Distribution-free and link-free estimation for a multivariate semiparametric sample selection model

Marie Chavent, Jérôme Saracco

HAL Id: hal-00278290

https://hal.archives-ouvertes.fr/hal-00278290

Submitted on 10 May 2008

Distribution-free and link-free estimation for a multivariate semiparametric sample selection model[⋆]

Marie Chavent[1] and Jérôme Saracco[1,2]

[1] Institut de Mathématiques de Bordeaux
    Université Bordeaux 1
    351 cours de la libération
    33405 TALENCE Cedex
    (e-mail: `Marie.Chavent@math.u-bordeaux1.fr`,
    `Jerome.Saraccomath.u-bordeaux1.fr`)
[2] GREThA,
    Université Montesquieu - Bordeaux IV
    Avenue Léon Duguit
    33608 PESSAC Cedex
    (e-mail: `Jerome.Saracco@u-bordeaux4.fr`)

Abstract. Most of the prevalent estimation methods for sample selection model rely heavily on parametric assumptions. We consider in this communication a multivariate semiparametric sample selection model and we 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. 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 give results from a simulation study in order to illustrate the estimation method.
Keywords: Multivariate Sliced Inverse Regression, Canonical Analysis, Semiparametric Regression Models.

1 Introduction

In this communication, we consider sample selection models (SSM). Basically they are described by two equations. A selection equation gives the state (missing / non missing) of the dependent variable y as a function of explanatory variables, and an outcome equation gives the value of the multivariate dependent variable, when observed, as another function of some explanatory variables x. Numerous papers dealing with univariate SSM have been published. The adjective "univariate" refers to $y \in \Re$. Here, we focus on multivariate SSM, that is when $y \in \Re^q$, $q > 1$.

[⋆] Proceedings of the XIIth International Symposium on Applied Stochastic Model and Data Analysis, Chania (2007)

Let us first give a brief overview of univariate SSM. When the dependent variable is univariate, [Heckman, 1979] introduced what is now regarded as the prototype selection model. [Amemiya, 1985] refers to this model as the type II Tobit Model:

$$
\begin{aligned}
(E1) \ &: y_{1i}^* = x_{1i}'\beta_1 + \varepsilon_{1i} \\
(E2) \ &: y_{2i}^* = x_{2i}'\beta_2 + \varepsilon_{2i} \\
(E3) \ &: y_{2i} = \mathbb{I}[y_{2i}^* > 0] \\
(E4) \ &: y_{1i} = y_{1i}^* y_{2i} \\
(E5) \ &: (\varepsilon_{1i}, \varepsilon_{2i})'|x_i \sim N(0, \Sigma), \quad \Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix}
\end{aligned}
$$

The observed variables are y_{1i}, y_{2i} and $x_i = (x_{1i}', x_{2i}')'$. Equation $(E3)$ is the selection equation, and equation $(E4)$ is the outcome equation. Maximum Likelihood method is generally used to estimate such models. The score equation 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 the 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 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, one may read [Amemiya, 1985], [Maddala, 1983], [Maddala, 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 the error specification. [Melenberg and van Soest, 1993] give a panorama of the semiparametric estimation methods for SSM. Most semiparametric estimation techniques of the SSM proceed in two stages. The first one gives a consistent estimate of the slope of the selection equation. Indeed, this equation can, on its own, be considered as a Probit Model. The second stage works with the non missing y's only, (i) building a biased estimate of the slope of the outcome equation, and (ii) correcting for this bias with the help of the first step estimated slope. [Duan and Li, 1987], [Ahn and Powell, 1993], [Lee, 1994] follow such a scheme. [Ichimura and Lee, 1991] estimate the two slopes simultaneously in a reasonable time, but calculating the asymptotic covariance of the estimators requires lengthy computations.

In this communication, we study a multivariate semiparametric SSM. We propose a geometric approach to the estimation of the slopes of the outcome and selection equations. As [Duan and Li, 1987] in the univariate case, we do not need any assumption about the link functions or the error distribution. Moreover, contrary to most existing methods, we deal symmetrically with both slopes. The method works in two steps. The first one performs a Multivariate Sliced Inverse Regression (MSIR) analysis. The second step converts the MSIR indices to estimates of the slopes by means of two Canon-

ical Analyses. The corresponding numerical algorithm is fast and does not require starting values.

2   A semiparametric multivariate sample selection model

We consider the following semiparametric multivariate sample selection model:

$$y = \begin{cases} g_1(\tilde{x}_1^{'e}\tilde{\gamma}_1, \varepsilon_1) & \text{if } g_2(\tilde{x}_2^{'e}\tilde{\gamma}_2, \varepsilon_2) > 0 \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

where:

- The dependent variable is multivariate: $y \in \Re^q$. The value "0" for y in equation (1) symbolically indicates a missing (non observed) value. Another symbol might be used to avoid any confusion with an observed value, 0, of y.
- The functions g_1 and g_2 are unknown link functions, g_1 is called the observation link function and g_2 the selection link function.
- The variables $\tilde{x}_1 \in \Re^{p_1}$ and $\tilde{x}_2 \in \Re^{p_2}$ are subvectors of the random vector $x \in \Re^p$, assumed to be elliptically distributed with parameters $\mu = \mathbb{E}[x]$ and $\mathbb{V}(x) = \Sigma$.
  Let A_j', $j = 1, 2$ be the matrices which select the components of $\tilde{x}_j$, $j = 1, 2$ in x. A_j is $p \times p_j$, $2 \leq p_j < p$. A_j has exactly one "1" in each column and at most one "1" in each row, and its other elements are "0". So A_j is of full column rank and such that $\tilde{x}_j = A_j'x$.
  It follows that $\tilde{x}_1$ and $\tilde{x}_2$ are elliptically distributed with parameters $\mu_j = A_j'\mu$, $j = 1, 2$ and $\Sigma_j = A_j'\Sigma A_j$, $j = 1, 2$. Moreover, from the definition of A_j's we have: $A_j'A_j = I_{p_j}$.
- The couple $(\varepsilon_1, \varepsilon_2)$ is a random error vector independent of x with an unknown distribution.
- The slope parameters $\tilde{\gamma}_1 \in \Re^{p_1}$ and $\tilde{\gamma}_2 \in \Re^{p_2}$ are two unknown vectors.

In this model, our main purpose is to estimate the directions of the slopes $\tilde{\gamma}_1$ and $\tilde{\gamma}_2$. Then, the link function g_1 and the state of y (missing/non missing) probabilities can be nonparametrically estimated.

One can observe that model (1) is a particular case of a multivariate two indices semiparametric regression models of the form

$$y = f(x^{'e}\beta_1, x^{'e}\beta_2, \varepsilon). \tag{2}$$

Model (2) has been introduced by Li (1991) when $y \in \Re$. Li (1991) proposed the sliced inverse regression in order to estimate the subspace of $\Re^p$, spanned by β_1 and β_2, which is called the e.d.r. (effective dimension reduction) space.

Some extensions of the SIR approach to multivariate y have been studied by [Aragon, 1997], [Li *et al.*, 2003], [Saracco, 2005], or [Barreda *et al.*, 2006].

We are going to consider model (1) as a particular kind of model (2), with extra information about the e.d.r. space, namely, structural zeros in the coefficients. To do this, we define $\gamma_j = A_j \tilde{\gamma}_j \in \Re^p$, $j = 1, 2$, that is we expand $\tilde{\gamma}_j$ to a $p \times 1$ vector with zeros corresponding to the non-selected components. So model (1) is writen:

$$y = g(x^{'e}\gamma_1, x^{'e}\gamma_2, \varepsilon) \tag{3}$$

with $\varepsilon = (\varepsilon_1, \varepsilon_2)$ and $g(t, u, e) = g_1(t, e_1)\mathbb{I}[g_2(u, e_2) > 0]$, where $e = (e_1, e_2)$ and $\mathbb{I}[.]$ denotes the indicator function. Let us define $E = \text{Span}(\gamma_1, \gamma_2) \subset \Re^p$. Without additional conditions, we have $\dim(E) \leq 2$. If γ_1 and γ_2 are linearly independent, then $\dim(E) = 2$, and $\{\gamma_1, \gamma_2\}$ determines a basis of the e.d.r. space.

Let us assign *identifiability conditions* which ensure that we are working on a two indices model (that is $\dim(E) = 2$):

(*i*) Each vector x_j, $j = 1, 2$, has at least an x-component not present in the other $x_{j'}$, $j' \neq j$; such a component could be called j-specific.
(*ii*) At least one component of γ_j among the j-specific component is non null, $j = 1, 2$.

We now bring these conditions into a geometric perspective. Let $E_j = \text{Span}(A_j)$, $E_j \subset \Re^p$, $\dim(E_j) = p_j$. The identifiability conditions are:

(*i*) $E_1 \not\subset E_2$ and $E_2 \not\subset E_1$,
(*ii*) $E \cap E_1 \neq E$ and $E \cap E_2 \neq E$.

Let us consider more closely the linear subspace $E \cap E_j$. Since $\dim(E) = 2$, $\dim(E \cap E_j) \leq 2$. From the definition of E and E_j, $\gamma_j \in E \cap E_j$, thus $\dim(E \cap E_j) \geq 1$. But because of the identifiability conditions, for $j^{'e} \neq j$, $\gamma_{j'e} \in E$ and $\gamma_{j'e} \notin E_j$, thus $\gamma_{j'e} \notin E \cap E_j$ and $\dim(E \cap E_j) < 2$. Finally, $\dim(E \cap E_j) = 1$ and $E \cap E_j \subset \Re^p$ is spanned by γ_j.

3   Population and sample approaches

The idea is to use multivariate sliced inverse regression in order to find another basis of $E = \text{Span}(\gamma_1, \gamma_2)$, the e.d.r. space. Let us call this basis $\{b_1, b_2\}$. These vectors are Σ-orthogonal. Since the matrices A_1 and A_2 are known (and so are the subspaces E_1 and E_2), two canonical analysis of the couples (E, E_1) and (E, E_2) give us bases of $E \cap E_1$ and $E \cap E_2$.

3.1 Population version

- For model (3), [Saracco, 2005] has shown that pooled marginal sliced inverse regression provides a basis denoted $\{b_1, b_2\}$ of the e.d.r. space E, the only novelty is to consider a transformation (slicing) $T(.)$ which does not modify the missing y value. The vector b_j are the eigenvectors corresponding to the two largest eigenvalues of $\Sigma^{-1}M_T$. Lest us denote $B = [b_1, b_2]$, we have $\mathrm{Span}(B) = E$.
- Let us consider two subspaces F and G of $\Re^p$ equipped with the inner product Σ. Canonical analysis is a useful tool to find out a Σ-orthogonal basis of $F \cap G$. This basis is formed by the eigenvectors corresponding to the eigenvalue 1 of $P_F P_G$, where P_F and P_G are the Σ-orthogonal projectors onto F and G. Specifically, we take $F = E$ and $G = E_j$, $j = 1, 2$. Thus, $P_E = B(B'\Sigma B)^{-1}B'\Sigma = BB'\Sigma$ and and $P_{E_j} = A_j(A_j'\Sigma A_j)^{-1}A_j'\Sigma$. It is equivalent and simpler to diagonalize $P_{E_j}P_E P_{E_j}$ which is a Σ-symmetric matrix. Let us call v_j the unique eigenvector corresponding to the eigenvalue 1; v_j is colinear to γ_j and is normalized: $v_j'\Sigma v_j = 1$. We next derive a vector, $\tilde{v}_j$, colinear to $\tilde{\gamma}_j$:

$$\tilde{v}_j = A_j' v_j.$$

This vector $\tilde{v}_j$ is normalized: $\tilde{v}_j'\Sigma_j\tilde{v}_j = 1$ where $\Sigma_j = A_j'\Sigma A_j$.

3.2 Estimation of the directions

As was precised in the former section, the directions are obtained from computations based only on covariance matrices. Substituting estimates in place of these matrices yields estimated directions.

Let $\{(y_i, x_i),\ i = 1, ..., n\}$ be a sample from the reference model (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 pooled marginal sliced inverse regression.** We can apply the usual procedure in order to obtain $\widehat{M_T}$, the estimates of the matrix M_T. The only constraint is about the slicing of each y component. Let $H_j + 1$ be a fixed number of slices for the jth component of y. One of them, say s_0^j, contains the cases with $y = 0$ (a missing value of the jth component of y), the other slices, s_h^j, $h = 1, ..., H_j$, are made by splitting the range of the non-missing values of the jth component of y into slices of nearly equal weight.

  The two estimated e.d.r. directions, $\hat{b}_1$ and $\hat{b}_2$, are the eigenvectors corresponding to the two largest eigenvalues of $\hat{\Sigma}^{-1}\widehat{M_T}$. These vectors form a $\hat{\Sigma}$-orthonormal basis of the estimated e.d.r. space $\hat{E} = \mathrm{Span}(\hat{B})$ where $\hat{B} = [\hat{b}_1, \hat{b}_2]$.

- **Step 2: Estimating the direction of** γ_j, $j = 1, 2$. We get these directions by the canonical analyses of $(\hat{E}, E_1)$ and $(\hat{E}, E_2)$. Let $\hat{v}_j$ be the eigenvector associated with the largest eigenvalue of $\hat{P}_{E_j} \hat{P}_{\hat{E}} \hat{P}_{E_j}$, where $\hat{P}_{\hat{E}} = \hat{B}(\hat{B}'^e \hat{\Sigma} \hat{B})^{-1} \hat{B}'^e \hat{\Sigma} = \hat{B}\hat{B}'^e \hat{\Sigma}$ and $\hat{P}_{E_j} = A_j(A_j'^e \hat{\Sigma} A_j)^{-1} A_j \hat{\Sigma}$. An estimation of the direction of $\tilde{\gamma}_j$ is given by

$$\hat{\tilde{\gamma}}_j = A_j'^e \hat{v}_j.$$

Asymptotics. With classical asymptotic theory, we can obtain the convergence in probability of the estimated directions to the true directions at rate $n^{-1/2}$:

$$\hat{\tilde{\gamma}}_j = v_j + O_p(1/\sqrt{n}), \ j = 1, 2.$$

Asymptotic normality of $\hat{\tilde{\gamma}}_j$, $j = 1, 2$ can be also derived from the asymptotic distribution of the canonical analysis matrix (that is $\sqrt{n}(\hat{P}_{E_j} \hat{P}_{\hat{E}} \hat{P}_{E_j} - P_{E_j} P_E P_{E_j})$) and the asymptotic distribution of the corresponding two major eigenvectors. The estimates of the asymptotic covariances of the estimators may be obtained through standard matrix calculus.

Simulation results. In order to evaluate the numerical behaviour of our approach, we conduct a simulation study with sample sizes $n = 100$ and 300 and percentage of missing values of y around 50%. The quality of the estimates has been measured by the square cosine between the true direction slope and its estimate.

References

[Ahn and Powell, 1993]H. Ahn and J. Powell. Semiparametric estimation of censored selection models with a nonparametric selection mechanism. *Journal of Econometrics*, 58:3–29, 1993.

[Amemiya, 1985]T. Amemiya. *Advanced econometrics*. Basil Blackwell, Oxford, 1985.

[Aragon, 1997]Y. Aragon. A gauss implementation of multivariate sliced inverse regression. *Computational Statistics*, 12:355–372, 1997.

[Barreda *et al.*, 2006]L. Barreda, A. Gannoun, and J. Saracco. Some extensions of multivariate sir. *Journal of Statistical Computation and Simulation*, To appear, 2006.

[Blundell and Smith, 1993]R.W. Blundell and R.J. Smith. Simultaneous microeconometric models with censored or qualitative dependent variables. *In: G. S. Maddala, C. R. Rao and H. D. Vinod, eds., Handbook of statistics*, 11:117–143, 1993.

[Duan and Li, 1987]N. Duan and K.C. Li. Distribution-free and link-free estimation method for the sample selection model. *Journal of Econometrics*, 53:25–35, 1987.

[Goldberger, 1983]A.S. Goldberger. Abnormal selection bias. *In: S. Karlin, T. Amemiya and L.A. Goodman, eds., Studies in econometrics, times series, and multivariate statistics*, 1983.

[Heckman, 1979]J. Heckman. Sample selection bias as a specification error. *Econometrica*, 47:153–161, 1979.

[Ichimura and Lee, 1991]H. Ichimura and F. Lee. Semiparametric least squares of multiple index models: Single equation estimation. *In: W.A. Barnett et al., eds., Nonparametric and semiparametric methods in econometrics ans statistics*, 1991.

[Lee, 1994]L. Lee. Semiparametric two-stage estimation of sample selection models subject to tobit-type selection rules. *Journal of Econometrics*, 61:305–344, 1994.

[Li *et al.*, 2003]K. C. Li, Y. Aragon, K. Shedden, and C. Thomas Agnan. Dimension reduction for multivariate response data. *Journal of the American Statistical Association*, 98:99–109, 2003.

[Maddala, 1983]G.S. Maddala. *Limited-dependent and qualitative variables in Econometrics*. Cambridge University Press, Cambridge, 1983.

[Maddala, 1993]G.S. Maddala. Estimation of limited-dependent variable models under rational expectations. *In: G. S. Maddala, C. R. Rao and H. D. Vinod, eds., Handbook of statistics*, 11:175–194, 1993.

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

[Saracco, 2005]J. Saracco. Asymptotics for pooled marginal slicing estimator based on sir$_\alpha$ approach. *Journal of Multivariate Analysis*, 96:117–135, 2005.