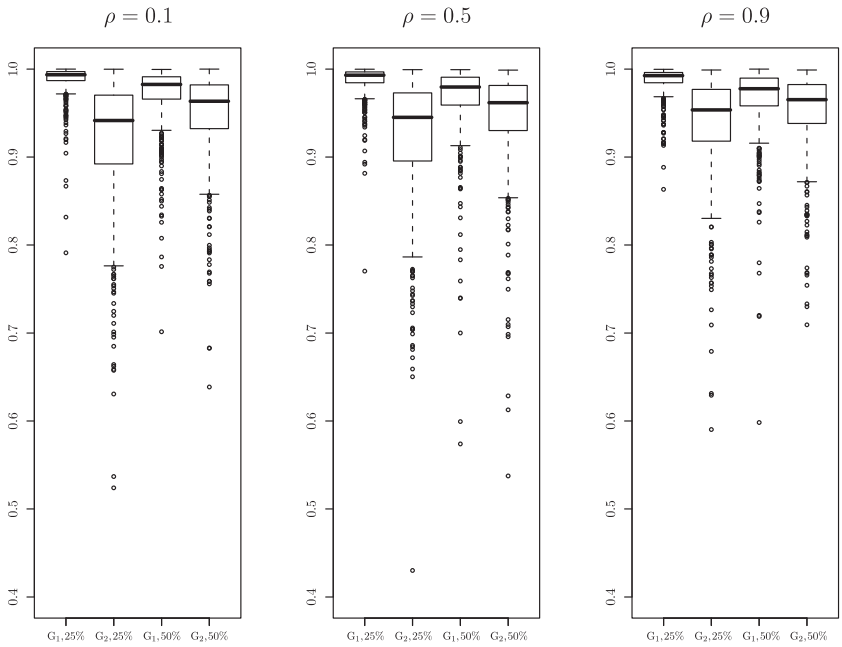

Figure 5.6. Boxplot of the squared cosines when x follows a (discrete) rectangular distribution on $\{1, \ldots, 4\}$.

5.4. Comparison with a parametric approach

We compared in this simulation the parametric Tobit II model (implemented in Henningsen and Toomet (2008)) with our semiparametric approach using various error distributions and various selection and observations link functions. We considered two models, $(M1)$ and $(M2)$, from the sample models defined in (2.1) with $q = 1$:

$$(M1): \begin{cases} g_1(\tilde{x}_1^{'}\tilde{\gamma}_1, \epsilon_1) = \tilde{x}_1^{'}\tilde{\gamma}_1 + \epsilon_1 \\ g_2(\tilde{x}_2^{'}\tilde{\gamma}_2, \epsilon_2) = \tilde{x}_2^{'}\tilde{\gamma}_2 + \epsilon_2 \end{cases} \text{ and } (M2): \begin{cases} g_1(\tilde{x}_1^{'}\tilde{\gamma}_1, \epsilon_1) = \exp(\tilde{x}_1^{'}\tilde{\gamma}_1) + \epsilon_1 \\ g_2(\tilde{x}_2^{'}\tilde{\gamma}_2, \epsilon_2) = \exp(\tilde{x}_2^{'}\tilde{\gamma}_2) + \epsilon_2 \end{cases}.$$

Model $(M1)$ is in favour of the parametric approach with linear link functions, whereas model $(M2)$ has non-linear link functions. For these two models, the error term $\epsilon = (\epsilon_1, \epsilon_2)$ was normally distributed as in the previous simulation study, x followed a five-dimensional standardized normal distribution, $\tilde{x}_1$ (resp. $\tilde{x}_2$) the 4-dimensional vector corresponding to the first (resp. last) four coordinates of x. To control the level $\mathcal{L}$ of non-observed values for y, we used different values of μ_ϵ. For the slope parameters , we took $\tilde{\gamma}_1 = (1, 1, -1, -1)^{'}$ and $\tilde{\gamma}_2 = (1, -1, 1, -1)^{'}$.

We present in Figure 5.7 only the results for $n = 200$, $\rho = 0.9$ and $\mathcal{L} = 50\%$, over $N = 500$ replicated samples. As expected, the Tobit II approach performed poorly for model $(M2)$ only for the outcome equation, not for the

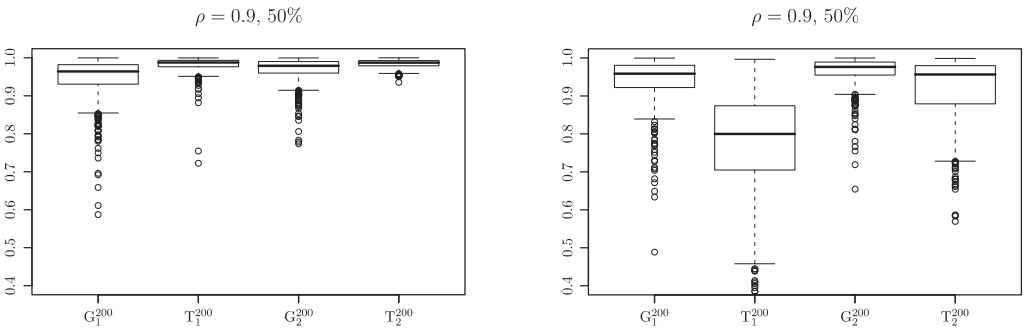

Figure 5.7. Boxplot of the squared cosines $n = 200$, where the notation G (resp. T) is used for our proposed estimators (resp. Tobit II estimators).

selection equation, and the proposed method was somewhat inferior to Tobit II approach for model $(M1)$.

We also considered various combinations of the simulation parameters: the level $\mathcal{L}$ of non-observed values for $y^{(j)}$ (25% or 50%), the error term correlation $\rho = 0.1$, 0.5 or 0.9. In any case, we observed very similar results. Moreover, we compared the two approaches with the linear model $(M1)$ when the error term was non-normally distributed. The Tobit II method appears to be robust to mild violations of the normality assumption like our approach (which does not rely on this kind of assumption).

6. Concluding Remarks

We have proposed a semi-parametric estimation method for a multivariate sample selection model (MSSM). The proposed geometric approach to the estimation of the slope vectors in the outcome equation and in the selection models has the advantage of dealing symmetrically with both slope vectors. The convergence in probability at root n rate, and the asymptotic normality of the slope estimators, has been established. This estimation method is numerically very fast since it is based on only a few matrix operations and eigen-decompositions and does not demand any time-consuming iterative computations. The corresponding algorithm is easy to implement and the R source code is available from the authors. From a practical point of view, a simulation study has highlighted good behaviour of the estimation method even for non-elliptical distribution of the covariate. Finally an interesting thought direction would be to develop another two-step semi-parametric estimation methods: in a first step, we could take into account the MSIR estimator of the selection slope parameter since the selection probability only depends on the index $\tilde{x}_2'\tilde{\gamma}_2$; in a second step, we could incorporate the additional information in order to get the observation slope vector from the entire e.d.r. space.

Acknowledgement

The authors are very grateful to the Editor, an associate editor, and the three referees for their valuable comments and constructive suggestions. They thank Jean Belin, economics researcher of the GREThA laboratory of Bordeaux 4 University, for numerous discussions and future work on applications in economics of the proposed semiparametric multivariate sample selection model.

Appendix: Asymptotic Distribution of $\hat{P}_{E_k}$

Let $\hat{P}_{E_k} = A_k(A_k'\hat{\Sigma}A_k)^{-1}A_k'\hat{\Sigma}$ (resp. $P_{E_k} = A_k(A_k'\Sigma A_k)^{-1}A_k'\Sigma$) be the $\hat{\Sigma}$ (resp. Σ) orthogonal projector onto the linear subspace E_k spanned by the columns of A_k.

For an elliptically distributed x, with covariance matrix Σ and kurtosis parameter κ, Tyler (1981) gave the following asymptotic distribution: $\sqrt{n}(\hat{\Sigma} - \Sigma) \longrightarrow_d \Phi_\Sigma$ where $\text{vec}(\Phi_\Sigma) \sim N(0, C_\Sigma)$ and $C_\Sigma = (1+\kappa)(I_{p^2} + K_p)(\Sigma \otimes \Sigma) + \kappa \text{vec}(\Sigma)[\text{vec}(\Sigma)]'$. K_p is the $p^2 \times p^2$ commutation matrix (see Magnus and Neudecker (1979)).

We obtain the asymptotic distribution of $\hat{P}_{E_k}$ through the following three steps.

Step 1. Let $f_1 : \mathbb{R}^{p^2} \longrightarrow \mathbb{R}^{p_k p + p_k^2}$ be defined by

$$f_1(\text{vec}(M)) = \begin{bmatrix} (I_p \otimes A_k')\text{vec}(M) \\ (A_k' \otimes A_k')\text{vec}(M) \end{bmatrix}.$$

Then, from the Delta method, we get:

$$\sqrt{n}\left(\begin{bmatrix} \text{vec}(A_k'\hat{\Sigma}) \\ \text{vec}(A_k'\hat{\Sigma}A_k) \end{bmatrix} - \begin{bmatrix} \text{vec}(A_k'\Sigma) \\ \text{vec}(A_k'\Sigma A_k) \end{bmatrix} \right) \longrightarrow_d U_{1k},$$

where $U_{1k} \sim \mathcal{N}(0, C_{1k})$ with $C_{1k} = \begin{bmatrix} I_p \otimes A_k' \\ A_k' \otimes A_k' \end{bmatrix} C_\Sigma \begin{bmatrix} I_p \otimes A_k & A_k \otimes A_k \end{bmatrix}$.

Step 2. From the following first order approximation:

$$\begin{aligned} &\sqrt{n}((A_k'\hat{\Sigma}A_k)^{-1} - (A_k'\Sigma A_k)^{-1}) \\ &\doteq -(A_k'\Sigma A_k)^{-1}\left[\sqrt{n}(A_k'\hat{\Sigma}A_k - A_k'\Sigma A_k)\right](A_k'\Sigma A_k)^{-1}, \end{aligned}$$

we derive:

$$\sqrt{n}\left(\begin{bmatrix} \text{vec}(A_k'\hat{\Sigma}) \\ \text{vec}((A_k'\hat{\Sigma}A_k)^{-1}) \end{bmatrix} - \begin{bmatrix} \text{vec}(A_k'\Sigma) \\ \text{vec}((A_k'\Sigma A_k)^{-1}) \end{bmatrix} \right) \longrightarrow_d U_{2k} = S_k U_{1k},$$

where $U_{2k} \sim N(0, C_{2k})$ with $C_{2k} = S_k C_{1k} S_k'$ and

$$S_k = \begin{bmatrix} I_{p_k p} & 0_{p_k p + p_k^2} \\ 0_{p_k^2 + p_k p} & -(A_k'\Sigma A_k)^{-1} \otimes (A_k'\Sigma A_k)^{-1} \end{bmatrix}.$$

STEP 3. Let us introduce the function $f_2 : \mathbb{R}^{p_k p + p_k^2} \longrightarrow \mathbb{R}^{p^2}$ defined by $f_2 \begin{pmatrix} \text{vec}(M_1) \\ \text{vec}(M_2) \end{pmatrix} = \text{vec}(A_k M_2 M_1)$. Then from a second application of the Delta method, we derive:

$$\text{vec}(\sqrt{n}[\hat{P}_j - P_{E_k}]) \longrightarrow_d U_k,$$

where $U_k \sim N(0, C_{U_k})$ with $C_{U_k} = N_k C_\Sigma N_k'$ and

$$\begin{aligned} N_k &= I_p \otimes A_k(A_k'\Sigma A_k)^{-1}A_k' - P_{E_k}' \otimes A_k(A_k'\Sigma A_k)^{-1}A_k' \\ &= (I_p - P_{E_k}') \otimes [A_k(A_k'\Sigma A_k)^{-1}A_k']. \end{aligned} \tag{A.1}$$

References

Ahn, H. and Powell, J. (1993). Semiparametric estimation of censored selection models with a nonparametric selection mechanism. *J. Econometrics* **58**, 3-29.

Amemiya, T. (1985). *Advanced Econometrics*. Basil Blackwell, Oxford.

Aragon, Y. (1997). A Gauss implementation of multivariate sliced inverse regression. *Comput. Statist.* **12**, 355-372.

Barreda, L., Gannoun, A. and Saracco, J. (2007). Some extensions of multivariate sliced inverse regression. *J. Statist. Comput. Simulation* **77**, 1-17.

Blundell, R. W. and Smith, R. J. (1993). Simultaneous microeconometric models with censored or qualitative dependent variables. In *Handbook of Statistics* **1** (Edited by G. S. Maddala, C. R. Rao and H. D. Vinod), 117-143. North-Holland.

Chen, C. and Li, K. C. (1998). Can SIR be as popular as multiple regression? *Statist. Sinica* **8**, 289-316.

Diaconis, P. and Freedman, D. (1984). Asymptotics of graphical projection pursuit. *Ann. Statist.* **12**, 793-815.

Duan, N. and Li, K. C. (1987). Distribution-free and link-free estimation method for the sample selection model. *J. Econometrics* **53**, 25-35.

Duan, N. and Li, K. C. (1991). Slicing regression: a link-free regression method. *Ann. Statist.* **19**, 505-530.

Eiswerth, M. E. and Shonkwiler, J. S. (2006). Examining post-wildfire reseeding on arid rangeland: A multivariate tobit modelling approach. *Ecological Modelling* **192**, 286-298.

Eubank, R. (1988). *Spline Smoothing and Nonparametric Regression.* Dekker, New York.

Goldberger, A. S. (1983). Abnormal selection bias. In *Studies in Econometrics, Times Series, and Multivariate Statistics.* (Edited by S. Karlin, T. Amemiya and L.A. Goodman). Academic Press, New York.

Haerdle, W. (1990). *Applied Nonparametric Regression*. Cambridge University Press.

Hall, P. and Li, K. C. (1993). On almost linearity of low dimension projection from high dimensional data. *Ann. Statist.* **21**, 867-889.

Heckman, J. (1979), Sample selection bias as a specification error. *Econometrica* **47**, 153-161.

Henningsen, A. and Toomet, O. (2008). sampleSelection - A package for the free Statistical Software "R" for estimating sample selection models. Published on: CRAN (The Comprehensive R Archive Network). Further information is available at `http://www.sampleSelection.org/`.

Lee, L. (1994). Semiparametric two-stage estimation of sample selection models subject to Tobit-type selection rules. *J. Econometrics* **61**, 305-344.

Li, K. C. (1991). Sliced inverse regression for dimension reduction, with discussions. *J. Amer. Statist. Assoc.* **86**, 316-342.

Li, K. C. (1997). Nonlinear confounding in high-dimensional regression. *Ann. Statist.* **25**, 577-612.

Li, K. C., Aragon, Y., Shedden, K. and Thomas Agnan, C. (2003). Dimension reduction for multivariate response data. *J. Amer. Statist. Assoc.* **98**, 99-109.

Maddala, G. S. (1983). *Limited-dependent and Qualitative Variables in Econometrics*. Cambridge University Press, Cambridge.

Maddala, G. S. (1993). Estimation of limited-dependent variable models under rational expectations. In *Handbook of Statistics*, **11** (Edited by G. S. Maddala, C. R. Rao and H. D. Vinod), 175-194. North-Holland.

Magnus, J. R. and Neudecker, H. (1979). The commutation matrix: some properties and applications. *Ann. Statist.* 381-394.

Melenberg, B. and van Soest, A. (1993). Semi-parametric estimation of the sample selection model. Discussion Paper 9334, CentER, Tilburg University.

Newey, W. K. (1991). Two-step series estimation of sample selection models. Manuscript (Department of Economics, MIT, Cambridge, MA).

Saracco, J. (1997). An asymptotic theory for Sliced Inverse Regression. *Comm. Statist. Theory Methods* **26**, 2141-2171.

Saracco, J. (2001). Pooled slicing methods versus slicing methods. *Comm. Statist. Theory Methods* **30**, 489-511.

Saracco, J. (2005). Asymptotics for pooled marginal slicing estimator based on SIR_α approach. *J. Multivariate Anal.* **96**, 117-135.

Tyler, D. E. (1981). Asymptotic inference for eigenvectors. *Ann. Statist.* **9**, 725-736.

CNRS, UMR 5251, Bordeaux, F-33000, France.

INRIA, research project team CQFD, Bordeaux, F-33000, France.

E-mail: Marie.Chavent@math.u-bordeaux1.fr

INSERM U897, Equipe de Biostatistique, ISPED, Université Victor Ségalen - Bordeaux 2, 146 rue Léo Saignat, 33076 Bordeaux Cedex, France.

E-mail: Benoit.Liquet@isped.u-bordeaux2.fr

GREThA, Université Montesquieu - Bordeaux IV, Avenue Léon Duguit, 33608 Pessac Cedex, France.

E-mail: jerome.saracco@math.u-bordeaux1.fr

(Received November 2007; accepted January 2009)