Two-step Estimation in a Multivariate Semiparametric Sample Selection Model - *Estimation en deux étapes dans un modèle de sélection semi-paramétrique multivarié*

Marie Chavent[1], Benoît Liquet[2] and Jérôme Saracco[1,3]

[1] Institut de Mathématiques de Bordeaux (IMB), Université Bordeaux 1
[2] ISPED, Université Bordeaux 2
[3] GREThA, Université Montesquieu Bordeaux IV

Joint Meeting of the Statistical Society of Canada and the Société Française de Statistique - Mai 2008

Plan

Plan

Tobit model or sample selection model (SSM)

Basically sample selection models (SSM) are described by two equations.
A selection equation gives the state "observed / non observed (missing)" of the dependent variable y as a function of explanatory variables x.
An outcome equation gives the value of the dependent variable, when observed, 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 communication, we focus on multivariate SSM, that is when $y \in \mathbb{R}^q,\ q > 1$.

A multivariate semiparametric sample selection model

For $j = 1, \ldots, q$,

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

- $y = (y^{(1)}, \ldots, y^{(q)})$ is a q-dimensional random vector.
- $\tilde{x}_1 \in \mathbb{R}^{p_1}$ and $\tilde{x}_2 \in \mathbb{R}^{p_2}$, are subvectors of a random vector $x \in \mathbb{R}^p$, assumed to be elliptically distributed with parameters $\mu = \mathbb{E}[x]$ and $\Sigma = \mathbb{V}(x)$ positive definite, that is $\tilde{x}_k = A_k' x$.
- The parameters $\tilde{\gamma}_1$ and $\tilde{\gamma}_2$ are the $p_1 \times 1$ and $p_2 \times 1$ real unknown slope parameters.

A multivariate semiparametric sample selection model

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

- $g_1^{(j)}$ and $g_2^{(j)}$ are unknown link functions called the observation link and the selection link functions $\longrightarrow$ link-free approach
- Let $\varepsilon = \left(\varepsilon_1^{(1)}, \varepsilon_2^{(1)}, \ldots, \varepsilon_1^{(q)}, \varepsilon_2^{(q)}\right)$ be a random error term independent of x with an unknown distribution $\longrightarrow$ distribution-free approach

We will focus on the parametric part
$\hookrightarrow$ estimation of the direction of the selection and observation slope vectors $\tilde{\gamma}_1$ and $\tilde{\gamma}_2$

A multivariate semiparametric sample selection model

For $j = 1, \ldots, q$,

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

This model is a particular case of a general multivariate two indices semiparametric regression model :

$$y = f(x'\gamma_1, x'\gamma_2, \varepsilon) \tag{2}$$

where $\gamma_k = A_k\tilde{\gamma}_k \in \mathbb{R}^p,\ k = 1, 2$.
(We only expand $\tilde{\gamma}_k$ to a $p \times 1$ vector with zeros corresponding to the non-selected components.)

Identifiability conditions :

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

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

Let us define the linear subspaces $E = \text{Span}(\gamma_1, \gamma_2)$ and $E_k = \text{Span}(A_k)$ of $\mathbb{R}^p$.

We can now bring these conditions into a geometrical perspective.

Theorem 1. *Under the assumptions of model (1) and the identifiability conditions, we have : for* $k = 1, 2$,

$$E \cap E_k = Span(\gamma_k).$$

Plan

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 $E = \mathrm{Span}(\gamma_1, \gamma_2)$.
- In the second step, since the linear subspaces E_1 and E_2 are known (since A_1 and A_2 are chosen by the user), two canonical analysis of (E, E_1) and (E, E_2) can provide bases of $E \cap E_1 = \mathrm{Span}(\gamma_1)$ and $E \cap E_2 = \mathrm{Span}(\gamma_2)$.

Population version

Step 1 : Pooled marginal sliced inverse regression $\hookrightarrow$ provide a Σ-orthogonal basis of E

Method : PMS_α (Pooled marginal sliced inverse regression based on SIR_α, see Saracco, 2005)
Major novelty : consider a transformation (slicing) $T_j(.)$ of $y^{(j)}$ with a specific slice for "missing" $y^{(j)}$ value.
Results : Under usual assumptions for MSIR approach and model (1), the eigenvectors v_1, v_2 associated with the largest two eigenvalues of $\Sigma^{-1} M_{\alpha,P}$ span the e.d.r. space :

$$\text{Span}(v_1, v_2) = E$$

where $M_{\alpha,P} = \sum_{j=1}^{q} w_j M_{\alpha_j}^{(j)}$, for positive weigths w_j and parameters $\alpha_j \in [0, 1]$.

Step 2a : Two canonical analysis $\hookrightarrow$ provide a Σ-orthogonal basis of $E_k \cap E$

We consider the subspaces E_k and E of $\mathbb{R}^p$ equipped with the inner product Σ.
This basis is formed by the eigenvector $b_k \in \mathbb{R}^p$ corresponding to the eigenvalue 1 of $P_{E_k} P_E P_{E_k}$, where P_{E_k} and P_E are respectively the Σ-orthogonal projectors onto E_k and E.
From Theorem 1, this 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 $\tilde{\gamma}_k \in \mathbb{R}^{p_k}$.

$$\tilde{b}_k = A_k' b_k \in \mathbb{R}^{p_k}$$

This vector $\tilde{b}_k$ is colinear to $\tilde{\gamma}_k$ and is Σ_k-normalized.

Estimation of the directions

Let $\{(y_i, x_i),\ i = 1, ..., n\}$ be a sample. 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 PMS$_\alpha$

$\hookrightarrow$ $\hat{E} = \text{Span}(\hat{v}_1, \hat{v}_2)$

where $\hat{v}_1$ and $\hat{v}_2$, are the eigenvectors associated with the two largest eigenvalues of $\hat{\Sigma}^{-1}\hat{M}_{\hat{\alpha},P}$.

Step 2a : Estimating the direction of $\gamma_k,\ k = 1, 2$ via canonical analysis of $(\hat{E}, E_k)$

$\hookrightarrow$ 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 $\tilde{\gamma}_k,\ k = 1, 2$.

$$\hat{\tilde{b}}_k = A_k'\hat{b}_k.$$

Plan

Convergence in probability of the estimated directions

Theorem 2. *Under classical assumptions for MSIR approach, we have : for* $k = 1, 2$,

$$\hat{\tilde{b}}_k = \tilde{b}_k + O_p(n^{-1/2}), \qquad \text{with the vector } \tilde{b}_k \text{ colinear to } \tilde{\gamma}_k.$$

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

Theorem 3. *Under classical assumptions for MSIR approach, we have : for* $k = 1, 2$,

$$\sqrt{n}(\hat{\tilde{b}}_k - \tilde{b}_k) \longrightarrow_d \mathcal{N}(0, C_k),$$

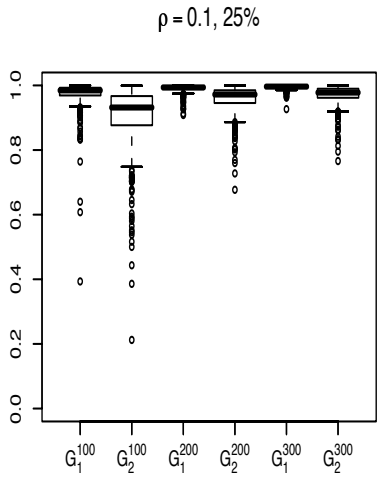
where the expression of C_k *can be found in Chavent et al. (2008).*

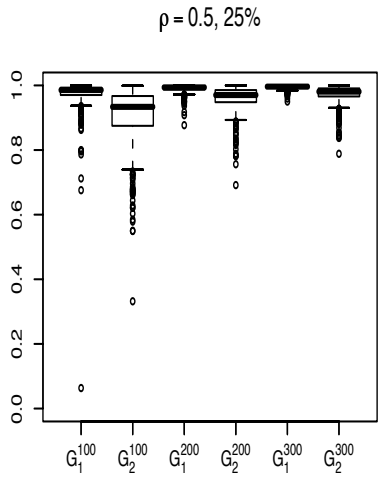
Plan

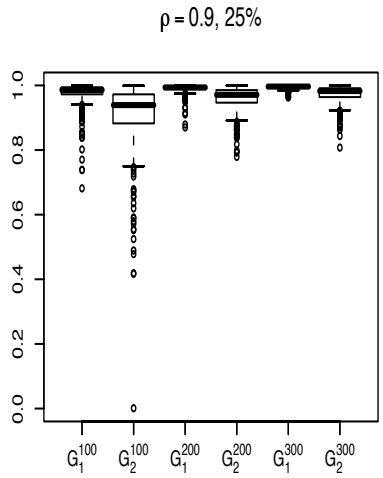
Simulated model : semiparametric multivariate ($q = 2$) model

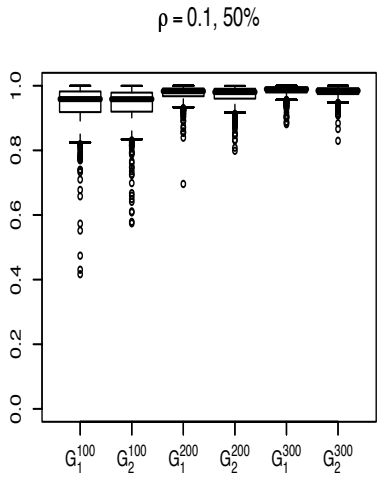
$$\begin{cases} g_1^{(1)}(\tilde{x}_1'\tilde{\gamma}_1, \varepsilon_1^{(1)}) = \exp(\tilde{x}_1'\tilde{\gamma}_1) + \varepsilon_1^{(1)} \text{ (observation)} \\ g_2^{(1)}(\tilde{x}_2'\tilde{\gamma}_2, \varepsilon_2^{(1)}) = \tilde{x}_2'\tilde{\gamma}_2 + \varepsilon_2^{(1)} \text{ (selection)} \\ \\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

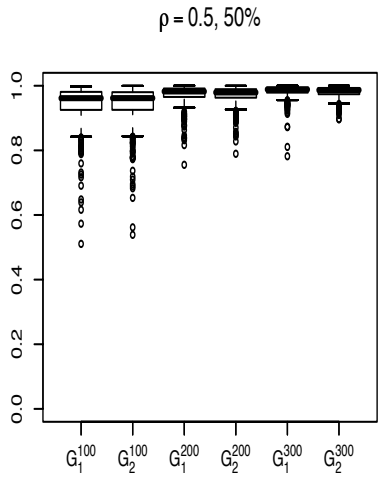
- Quality measure of the estimates : $\cos^2(\hat{\tilde{b}}_k^l, \tilde{\gamma}_k)$, $k = 1, 2$.
- Different percentages of non-observed values (25% and 50%)
- Various dimensions of the explanatory variable ($p = 5$, 10)
- Various correlation of the error term between the obsevation equation and the selection equation ($\rho = 0.1,\ 0.5,\ 0.9$)
- Different sample sizes ($n = 100,\ 200$ and 300)

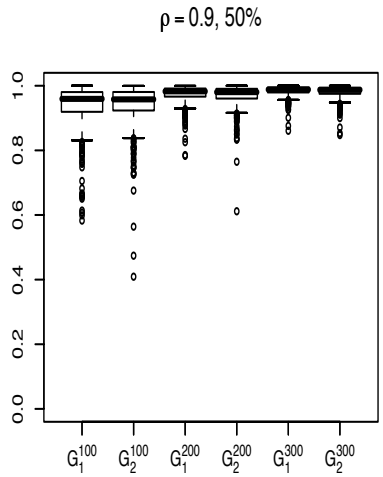

ρ = 0.1, 25%
0.0
0.2
0.4
0.6
0.8
1.0
G_1^{100} G_2^{100} G_1^{200} G_2^{200} G_1^{300} G_2^{300}


ρ = 0.5, 25%
0.0
0.2
0.4
0.6
0.8
1.0
G_1^{100} G_2^{100} G_1^{200} G_2^{200} G_1^{300} G_2^{300}


ρ = 0.9, 25%
0.0
0.2
0.4
0.6
0.8
1.0
G_1^{100} G_2^{100} G_1^{200} G_2^{200} G_1^{300} G_2^{300}


ρ = 0.1, 50%
0.0
0.2
0.4
0.6
0.8
1.0
G_1^{100} G_2^{100} G_1^{200} G_2^{200} G_1^{300} G_2^{300}

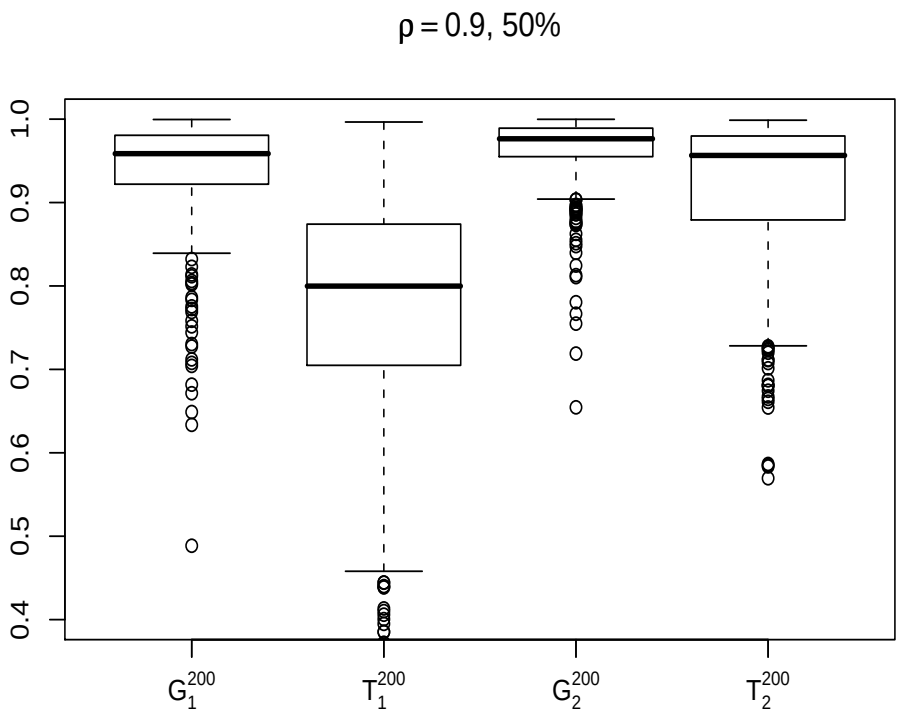

ρ = 0.5, 50%
0.0
0.2
0.4
0.6
0.8
1.0
G_1^{100} G_2^{100} G_1^{200} G_2^{200} G_1^{300} G_2^{300}


ρ = 0.9, 50%
0.0
0.2
0.4
0.6
0.8
1.0
G_1^{100} G_2^{100} G_1^{200} G_2^{200} G_1^{300} G_2^{300}

Comparison with a parametric approach : Tobit II model

Simulated model ($q = 1$) :

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



Concluding remarks

- Main advantages :
  - Link-free and distribution-free method.
  - Geometric approach which deals symmetrically with both selection and observation slope vectors.
  - Estimation method numerically very fast.
- The R source code is available from the authors.
- The simulation study has highlighted a good behaviour of the method even for non-elliptical distribution of the covariate.
- A real economic application is currently under investigation.

References

Chavent, M., Liquet, B. and Saracco, J, 2008, A semiparametric approach for multivariate sample selection model. *In revision for Statistica Sinica*.

Saracco, J., 2005, Asymptotics for pooled marginal slicing estimator based on SIR_α approach. *Journal of Multivariate Analysis*, **96**, 117-135.

Remark : For model (2), two crucial conditions for the theoretical success of SIR_α and PMS_α methods are the following :
a linearity condition

$$\mathbb{E}(v'x|\gamma_1'x, \gamma_2'x) \text{ is linear for any } v, \tag{3}$$

and a constant variance condition

$$\mathbb{V}(x|\gamma_1'x, \gamma_2'x) \text{ is non-random.} \tag{4}$$

Note that (3) is satisfied when x has an elliptically symmetric distribution and (4) is satisfied when x follows a multivariate normal distribution (which is an elliptically one).

Under usual assumptions for MSIR approach (satisfied when x follows a multivariate normal distribution) and model (1), 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 E :

$$\text{Span}(v_1, v_2) = E$$

Let us first give a brief overview of 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 :

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

The **observed variables** are : $y_1 \in \mathbb{R}$, $y_2 \in \{0, 1\}$ and $x \in \mathbb{R}^p$.

Equation $(E3)$ is the selection equation, and equation $(E4)$ is the outcome equation.

Simulated model :

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

where $x \sim \mathcal{N}_p(0, I_p)$, $\tilde{x}_1$ (resp. $\tilde{x}_2$) is the $(p-1)$-dimensional vector corresponding to the first (resp. last) $(p-1)$ coordinates of x.

The error term $\varepsilon = (\epsilon_1^{(1)}, \epsilon_2^{(1)}, \epsilon_1^{(2)}, \epsilon_2^{(2)})'$ is normally distributed : $\epsilon \sim \mathcal{N}_4(\mu_\epsilon, \Sigma_\epsilon)$.

Two design of the covariance of ϵ will be considered :

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

with different values of ρ (0.1, 0.5 and 0.9).

In the matrix Σ_ϵ^I the error term associated with the two component $y^{(1)}$ and $y^{(2)}$ are assumed to be independant, which is not the case with the covariance matrix Σ_ϵ^{II}.

To control the number of non observed values for the $y^{(j)}$'s component, we will use two different values of μ_ϵ in order to obtain around 25% (resp. 50%) of non observed values for $y^{(1)}$ and $y^{(2)}$.

For the slope parameters, we take $\tilde{\gamma}_1 = (1, 1, -1, -1, 0, \ldots, 0)'$ and $\tilde{\gamma}_2 = (0, \ldots, 0, 1, -1, 1, -1)'$.

Simulated example

We consider a simulated sample of $n = 100$ data points from the previous model for $p = 5$, $\Sigma_\epsilon = \Sigma_\epsilon^{II}$, $\rho = 0.5$ and $\mathcal{L} = 25\%$.

Let us introduce the two variables $y_*^{(1)} = g_2^{(1)}(\tilde{x}_2'\tilde{\gamma}_2, \varepsilon_2^{(1)})$ and $y_*^{(2)} = g_2^{(2)}(\tilde{x}_2'\tilde{\gamma}_2, \varepsilon_2^{(2)})$, which are called in the literature latent variables (since in practice the values of these variables are never available in the sample).

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.

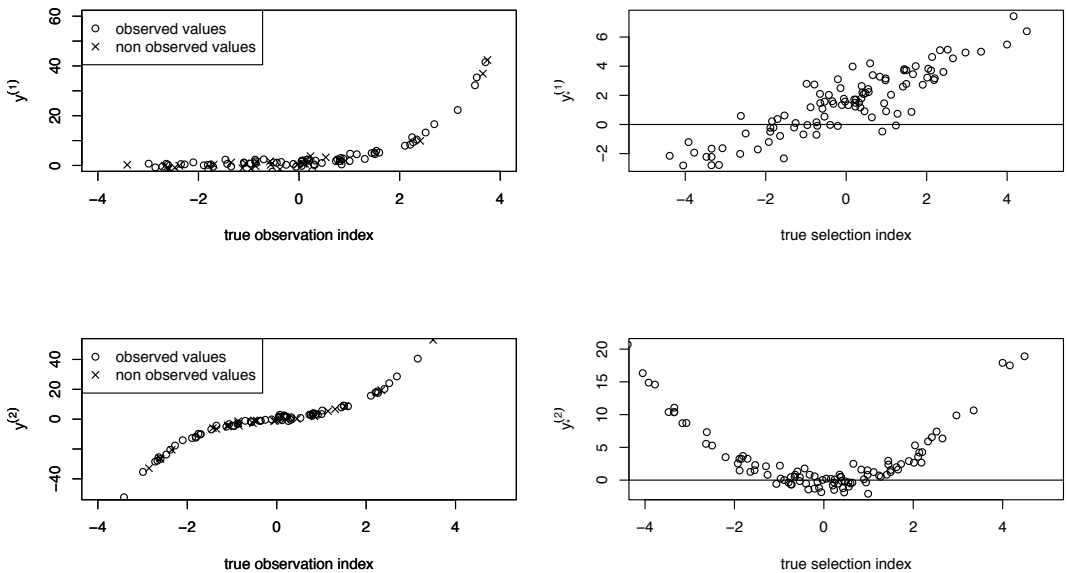


FIG.: *Plots of $y^{(j)}$ versus the true "observation" index $\tilde{x}_1'\tilde{\gamma}_1$ (on the left) and plots of the latent variables $y_*^{(j)}$ versus the true "selection" index $\tilde{x}_2'\tilde{\gamma}_2$ (on the right).*

The direction of $\tilde{\gamma}_1$ and $\tilde{\gamma}_2$ are then estimated and we get

$$\hat{\tilde{b}}_1 = (-0.483, -0.565, 0.447, 0.497)' \quad \text{and} \quad \hat{\tilde{b}}_2 = (-0.613, 0.539, -0.350, 0.459)'.$$

The corresponding squared cosines are respectively equal to 0.993 and 0.962. Moreover, we compute the quality of the estimation $\hat{E}$ of the e.d.r. space E using $\text{Trace}(P_E P_{\hat{E}})/2$ which is equal to 0.886 for this simulated sample. Even if this subspace is relatively poorly estimated compared with the quality of each estimated direction, the second step (which takes into account additional information) ensures to recover the good directions of the observation and selection slope vectors.

In Figure 2, we represent on the left hand side the plots of the response variable $y^{(j)}$ versus the estimated "observation" index $\tilde{x}_1' \hat{\tilde{b}}_1$. Note that the scatterplots of the (left hand side) Figures 1 and 2 have not the same orientation. We add on these plots the Nadaraya-Watson estimation 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 the Nadaraya-Watson estimation of the probability to observe $y^{(j)}$.

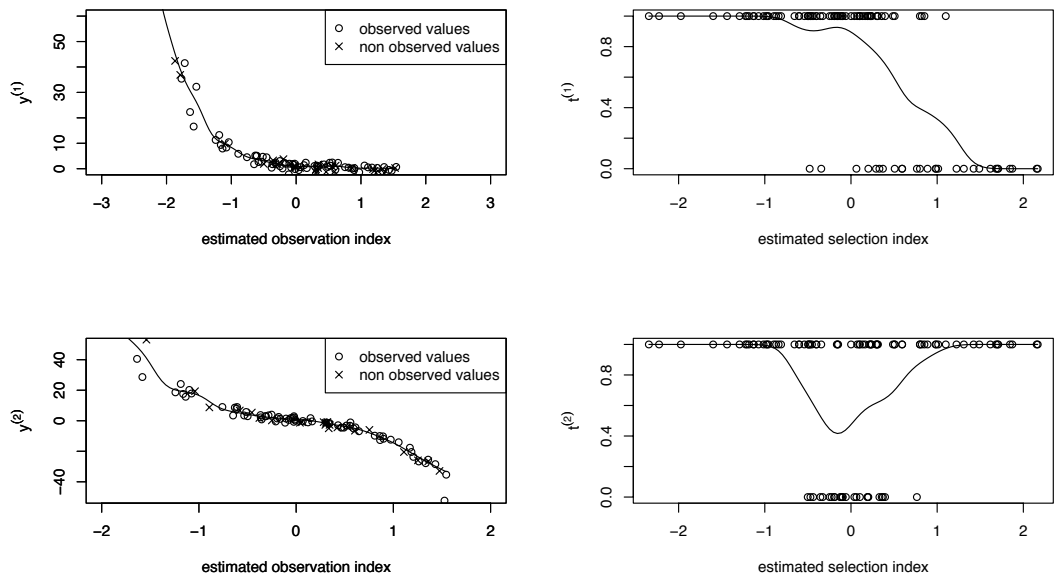


FIG.: *Kernel estimation of the observation link functions (left hand side) and Nadaraya-Watson estimation of the probability of* $t^{(j)} = 1$ *(that is* $y^{(j)}$ *observed)*

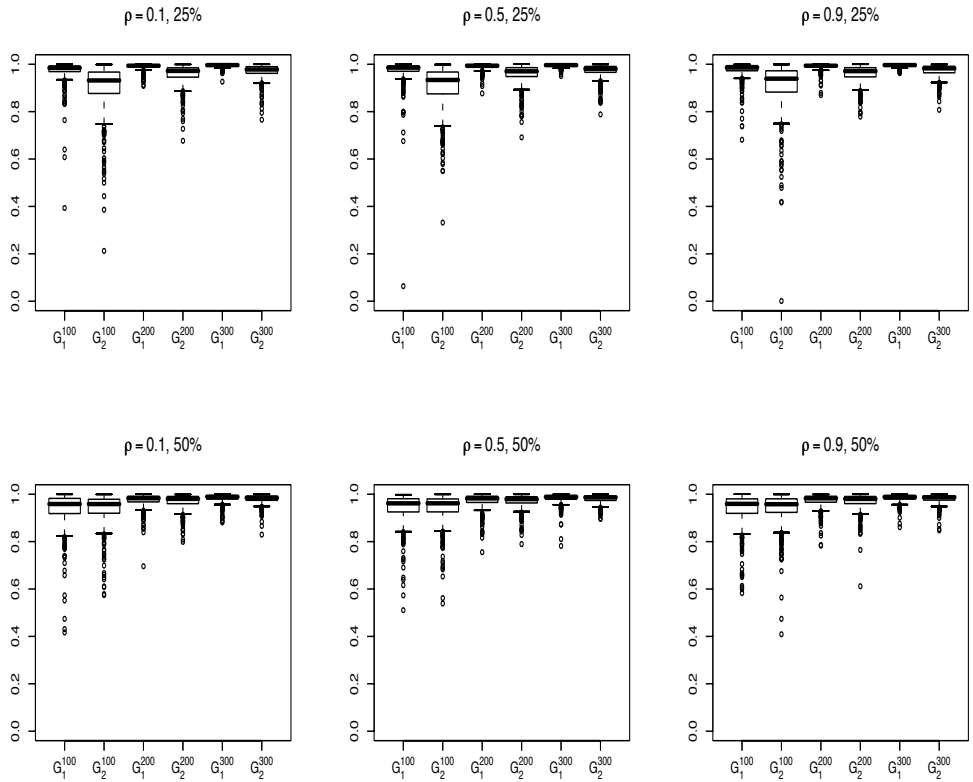
Simulation study

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

Results of the simulation study

For each combination of the simulation parameters (p, n, ρ, $\mathcal{L}$, ...), $N = 500$ samples have been generated. For each sample $l = 1, \ldots, N$, the directions of the slope vectors $\tilde{\gamma}_1$ and $\tilde{\gamma}_2$ have been estimated and we get $\hat{\tilde{b}}_1^l$ and $\hat{\tilde{b}}_2^l$. Then, we evaluate 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 $l = 1, \ldots, N$. The closer the squared cosine is to one, the better the estimation.

We exhibit the results via the boxplots of these squared cosines for only one combination.



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

For others simulation parameters, we have observed that the results were also very good. More precisely, one can note that :

- the form of the covariance matrix of the error term ε and the value of the parameter ρ do not seem to influence the quality of the estimates.
- the level $\mathcal{L}$ of the non observed values for the $y^{(j)}$'s only have a gentle influence on the quality of the estimation of the selection slope vectors $\tilde{\gamma}_2$, especially in terms of spread ot the squared cosine values.
  When this level is low ($\mathcal{L} = 25\%$), there is less information on the selection part of the model and then the quality of the $\tilde{\gamma}_2$ estimates are slightly lower than when this level is larger ($\mathcal{L} = 50\%$).

  On the other hand, not surprisingly, one can observe 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 has a quite predictable influence of the quality of the estimates : the largest is the sample size, the greatest are the squared cosines.
- the dimension p of the explanatory variable x does seem to have any effect on the quality of the estimates.
- In order to investigate the robustness of the method when x does not follow a multivariate normal distribution, we consider here a discrete distribution for x. One can see that the estimations of the directions of the slopes for the selection equations and the outcome equations are quite good, even for a discrete x.