INSTITUT DE
STATISTIQUE

UNIVERSITÉ CATHOLIQUE DE LOUVAIN

DISCUSSION
PAPER

0807

SINGLE-INDEX MODELLING OF CONDITIONAL PROBABILITIES IN TWO-WAY CONTINGENCY TABLES

GEENENS, G. and L. SIMAR

Single-index modelling of conditional probabilities in two-way contingency tables

Gery Geenens and Léopold Simar
Institut de Statistique
Université Catholique de Louvain, Belgium

February 8, 2008

Abstract

When analyzing a contingency table, it is often worth relating the probabilities that a given individual falls into the different cells to a set of predictors. These conditional probabilities are usually estimated using appropriate regression techniques. In particular, in this paper, a semiparametric model is developped. Essentially, it is only assumed that the effect of the vector of covariates on the probabilities can entirely be captured by a single index, which is a linear combination of the initial covariates. The estimation is then twofold : the coefficients of the linear combination and the functions linking this index to the related conditional probabilities have to be estimated. Inspired by the estimation procedures already proposed in the literature for Single-Index regression models, four estimators of the index coefficients are proposed and compared, from a theoretical point-of-view, but also practically, with the aid of simulations. Estimation of the link functions is also adressed.

Key words : contingency table; conditional probabilities; semiparametric regression; single-index model; semiparametric maximum likelihood; semiparametric least squares; average derivatives; sliced inverse regression.

Acknowledgement : Research support from the "Interuniversity Attraction Pole", Phase VI (No. P06/03) from the Belgian Science Policy is acknowledged.

1 Introduction

Consider the contingency table built by cross-classifying a sample of n individuals with respect to the levels of two categorical variables R and S, having r and s levels respectively. Quantity of interest facing such a table is typically the joint probability distribution $\pi = \{\pi_{ij} : 1 \leq i \leq r, 1 \leq j \leq s\}$ of R and S, with

$$\pi_{ij} = \mathbb{P}\left(R = i, S = j\right),$$

i.e. the probability that a given individual falls into cell (i, j) of the table. The analysis of such a table is an over-studied problem in statistical literature. See e.g. Everitt (1992). However, it is often the case that for each individual of the sample are known not only R and S, but also a set of p explanatory variables, say X, characterizing him. In most of the situations, it is useful to relate the cell probabilities to these characteristics. At first, to check an eventual effect of X on R and S, and next to perform the classical analyses on contingency tables, taking this eventual heterogeneity of the population into account. In these purposes are needed reliable estimates of the conditional distribution of R and S given X, denoted $\pi(x) = \{\pi_{ij}(x) : 1 \leq i \leq r, 1 \leq j \leq s\}$, with

$$\pi_{ij}(x) = \mathbb{P}\left(R = i, S = j \middle| X = x\right). \tag{1.1}$$

This paper adresses the problem of estimation of such conditional probabilities.

For a long time, the relation between a categorical response and some explanatory variables was almost always analysed via parametric methods, mainly logistic regression and its generalizations. McCullagh and Nelder (1989, section 6.5.4) proposed to generalize the basic idea of binary logistic regression to multivariate categorical response models. In a general way, their multivariate logistic regression model, also discussed by Glonek and McCullagh (1995) and Glonek (1996), is written

$$C^t \log(L\pi(X)) = \Theta X, \tag{1.2}$$

where L and C are appropriately chosen matrix of 0, 1 and -1, and Θ is a $(rs-1)\times(p+1)$ matrix of unknown parameters. As illustration, in the simplest case $r = s = 2$, it becomes[†]

$$\begin{aligned} \text{logit}\left(\pi_{1\cdot}(X)\right) &= \theta_R^t X & (1.3)\\ \text{logit}\left(\pi_{\cdot 1}(X)\right) &= \theta_S^t X & (1.4)\\ \log\left(\frac{\pi_{11}(X)\pi_{22}(X)}{\pi_{12}(X)\pi_{21}(X)}\right) &= \theta_{RS}^t X, \end{aligned}$$

[†]We adopt the usual following subscript convention: "$\cdot$" denotes the sum over the index it replaces, e.g. $\pi_{i\cdot} = \sum_{j=1}^{s} \pi_{ij}$.

with θ_R, θ_S and θ_{RS} vectors of unknown parameters to be estimated. This model rests on three structural assumptions. As it implies the univariate logistic model marginally for R and S, it first supposes that the vector of covariates X influences the distribution of R (resp. the one of S) only through a linear combination $\theta_R^t X$ (resp. $\theta_S^t X$), and second, that the functions linking these linear combinations to the resulting marginal conditional probabilities are the same, namely the logit function. Finally, it moreover assumes that the log of the odds, given X, is also a linear function of X.

The lack of flexibility implied by this structure seems to be a serious limitation of this model. For example, once a marginal conditional probability to fall in some level of R or S is not a monotonic function of one of the covariate, expressions like (1.3) or (1.4) are not appropriate. In classical regression, such kind of fit problem is often overcome by using nonparametric methods. These ones require no structural assumption on the underlying functions, except very mild properties such as a certain amount of smoothness. For categorical response model, the use of nonparametric regression techniques was first studied by Copas (1983). Later, Azzalini et al (1989), Rodriguez-Campos and Cao-Abad (1993) and Chu and Cheng (1995) a.o. used a Nadaraya-Watson (NW) estimator in this context. Recently, Geenens and Simar (2008) developped a nonparametric test of independence between two categorical variables, conditionally to a set of explanatory variables, using this estimator. This one can be defined in the following way. Consider the random vector

$$Z = (Z^{(11)}, Z^{(12)}, \ldots, Z^{(rs)})^t, \tag{1.5}$$

with $Z^{(ij)}$ taking the value 1 if the individual belongs to cell (i, j) and 0 otherwise. Note that, for ease of derivation, the components of Z are indexed by the pairs ij, so that the subset (ij) denotes the $((i-1)s+j)$th element of Z. From a sample $\{(X_k, Z_k), k = 1, \ldots, n\}$, and given a kernel function K and a bandwidth h, the NW estimator of $\pi_{ij}(x)$ is given by[†]

$$\hat{p}_{ij}(x) = \frac{\sum_{k=1}^{n} K_h(x - X_k) Z_k^{(ij)}}{\sum_{k=1}^{n} K_h(x - X_k)}, \tag{1.6}$$

that is a weighted average of the 0-1 responses $Z_k^{(ij)}$, with weights varying according to the distance between x and X_k. For more details, see the classical references dealing with nonparametric regression, such as Härdle (1990) or Wand and Jones (1995).

Although the great amount of flexibility offered by this kind of estimator, it also suffers from an important disadvantage: it can be shown that it gets very demanding with respect to the

[†] K_h is the usual normalized version of K: $K_h(\cdot) = \frac{1}{h} K(\cdot/h)$.

number of observations when increasing the number of regressors. Specifically, the fastest achievable rate of convergence of nonparametric regression function estimators towards the true curve decreases as the number of continuously distributed components of X increases (see Stone (1980)). Hence, the larger the number of regressors, the larger the dimension of data samples needed in order to achieve reasonable estimates, so that in practice, for data sample of usual size, nonparametric estimators are considered as reliable if p, the number of regressors, is 1 or 2 only. This phenomenon is known as the "curse of dimensionality", and affects any nonparametric method.

The general idea developped in this paper is to propose a semiparametric method for estimating the conditional probabilities (1.1). Semiparametric models can be seen as a mix of parametric and nonparametric approaches, with the aim to compensate for their respective drawbacks. Formally, they are characterized by a twofold parametrization, say θ and γ, where θ lies in a finite-dimensional space Θ (parametric part of the model) and γ lies in an infinite-dimensional space Γ (nonparametric part). This permits to relax some of the restrictive shape assumptions of parametric models, keeping the model as flexible as possible via γ, but in the same time to maintain many of the desirable features of them, essentially their good properties and their ease of computation and interpretation. In particular, one objective is to avoid the above-metionned curse of dimensionality. In this work, we formulate a Single-Index assumption, made explicit in the next section, on the conditional probabilities (1.1). Note that the use of semiparametric models in case of discrete response was already discussed by Manski (1985), Klein and Spady (1993), Thompson (1993), Lee (1995) or Pagan and Ullah (1999, section 7), among others.

The paper is organized as follows. Section 2 describes the assumed semiparametric model for the conditional probabilities. As it will appear, the estimation is twofold: some link functions and a vector of parameters need to be estimated. Section 3 is concerned with the estimation of the link functions, while section 4 deals with the estimation of the parametric part of the model. Section 5 proposes a simulation study, which illustrates the practical performances in finite sample of the different estimators described in section 4. Section 6 concludes.

2 Single-index modelling of the conditional probabilities

As mentionned in the introduction, this paper develops a semiparametric model for the conditional probabilities based on a Single-Index assumption. Single-Index Models (SIM)

were first introduced as such by Ichimura (1987, 1993). In a classical regression setting, where Y is a continuous response and X a vector of covariates, a SIM is given by

$$Y = g_0(\theta_0^t X) + \varepsilon, \tag{2.1}$$

with ε a random disturbance such that $\mathbb{E}(\varepsilon|X) = 0$, θ_0 an unknown p-vector of parameters and g_0 an unknown link function. An equivalent formulation is

$$\exists\, \theta_0 \in \mathbb{R}^p : m(x) \doteq \mathbb{E}(Y|X = x) = \mathbb{E}(Y|\theta_0^t X = \theta_0^t x) = g_0(\theta_0^t x), \tag{2.2}$$

i.e. the vector of covariates X influences the conditional mean of Y only through a linear combination $\theta_0^t X$ of its components, called the index. In this model, the index coefficient vector θ_0, forming the parametric part of the model, and the link function g_0, the non-parametric part, have to be estimated, contrary to what happens in a Generalized Linear Model, where the function g_0 is a priori fixed and supposed invertible. At this point, an important remark is that any pair (index coefficients vector, link function) from the set $\{(c\theta_0, g_c(.) \doteq g_0(./c)), c \in \mathbb{R}_0\}$ should exactly lead to the same regression function $m(x)$, so that they could not be distinguished. Hence, for identifiability purpose, it is necessary to fix the scale of θ_0, for example by fixing $\theta_0^{(1)} = 1$, where $\theta_0^{(1)}$ is the first component of the vector θ_0. Many estimators of this θ_0 have been proposed in the literature. In the remainder, we will mainly be interested in the Semiparametric (or Pseudo) Maximum Likelihood of Klein and Spady (1993), Ai (1997) or Delecroix et al (2003), the Semiparametric Least Squares estimator of Ichimura (1993), the Average Derivative estimator of Powell et al (1989) and the Sliced Inverse Regression estimator of Duan and Li (1991). Those are the most popular methods in this context, and are comprehensively reviewed in Horowitz (1998, chapter 2), Härdle et al (2004, chapter 6) or Geenens and Delecroix (2006).

In this paper, we propose to adapt those ideas to the case of the estimation of the conditional probabilities in a two-way contingency table. The main difference is that only one unknown link function is concerned in model (2.1), while it will appear that, in our framework, rs link functions are needed. With the vector Z defined in (1.5), we will first suppose :

Assumption 1. *The sample $\{(X_k, Z_k), k = 1, \ldots, n\}$ is formed by i.i.d. replications of (X, Z), a random vector of compact support $D = S_X \times \{z \in \{0,1\}^{rs} : \sum_q z^{(q)} = 1\}$, with S_X not contained in any proper linear subspace of $\mathbb{R}^p$, and such that $Z|X$ follows a multinomial distribution, with parameters $(1; \pi(X))$.*

Since we have obviously $\mathbb{E}(Z^{(ij)}|X = x) = \pi_{ij}(x)$, analogy with (2.2) leads to write the single-index assumption as

Assumption 2. *There exists* $\theta_0 \in \Theta \subset \{\theta \in \mathbb{R}^p : \theta^{(1)} = 1\}$ *and* rs *functions* $g_{ij} : \mathbb{R} \to \mathbb{R}$, $i = 1, \ldots, r$, $j = 1, \ldots, s$ *such that*

$$\pi_{ij}(x) = g_{ij}(\theta_0^t x) \qquad \forall\, x \in S_X,\, \forall\, (i,j).$$

In other words, this assumption amounts to say that

(i) the influence of the vector of covariates X on the joint distribution of R and S can entirely be captured by some linear combination $\theta_0^t X$, and

(ii) no structural assumptions on the functions g_{ij}, linking this index to the related conditional probabilities, are made.

Note that (i) is quite different to what model (1.2) assumed: here, one single index $\theta_0^t X$ is concerned for the whole joint distribution of R and S, while in the former $(r-1)+(s-1)$ linear combinations were needed to model the marginal distributions, plus $(r-1)(s-1)$ for the interactions. On the other hand, (ii) requires the nonparametric estimation of the functions $\{g_{ij}\}$, and therefore the use of estimators such as (1.6). This grants the model a great flexibility. In particular, monotonicity of functions $\{\pi_{ij}\}$ is not required. In the remainder, it will also be assumed the following regularity conditions.

Assumption 3. *The random variable* $U_0 = \theta_0^t X$ *admits a bounded density* f_0*, which has two bounded continuous derivatives. Moreover,* $f_0(u) > 0$ *for any* u *in the support of* U_0*, denoted* S_0*.*

Assumption 4. *The functions* g_{ij} *have two bounded continuous derivatives, and* $0 < g_{ij}(u) < 1$*,* $\forall u \in S_0$*,* $\forall (i,j)$*. Moreover, there is at least one* g_{ij} *which is not a constant function.*

These conditions are sufficient in order to ensure the identification of the model. This can easily be shown in the same way as the proof of theorem 4.1 of Ichimura (1993), starting from the fact that θ_0 minimizes $\mathbb{E}\left((Z - \mathbb{E}(Z|\theta^t X))^t (Z - \mathbb{E}(Z|\theta^t X))\right)$ on Θ. Note that a necessary condition for Assumption 3 is that there exists at least one continuously distributed regressor, say $X^{(1)}$. This therefore allows some explanatory variables to be discrete. However, in this latter case, identification of the model requires two extra conditions: (a) varying the values of the discrete regressors does not divide S_0 into disjoints subsets, and (b) there is at least one g_{ij} which is not a periodic function. See Ichimura (1993) or Horowitz (1998, section 2.4) for details. In the sequel, we will also refer to f_θ, g_{ij}^θ and S_θ as the density of $\theta^t X$, the conditional expectation of $Z^{(ij)}$ given $\theta^t X$ and the support of f_θ, respectively. Note that $f_0 \equiv f_{\theta_0}$, $g_{ij} \equiv g_{ij}^{\theta_0}$ and $S_0 \equiv S_{\theta_0}$. Define also $S_0^{(h)}$, the "interior" of the support S_0, as

$S_0^{(h)} \doteq \{u \in S_0 : m_U + h \leq x \leq M_U - h\}$, where m_0 and M_0 are the lower and the upper bound of S_0[†]. Such a set needs to be defined as it is well known that the behavior of the Nadaraya-Watson estimator differs when computed at points close to the boundary of the support.

3 Estimation of the link functions

Suppose at first that the vector θ_0 appearing in Assumption 2 is known. Then, any function g_{ij} could be estimated via the regression of $Z^{(ij)}$ on the index $\theta_0^t X$. As this index is univariate, the curse of dimensionality mentionned in introduction would be avoided. From the sample $\{(X_k, Z_k), k = 1, \ldots, n\}$, the Nadaraya-Watson estimator of g_{ij} would be given by

$$\hat{g}_{ij}^{\theta_0}(u) = \frac{\sum_{k=1}^{n} K_h(u - \theta_0^t X_k) Z_k^{(ij)}}{\sum_{k=1}^{n} K_h(u - \theta_0^t X_k)}. \tag{3.1}$$

The asymptotic theory of such an estimator is well known. In particular, besides Assumptions 1-4, it is often assumed that

Assumption (link1). *The kernel K is a symmetric Lipschitz continuous probability density on $[-1, 1]$;*

and

Assumption (link2). *The bandwidth sequence is such that $nh^5 = O(1)$.*

Then, we have, for any $u \in S_0^{(h)}$, for all (i, j),

$$(nh)^{1/2} \left(\hat{g}_{ij}^{\theta_0}(u) - g_{ij}(u) - b_{ij}(u) \right) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \sigma_{ij}^2(u)\right), \tag{3.2}$$

where

$$b_{ij}(u) = \frac{1}{2}\kappa_2 h^2 \left(g_{ij}''(u) + 2 g_{ij}'(u) \frac{f_0'(u)}{f_0(u)} \right) \text{ and } \sigma_{ij}^2(u) = \nu_0 \frac{g_{ij}(u)(1 - g_{ij}(u))}{f_0(u)}, \tag{3.3}$$

with $\kappa_q = \int u^q K(u) du$ and $\nu_q = \int u^q K^2(u) du$.

Remark 3.1. *As the asymptotic bias and variance depend on the function g_{ij} itself, the asymptotic optimal value (in the sense of minimum MISE) for the bandwidth in estimator*

[†]As S_X is assumed compact (Assumption 1), its projection on θ_0 is also compact, that is a closed interval of $\mathbb{R}$.

(3.1) indicates that different values h_{ij} should be used for estimation of each function g_{ij}. Nevertheless, we argue it is preferable in practice to use the same bandwidth h for each cell (i,j). The reason is simple: it permits to keep, for $\{\hat{g}_{ij}^{\theta_0}(x)\}$, essential properties of the underlying $\{\pi_{ij}(x)\}$, mainly the fact that they sum to one for any x. It should not be the case if different bandwidths h_{ij} were used. See Geenens and Simar (2008, section 2.2) for a more detailed related discussion.

Due to the assumed multinomial sampling, standard developments show that the asymptotic covariance between $\hat{g}_{i_1j_1}^{\theta_0}(u)$ and $\hat{g}_{i_2j_2}^{\theta_0}(u)$ is $-\nu_0 \frac{g_{i_1j_1}(u) g_{i_2j_2}(u)}{nhf(u)}$, for $(i_1,j_1) \neq (i_2,j_2)$. Therefore, defining vectors

$$\begin{aligned} g(u) &= \left(g_{11}(u), g_{12}(u), \ldots, g_{r(s-1)}(u), g_{rs}(u)\right)^t \\ \hat{g}^{\theta_0}(u) &= \left(\hat{g}_{11}^{\theta_0}(u), \hat{g}_{12}^{\theta_0}(u), \ldots, \hat{g}_{r(s-1)}^{\theta_0}(u), \hat{g}_{rs}^{\theta_0}(u)\right)^t \\ b(u) &= \left(b_{11}(u), b_{12}(u), \ldots, b_{r(s-1)}(u), b_{rs}(u)\right)^t, \end{aligned}$$

the vector analogue of (3.2) is

$$(nh)^{1/2}\left(\hat{g}^{\theta_0}(u) - g(u) - b(u)\right) \xrightarrow{\mathcal{L}} \mathcal{N}_{rs}\left(0, \frac{\nu_0}{f(u)}\left(\mathrm{diag}(g(u)) - g(u)g(u)^t\right)\right),$$

with $\mathrm{diag}(g(u))$ being the diagonal matrix built on the elements of $g(u)$.

Obviously, as θ_0 is unknown in practice, estimator (3.1) is not feasible as such. However, suppose that a consistent estimator $\hat{\theta}$ is known. A natural estimator for g_{ij} then becomes

$$\hat{g}_{ij}^{\hat{\theta}}(u) = \frac{\sum_{k=1}^{n} K_h(u - \hat{\theta}^t X_k) Z_k^{(ij)}}{\sum_{k=1}^{n} K_h(u - \hat{\theta}^t X_k)}. \tag{3.4}$$

Assume moreover that $\hat{\theta}$ is root-n consistent, i.e.

$$\|\hat{\theta} - \theta_0\| = O_P(n^{-1/2}), \tag{3.5}$$

which is the typical rate of convergence for parametric estimators. It is well known that a nonparametric estimator cannot achieve this rate, so that the convergence of $\hat{\theta}$ towards the true θ_0 is in that case faster than the convergence of $\hat{g}_{ij}^{\theta_0}(u)$ towards $g_{ij}(u)$. Therefore, it is intuitively clear that the difference between $\hat{g}_{ij}^{\hat{\theta}}(u)$ and $\hat{g}_{ij}^{\theta_0}(u)$ would be asymptotically negligible. Specifically, we would have

$$(nh)^{1/2}\left(\hat{g}_{ij}^{\hat{\theta}}(u) - g_{ij}(u)\right) = (nh)^{1/2}\left(\hat{g}_{ij}^{\theta_0}(u) - g_{ij}(u)\right) + o_P(1)$$

for any u in S_0, so that the estimation of θ_0 by an estimator $\hat{\theta}$ would have no effect on the asymptotic distribution of the estimator of g_{ij}, provided that (3.5) holds. See a.o. Härdle and Stoker (1989, theorem 3.3) for a formal proof of this reasoning. Hence, we can write, for any $u \in S_0$, for all (i,j), for any root-n consistent estimator $\hat{\theta}$ of θ_0,

$$(nh)^{1/2}\left(\hat{g}_{ij}^{\hat{\theta}}(u) - g_{ij}(u) - b_{ij}(u)\right) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \sigma_{ij}^2(u)\right),$$

with $b_{ij}(u)$ and $\sigma_{ij}^2(u)$ defined in (3.3), and

$$(nh)^{1/2}\left(\hat{g}^{\hat{\theta}}(u) - g(u) - b(u)\right) \xrightarrow{\mathcal{L}} \mathcal{N}_{rs}\left(0, \frac{\nu_0}{f_0(u)}\left(\mathrm{diag}(g(u)) - g(u)g(u)^t\right)\right).$$

Besides, as kernel estimators such as $\hat{g}^{\hat{\theta}}$ inherits the smoothness properties of the kernel K, we have also, by assumption (link1), that $|\hat{g}^{\hat{\theta}}(\hat{\theta}^t x) - \hat{g}^{\hat{\theta}}(\theta_0^t x)| = O(\|\hat{\theta} - \theta_0\|)$, and therefore, under assumption 2, for any $x \in S_X$ such that $\theta_0^t x \in S_0^{(h)}$,

$$(nh)^{1/2}\left(\hat{g}^{\hat{\theta}}(\hat{\theta}^t x) - \pi(x) - b(\theta_0^t x)\right) \xrightarrow{\mathcal{L}} \mathcal{N}_{rs}\left(0, \frac{\nu_0}{f_0(\theta_0^t x)}\left(\mathrm{diag}(\pi(x)) - \pi(x)\pi(x)^t\right)\right).$$

Next section shows that estimators satisfying (3.5) actually exist.

4 Estimation of the index

In this section, the main estimation procedures of the index coefficients vector θ_0 in classical SIM are adapted to our setting. First of all, notice that those methods can be classified into two groups, according to whether they require solving a nonlinear optimization problem (M-estimators) or not (direct estimators). Examples of M-estimators are typically the Semiparametric Least Squares and the Semiparametric Maximum Likelihood estimators, direct ones are among others the Average Derivatives and the Sliced Inverse Regression estimators.

4.1 Semiparametric Maximum Likelihood estimator (SML)

Maximum Likelihood methods are well studied in semiparametric models. In the usual SIM context (2.1)-(2.2), Ai (1997) and Delecroix et al (2003) form a quasi-likelihood function by replacing the unknown density of the response conditional on the index by a nonparametric estimator. Beforehand, Klein and Spady (1993) developped this idea in a binary-response model, and Lee (1995) generalized it to the polychotomous case. An evident but interesting observation is that the model defined by Assumptions 1-2 is nothing else but a polychotomous choice model, so that those results directly adapt. This adaptation is presented hereafter.

If the link functions were known, the parametric multinomial likelihood would be

$$\Lambda(\theta) = \prod_{k=1}^{n} \prod_{ij=11}^{rs} g_{ij}(\theta^t X_k)^{Z_k^{(ij)}},$$

so that the log-likelihood would be written

$$L(\theta) = \sum_{k=1}^{n} \sum_{ij=11}^{rs} Z_k^{(ij)} \log g_{ij}(\theta^t X_k) \tag{4.1}$$

and the estimator would be the value which maximizes $L(\theta)$. This parametric maximum likelihood estimator is known to have excellent asymptotic properties, in particular its asymptotic efficiency.

When the functions $\{g_{ij}\}$ are unknown, the semiparametric maximum likelihood estimator is then defined as

$$\hat{\theta} = \arg\max_{\theta \in \Theta} \sum_{k=1}^{n} \sum_{ij=11}^{rs} Z_k^{(ij)} \log \hat{g}_{ij}^{\theta}(\theta^t X_k) \mathbb{1}(X_k \in \mathcal{X}_n), \tag{4.2}$$

that is the value which maximizes log-likelihood (4.1) where the unknown links have been replaced by some nonparametric estimators, usually taken to be Nadaraya-Watson estimators like

$$\hat{g}_{ij}^{\theta}(\theta^t X_k) = \frac{\sum_{k' \neq k} Z_{k'}^{(ij)} K_1\left(\frac{\theta^t X_{k'} - \theta^t X_k}{h_1}\right)}{\sum_{k' \neq k} K_1\left(\frac{\theta^t X_{k'} - \theta^t X_k}{h_1}\right)}, \tag{4.3}$$

with K_1 a kernel function and h_1 a bandwidth. Note that (4.2) is therefore the maximizer of a so-called pseudo- or profile-likelihood. Also, notice that (3.4) and (4.3) are two different estimators : the former gives the final estimator of the concerned link function once an estimator of θ_0 has been determined, while the latter defines a primary estimator of the link, precisely needed for deriving the estimator $\hat{\theta}$. No confusion is possible, as the final estimator (3.4) does not arise in the derivations of this section. In (4.3), see that the observation X_k is excluded from the calculation of $\hat{g}_{ij}^{\theta}(\theta^t X_k)$ ("leave-one-out" estimator), for intuitively clear bias reasons as well as for technical facilities. Also, a trimming term $\mathbb{1}(X_k \in \mathcal{X}_n)$ is added in (4.2), in order to avoid eventual problems from the random denominator of (4.3). In particular, it permits the convergence of estimator $\hat{g}_{ij}^{\theta}$ towards g_{ij}^{θ}, uniformly in x and θ. Lee (1995) proposes the following trimming set[†]:

$$\mathcal{X}_n = \{x \in S_X : \xi_{n\alpha} \leq x^{(1)} \leq \xi_{n(1-\alpha)}\}, \tag{4.4}$$

[†]Actually, a discretized version of this trimming set is considered, for technical reasons, without changing its purpose.

where α is a specified small positive number, and $\xi_{n\alpha}$ is the αth sample quantile of the observed $\{X_k^{(1)}\}$. Note that this trimming indeed implies that the density of the index is bounded away from zero on the probability limit set of $\mathcal{X}_n$, i.e.

$$\mathcal{X} = \{x \in S_X : \xi_\alpha \leq x^{(1)} \leq \xi_{(1-\alpha)}\},$$

where ξ_α is the αth-quantile of $X^{(1)}$. See Lee (1995) end of section 2. Another remark is that K_1 needs to be a "higher-order" kernel function, in order to reduce the bias of estimator $\hat{\theta}$ induced by the kernel estimation of $\{g_{ij}\}$. However, the use of such kernels, which take on positive and negative values, can cause some trouble since the estimated probabilities are not ensured to be positive. A possible solution to this problem is to replace nonpositive $\hat{g}_{ij}^{\theta}(\theta^t X_k)$ by a specified small positive number in (4.2).

For deriving the asymptotic properties of the estimator (4.2), the following extra regularity conditions are made.

Assumption (SML1). *The kernel function K_1 has a bounded support S_1 and is two times differentiable with a second derivative satisfying a Lipschitz condition. Besides, $\int K_1(u)du = 1$, $\int |K_1(u)|du < \infty$, $\int u^q K_1(u)du = 0$ for $q = 1$ and 2 and $K_1(u) = 0$ for $u \in \partial S_1$, the boundary of S_1.*

Assumption (SML2). *The bandwidth sequence h_1 is such that $nh_1^5/\log n \to \infty$, $nh_1^4 \to \infty$ and $nh_1^6 \to 0$.*

Assumption (SML3). *The set Θ, defined in assumption 2, is compact and convex, and $\theta_0 \in \text{int}(\Theta)$.*

Assumption (SML4). *The density of the first component of X conditional to the other components, say $f_1(x^{(1)}|X^{(-1)} = x^{(-1)})$, is positive $\forall x = (x^{(1)}, x^{(-1)})$ in the interior of S_X, and is differentiable with respect to x_1 up to order 4.*

Now, define the following matrices:

$$W(\theta_0^t x) = \text{diag}\left(g_{11}(\theta_0^t x), g_{12}(\theta_0^t x), \ldots, g_{rs}(\theta_0^t x)\right)^{-1},$$

$$\Gamma(\theta^t x) = \begin{pmatrix} \frac{\partial g_{11}^{\theta}(\theta^t x)}{\partial \theta^{(2)}} & \frac{\partial g_{12}^{\theta}(\theta^t x)}{\partial \theta^{(2)}} & \cdots & \frac{\partial g_{rs}^{\theta}(\theta^t x)}{\partial \theta^{(2)}} \\ \vdots & & \ddots & \vdots \\ \frac{\partial g_{11}^{\theta}(\theta^t x)}{\partial \theta^{(p)}} & & \cdots & \frac{\partial g_{rs}^{\theta}(\theta^t x)}{\partial \theta^{(p)}}(\theta^t x) \end{pmatrix}, \tag{4.5}$$

$$\tilde{\Gamma}(\theta_0^t x) = \Gamma(\theta_0^t x)\, \mathbb{1}(x \in \mathcal{X}) - \mathbb{E}(\Gamma(\theta_0^t X)\, \mathbb{1}(X \in \mathcal{X})|\theta_0^t X),$$

$$\Sigma = \mathbb{E}\left(\Gamma(\theta_0^t X) W(\theta_0^t X) \Gamma(\theta_0^t X)^t\, \mathbb{1}(X \in \mathcal{X})\right)$$

and

$$\tilde{\Sigma} = \mathbb{E}\left(\tilde{\Gamma}(\theta_0^t X) W(\theta_0^t X) \tilde{\Gamma}(\theta_0^t X)^t \mathbb{1}(X \in \mathcal{X})\right).$$

Note that we have, for $q = 2, \ldots, p$,

$$\frac{\partial g_{ij}^{\theta}(\theta^t x)}{\partial \theta^{(q)}}|_{\theta=\theta_0} = g'_{ij}(\theta_0^t x)\left(x^{(q)} - \mathbb{E}(X^{(q)}|\theta_0^t X = \theta_0^t x)\right) \quad \forall\ (i,j)$$

while

$$\frac{\partial g_{ij}(\theta^t x)}{\partial \theta^{(1)}} \equiv 0 \quad \forall\ (i,j),$$

as $\theta^{(1)}$ is fixed to one for any $\theta \in \Theta$, what implies that matrix $\Gamma(\theta^t x)$ has only $(p-1)$ rows. Then, theorem 2 of Lee (1995) states :

Theorem 4.1. *Under Assumptions 1-4 and (SML1)-(SML4), the Semiparametric Maximum Likelihood estimator defined by (4.2) satisfies*

$$\sqrt{n}\left(\hat{\theta} - \theta_0\right) \xrightarrow{\mathcal{L}} \mathcal{N}_p(0, \Sigma_{SML}),$$

with $\Sigma_{SML} \doteq \left(\begin{array}{c|c} 0 & 0 \\ \hline 0 & \Sigma^{-1}\tilde{\Sigma}\Sigma^{-1} \end{array}\right)$.

The null first row and column of Σ_{SML} are obviously related to the first component of $\hat{\theta}$, fixed to one.

Given the asymptotic properties of its parametric counterpart, the question of the efficiency of this semiparametric maximum likelihood estimator is now adressed. A semiparametric estimator is efficient if its variance matrix equals the semiparametric variance bound, which is defined as the supremum of the Rao-Cramer bounds for all regular parametric submodels[†] of the considered semiparametric model. See Newey (1990) for discussion and detailed results about semiparametric efficiency bounds. Derivation of such bounds is not a trivial problem. Lee (1995) found that the semiparametric variance bound for estimators of $(\theta_0^{(2)}, \ldots, \theta_0^{(p)})$ in a polychotomous choice model is

$$V = \left[\mathbb{E}\left(\sum_{ij}^{rs} Z^{(ij)} \frac{\partial \log g_{ij}(\theta_0^t X)}{\partial \theta}\left(\frac{\partial \log g_{ij}(\theta_0^t X)}{\partial \theta}\right)^t\right)\right]^{-1},$$

where the differentiation with respect to θ starts from its second component. See that this can be written

$$V = \left[\mathbb{E}\left(\Gamma(\theta_0^t X) W(\theta_0^t X) \Gamma(\theta_0^t X)^t\right)\right]^{-1}, \tag{4.6}$$

[†] A parametric submodel is a parametric model that satisfies the semiparametric assumptions and contains the truth.

and that this bound would be attained by estimator (4.2) if the trimming terms $\mathbb{1}(X_k \in \mathcal{X}_n)$ in (4.2) were identically equal to 1. Indeed, in that case, the conditioning on the event $X \in \mathcal{X}$ would disappear from the different expectations in the expressions of $\tilde{\Gamma}$, $\tilde{\Sigma}$ and Σ, and as $\mathbb{E}(\Gamma(\theta_0^t X)|\theta_0^t X)$ equals zero, we would have $\Sigma = \tilde{\Sigma}$ and $\Sigma^{-1} = V$, so that

$$\Sigma_{SML} = \left(\begin{array}{c|c} 0 & 0 \\ \hline 0 & V \end{array}\right).$$

Hence, it is seen that the estimator (4.2) is not asymptotically efficient because of the sample information lost in the trimming process. Nevertheless, this loss of efficiency can be very small if the trimming quantile α appearing in (4.4) is very small, that is when the set $\mathcal{X}$ is very close to S_X. One can thus say that the proposed SML estimator is "nearly" asymptotically efficient.

Remark 4.1. *A natural idea should be to let α decreases to zero as n tends to infinity, to reach asymptotic efficiency. However, Lee (1995) points out that this design would create difficult analytic issues, and we do not develop further such idea. The quantile α is thus fixed to an arbitrarily small value.*

Remark 4.2. *As already mentionned, the higher order of the kernel K_1 can lead to practical troubles as the estimated probabilities are not ensured to belong to $[0, 1]$. Besides, criterion (4.2) becomes very unstable when using such kernels*[†]*. Therefore, it could be advantageous to work with a usual second order positive kernel, in order to stabilize the variance. In the simulation study in section 5, two SML estimators, one based on a second order kernel and another based on a higher order kernel, will be compared.*

4.2 Semiparametric Least Squares estimator (SLS)

Least squares methodology is another parametric regression method which can easily be adapted to our setting. As in the previous subsection, suppose at first that the link functions g_{ij} are known. Then θ_0 can be estimated via a classical non-linear (possibly weighted) least squares problem such as

$$\hat{\theta} = \arg\min_{\theta \in \Theta} \sum_{k=1}^{n} \sum_{ij=11}^{rs} w_{ij}(X_k)(Z_k^{(ij)} - g_{ij}(\theta^t X_k))^2, \tag{4.7}$$

that would yield a root-n consistent and asymptotically normal estimator, under usual mild conditions. The weighting allows to take an eventual heteroskedasticity into account, and to reach efficiency when optimal weights are used. When the link functions are unknown,

[†]Powell et al (1989) already pointed this out in the context of Average Derivatives Estimators.

problem (4.7) is solved with the g_{ij}'s replaced by consistent nonparametric estimators. As in Ichimura (1993), these links are estimated by their Nadaraya-Watson estimator

$$\hat{g}_{ij}^{\theta}(\theta^t X_k) = \frac{\sum_{k' \neq k} Z_{k'}^{(ij)} K_2\left(\frac{\theta^t X_{k'} - \theta^t X_k}{h_2}\right)}{\sum_{k' \neq k} K_2\left(\frac{\theta^t X_{k'} - \theta^t X_k}{h_2}\right)}, \tag{4.8}$$

with K_2 a kernel function and h_2 a bandwidth. As in (4.3), the observation X_k is excluded from the calculation of $\hat{g}_{ij}^{\theta}(\theta^t X_k)$. Ichimura (1993) added trimming terms $\mathbb{1}(X_k \in \mathcal{X}_n)$ in (4.7) and in (4.8), where

$$\mathcal{X}_n = \{x \in S_X : \exists x' \in \mathcal{X} : \|x - x'\| \leq 2h\}$$

and $\mathcal{X}$ is a compact subset of S_X such that the density of the index $\theta^t X$ is bounded away from zero, for any $\theta \in \Theta$. This prevents to find the denominator of (4.8) arbitrarily close to 0 as n increases, and permits to establish uniform convergence of the estimator towards its probability limit. Note that no guide is given on how to select the set $\mathcal{X}$ in Ichimura (1993), so that we propose to use the trimming scheme of Lee (1995) (see the previous subsection). This does not alter Ichimura's arguments, as the final aim of uniform convergence is the same.

Remark 4.3. *In the original version of Ichimura (1993), other extra weights appear in the definition of estimator (4.8), in order to increase efficiency and to reduce bias. However, as stated by its section 6, this inner weighting has no effect when the conditional variance of the response given the covariates only depends on the index. As this is the case in the considered setting - we have* $\mathbb{V}\mathrm{ar}(Z^{(ij)}|X) = g_{ij}(\theta_0^t X)(1 - g_{ij}(\theta_0^t X))$ *for any* (i, j) *by Assumption 1 -, these weights are not considered here.*

The Semiparametric Least Squares estimator of θ_0 is thus given by

$$\hat{\theta} = \arg\min_{\theta \in \Theta} \sum_{k=1}^{n} \mathbb{1}(X_k \in \mathcal{X}_n) \sum_{ij=11}^{rs} w_{ij}(X_k)(Z_k^{(ij)} - \hat{g}_{ij}^{\theta}(\theta^t X_k))^2. \tag{4.9}$$

Deriving asymptotic properties for this estimator requires the following conditions.

Assumption (SLS1). *The set* Θ*, defined in assumption 2, is compact, and* $\theta_0 \in \mathrm{int}(\Theta)$.

Assumption (SLS2). *The functions* $f_\theta(u)$ *and* $g_{ij}^{\theta}(u)$ *are three times continuously differentiable with respect to* u *and the third derivatives satisfy suitable Lipschitz conditions,* $\forall u \in S_\theta$ *uniformly in* θ.

Assumption (SLS3). *The kernel K_2 has support $[-1,1]$, is two times continuously differentiable, with the second derivative satisfying a Lipschitz condition, is such that $\int K(u)du = 1$ and $\int uK(u)du = 0$.*

Assumption (SLS4). *The bandwidth sequence h_2 satisfies $(\log h_2)/(nh_2^3) \to 0$ and $nh_2^8 \to 0$ as $n \to \infty$.*

Assumption (SLS5). *The weight functions w_{ij} appearing in (4.9) are bounded and positive.*

As our Assumptions 1-4 and (SLS1)-(SLS4) imply Assumptions 5.1-5.6 of Ichimura (1993) for any (i,j), we directly find, from his Lemma 5.1, that for any $\epsilon > 0$,

$$\mathbb{P}\left(\sup_{x\in\mathcal{X}}\sup_{\theta\in\Theta}|\hat{g}_{ij}^{\theta}(\theta^t x) - g_{ij}^{\theta}(\theta^t x)| > \epsilon\right) \to 0.$$

Moreover, Lemmas 5.5-5.10 of Ichimura (1993) also hold as such, while Lemmas 5.2-5.3 and Theorem 5.1 hold, with very slight modifications[†], for the criterion appearing in (4.9). Theorem 5.2, stating the root-n consistency as well as the asymptotic normality of the SLS estimator can also very easily be adapted. The matrices of interest are slightly modified in an evident way, and become

$$V = \mathbb{E}\left(\Gamma(\theta_0^t X)\mathcal{W}(X)(\Gamma(\theta_0^t X))^t \mathbb{1}(X \in \mathcal{X})\right) \tag{4.10}$$

and

$$\tilde{V} = \mathbb{E}\left(\Gamma(\theta_0^t X)\mathcal{W}(X)\Omega(\theta_0^t X)\mathcal{W}(X)(\Gamma(\theta_0^t X))^t \mathbb{1}(X \in \mathcal{X})\right), \tag{4.11}$$

with

$$\Gamma(\theta^t x) = \begin{pmatrix} \frac{\partial g_{11}^{\theta}(\theta^t x)}{\partial \theta^{(2)}} & \frac{\partial g_{12}^{\theta}(\theta^t x)}{\partial \theta^{(2)}} & \cdots & \frac{\partial g_{rs}^{\theta}(\theta^t x)}{\partial \theta^{(2)}} \\ \vdots & & \ddots & \vdots \\ \frac{\partial g_{11}^{\theta}(\theta^t x)}{\partial \theta^{(p)}} & & \cdots & \frac{\partial g_{rs}^{\theta}(\theta^t x)}{\partial \theta^{(p)}}(\theta^t x) \end{pmatrix},$$

as in (4.5),

$$\Omega(\theta^t x) = \mathrm{diag}(g(\theta^t x)) - g(\theta^t x)g(\theta^t x)^t \tag{4.12}$$

and

$$\mathcal{W}(x) = \mathrm{diag}\left(w_{11}(x), w_{12}(x), \ldots, w_{rs}(x)\right).$$

Remark 4.4. *Ichimura (1993) states that the presence of a tricky conditioning to $X \in \mathcal{X}$ can easily be eliminated by letting $\mathcal{X}$ grow very slowly with n. Nevertheless, Remark 4.1 seems to indicate the contrary. Therefore, the set $\mathcal{X}$ is here maintained fix.*

[†]With respect to the original proofs, only a no effect sum over ij is added in the criterion to be minimized.

Finally, it follows :

Theorem 4.2. *Under Assumptions 1-4 and (SLS1)-(SLS5), the Semiparametric Least Squares estimator defined by (4.9) satisfies*

$$\sqrt{n}\left(\hat{\theta}-\theta_0\right) \xrightarrow{\mathcal{L}} \mathcal{N}_p\left(0,\Sigma_{SLS}\right),$$

with $\Sigma_{SLS} \doteq \left(\begin{array}{c|c} 0 & 0 \\ \hline 0 & V^{-1}\tilde{V}V^{-1} \end{array}\right)$, *and matrices* V *and* $\tilde{V}$ *defined by (4.10) and (4.11).*

The problem of selecting the weights w_{ij} is now adressed. As usual, this weighting is introduced in the minimization problem (4.9) in order to increase efficiency. What follows shows that the choice

$$\mathcal{W}(x) = W(\theta_0^t x) = \operatorname{diag}\left(g_{11}(\theta_0^t x), g_{12}(\theta_0^t x), \ldots, g_{rs}(\theta_0^t x)\right)^{-1}$$

leads to a nearly efficient semiparametric estimator $\hat{\theta}$. The "nearly" term refers to the conditionning to the event $X \in \mathcal{X}$, see the comment following (4.6) at the end of the previous section. With this choice of matrix W, direct computations lead to

$$W(\theta_0^t x)\Omega(\theta_0^t x)W(\theta_0^t x) = W(\theta_0^t x) - 1_{rs}1_{rs}^t,$$

with 1_{rs} the rs-vector whose components are all equal to 1. Therefore,

$$\tilde{V} = \mathbb{E}\left(\Gamma(\theta_0^t X)W(\theta_0^t X)(\Gamma(\theta_0^t X))^t \mathbb{1}(X \in \mathcal{X})\right) - \mathbb{E}\left(\Gamma(\theta_0^t X)1_{rs}1_{rs}^t(\Gamma(\theta_0^t X))^t \mathbb{1}(X \in \mathcal{X})\right),$$

which is equal to (4.10), since $\Gamma(\theta_0^t x)1_{rs}$ is identically the null vector. Therefore, the non-zero part of Σ_{SLS} equals, up to the trimming term, the bound defined in (4.6). In applications, these weights can be replaced by consistent estimators without affecting the asymptotic distribution of the estimator, and therefore its asymptotic efficiency (usual result in M-estimation, see e.g. comments in Newey and Stoker (1993)). Easy to implement consistent estimators of the weights are given by the following two-steps procedure. In the first step, estimate θ_0 by the $n^{1/2}$-consistent, asymptotically normal but inefficient unweighted ($w \equiv 1$) SLS estimator, say $\hat{\theta}_1$, and build the corresponding Nadaraya-Watson estimators $\hat{g}_{ij}^{\hat{\theta}_1}(u)$, given by (3.4). In the second step, set $w_{ij}(X_k) = 1/\hat{g}_{ij}^{\hat{\theta}_1}(\hat{\theta}_1^t X_k)$ and derive the (nearly) efficient weighted SLS estimator $\hat{\theta}$. If some of the $\{\hat{g}_{ij}^{\hat{\theta}_1}(\hat{\theta}_1^t X_k)\}$ are zero, just replace them by a specified small positive number.

4.3 Average Derivatives Estimator (ADE)

Average Derivatives methods were introduced by Härdle and Stoker (1989) and Powell et al (1989). In the classical SIM context (2.1)-(2.2), they rest on the evident following fact:

$$\nabla m(x) = g'(\theta_0^t x)\theta_0 \quad \forall x \in S_X,$$

what induces that

$$\delta_w \doteq \mathbb{E}\left(w(X)\nabla m(X)\right) = \mathbb{E}\left(w(X)g'(\theta_0^t X)\right)\theta_0 \tag{4.13}$$

for any bounded continuous weight function w. Hence, any vector δ_w, called average derivative, is proportional to θ_0 provided that $\mathbb{E}\left(w(X)g'(\theta_0^t X)\right)$ is not zero, so that any estimator of δ_w easily leads to an estimator of θ_0. In practice, the choice $w(x) = f(x)$ appears to be judicious, as it permits to avoid the presence of a random denominator when estimating the average derivative (see Powell et al (1989)). Another important remark is that considering the gradient of m implies that X is a continuously distributed random vector[†].

In the setting concerned by Assumptions 1 and 2, we have that

$$\nabla\pi_{ij}(x) = g'_{ij}(\theta_0^t x)\theta_0 \quad \forall x \in S_X, \forall (i,j),$$

so that we can actually define rs (density-weighted) average derivatives

$$\delta_{ij} \doteq \mathbb{E}(f(X)\nabla\pi_{ij}(X)) = \mathbb{E}\left(f(X)g'_{ij}(\theta_0^t X)\right)\theta_0, \tag{4.14}$$

each proportional to θ_0, that is rs collinear vectors. Let Δ be the $(p \times rs)$-matrix

$$\Delta = (\delta_{11}, \delta_{12}, \ldots, \delta_{rs}), \tag{4.15}$$

that can clearly be written, in view of (4.14), as

$$\Delta = \theta_0\alpha^t, \tag{4.16}$$

with

$$\alpha = \left(\mathbb{E}\left(f(X)g'_{11}(\theta_0^t X)\right), \ldots, \mathbb{E}\left(f(X)g'_{rs}(\theta_0^t X)\right)\right)^t.$$

In addition, due to the identifiability condition $\theta_0^{(1)} = 1$, we directly have that

$$\alpha = (\delta_{11}^{(1)}, \delta_{12}^{(1)}, \ldots, \delta_{rs}^{(1)})^t.$$

Multiplying both sides of (4.16) by α, it follows

$$\theta_0\alpha^t\alpha = \Delta\alpha,$$

so that we have

$$\theta_0 = \frac{1}{\|\alpha\|^2}\Delta\alpha$$

[†]An extension to the case where some components of X are discrete is possible, see Horowitz and Härdle (1996). Nevertheless, such ideas are not pursued further in this work.

if $\|\alpha\| \neq 0$, what will be the case if at least one g_{ij} is such that $\mathbb{E}\left(f(X)g'_{ij}(\theta_0^t X)\right) \neq 0$. Finally, as α is simply the first line of Δ, it is seen that any estimator $\widehat{\Delta}$ of Δ easily leads to an estimator of θ_0 :

$$\hat{\theta} = \frac{1}{\|\hat{\alpha}\|^2}\widehat{\Delta}\hat{\alpha}. \tag{4.17}$$

Now, estimating Δ amounts to estimate any δ_{ij} defined by (4.14), so that the results of Powell et al (1989) in the classical context (4.13) are directly applicable for each of those estimations. In that purpose, the following conditions are needed. Let $P = (p+4)/2$ if p is even, and $P = (p+3)/2$ if p is odd.

Assumption (ADE1). *The random vector X is continuously distributed, and no component of X is functionally determined by others components. Besides, the support S_X of X is a convex subset of $\mathbb{R}^p$.*

Assumption (ADE2). *The density of X, denoted f, is continuously differentiable in the components of X, and $f(x) = 0$ for all $x \in \partial S_X$, where ∂S_X denotes the boundary of S_X. Also, the components of the matrix $\mathbb{E}\left(\nabla f(X)X^t\right)$ have finite second moment. In addition, all partial derivatives of f of order $P+1$ exist, and the expectation $\mathbb{E}\left(\frac{\partial^\rho f}{\partial x^{(l_1)}\ldots\partial x^{(l_p)}}(X)\right)$ exists for any $(l_1,\ldots,l_p)$ such that $l_1+\ldots+l_p = \rho$ and any $\rho \leq P+1$.*

Assumption (ADE3). *There exists at least one cell, say (i,j), for which $\mathbb{E}\left(f(X)g'_{ij}(\theta_0^t X)\right)$ is not zero.*

Remark that our Assumptions 1-4 and (ADE1)-(ADE3) imply Assumptions 1-3 of Powell et al (1989). Then, it easily follows by integration by parts (see their Lemma 2.1) that

$$\delta_{ij} = -2\mathbb{E}(Z^{(ij)}\nabla f(X)). \tag{4.18}$$

The idea is then to estimate this δ_{ij} by a sample analogue, where the unknown f is replaced by a consistent nonparametric estimate, e.g. the (multivariate) kernel density estimator

$$\hat{f}(x) = \frac{1}{nh_3^p}\sum_{k=1}^{n} K_3\left(\frac{x - X_k}{h_3}\right), \tag{4.19}$$

where K_3 and h_3 are a kernel function and a bandwidth sequence[†] satisfying the following conditions.

[†]For simplicty, a common bandwidth is used for all components of $x - X_k$. Different bandwidths, forming a bandwidth matrix, could also be used to take into account possible different scales for those components.

Assumption (ADE4). *The kernel function $K_3 : S_3 \subset \mathbb{R}^p \to \mathbb{R}$ is bounded, differentiable, symmetric*[†]*, on a convex support S_3, such that all moments of $K_3(x)$ of order P exist and $K_3(x) = 0$ for any x on the boundary of S_3. Besides, we have*

$$\int K_3(x)dx = 1, \int x^{(l_1)} \dots x^{(l_\rho)} K_3(x) = 0 \text{ for } \rho < P, \text{ and}$$
$$\int x^{(l_1)} \dots x^{(l_\rho)} K_3(x) \neq 0 \text{ for } \rho = P.$$

Assumption (ADE5). *The bandwidth sequence h_3 satisfies $nh_3^{p+2} \to \infty$ and $nh_3^{2P} \to 0$.*

The sample analogue of (4.18) is now given by

$$\hat{\delta}_{ij} = -\frac{2}{n} \sum_{k=1}^{n} Z_k^{(ij)} \nabla \hat{f}(X_k), \tag{4.20}$$

where $\nabla \hat{f}(X_k)$ is the gradient of the "leave-one-out" estimator (4.19) at X_k:

$$\nabla \hat{f}(X_k) = \frac{1}{(n-1)h_3^{p+1}} \sum_{k' \neq k} \nabla K_3 \left(\frac{X_k - X_{k'}}{h_3} \right).$$

By Theorem 3.3 of Powell et al (1989), it follows that, for any (i, j),

$$\sqrt{n}(\hat{\delta}_{ij} - \delta_{ij}) \xrightarrow{\mathcal{L}} \mathcal{N}_p(0, \Sigma_{ij}), \tag{4.21}$$

with

$$\Sigma_{ij} = 4\mathbb{E}(r_{ij}(X, Z) r_{ij}(X, Z)^t) - 4\delta_{ij}\delta_{ij}^t \tag{4.22}$$

and

$$r_{ij}(x, z) = f(x)\nabla \pi_{ij}(x) - (z^{(ij)} - \pi_{ij}(x))\nabla f(x).$$

Remark 4.5. *Assumptions (ADE4) and (ADE5) are designed in order to make the asymptotic bias of the estimator δ_{ij} vanish at rate $\sqrt{n}$. In particular, see that Assumption (ADE4) requires the use of a "higher-order" kernel of order P.*

The next lemma gives the variance-covariance matrix Σ_{ij} in a more tractable way.

Lemma 4.3.1. *The variance-covariance matrix Σ_{ij} given in (4.22) can be written*

$$\Sigma_{ij} = 4\,\mathbb{V}\text{ar}\left(f(X) g'_{ij}(\theta_0^t X)\right) \theta_0 \theta_0^t + 4\mathbb{E}\left(g_{ij}(\theta_0^t X)(1 - g_{ij}(\theta_0^t X)) \nabla f(X) \nabla f(X)^t\right). \tag{4.23}$$

[†]In the sense $K_3(u) = K_3(-u)$.

Proof: We have

$$\begin{aligned} r_{ij}(x,z) &= f(x)\nabla\pi_{ij}(x) - (z^{(ij)} - \pi_{ij}(x))\nabla f(x) \\ &= f(x)g'_{ij}(\theta_0^t x)\theta_0 - (z^{(ij)} - g_{ij}(\theta_0^t x))\nabla f(x), \end{aligned}$$

so that

$$\begin{aligned} r_{ij}(x,z)r_{ij}(x,z)^t = f^2(x)g'^2_{ij}(\theta_0^t x)\theta_0\theta_0^t + (z^{(ij)} - g_{ij}(\theta_0^t x))^2\nabla f(x)\nabla f(x)^t \\ - f(x)g'_{ij}(\theta_0^t x)(z^{(ij)} - g_{ij}(\theta_0^t x))\left(\theta_0\nabla f(x)^t + \nabla f(x)\theta_0^t\right). \end{aligned}$$

As $g_{ij}(\theta_0^t X) = \mathbb{E}(Z^{(ij)}|X)$, we have

$$\begin{aligned} \mathbb{E}(r_{ij}(X,Z)r_{ij}(X,Z)^t) &= \mathbb{E}\left(f^2(X)g'^2_{ij}(\theta_0^t X)\right)\theta_0\theta_0^t + \mathbb{E}\left((Z^{(ij)} - g_{ij}(\theta_0^t X))^2\nabla f(X)\nabla f(X)^t\right) \\ &= \mathbb{E}\left(f^2(X)g'^2_{ij}(\theta_0^t X)\right)\theta_0\theta_0^t + \mathbb{E}\left(\mathbb{V}\mathrm{ar}(Z^{(ij)}|X)\nabla f(X)\nabla f(X)^t\right) \\ &= \mathbb{E}\left(f^2(X)g'^2_{ij}(\theta_0^t X)\right)\theta_0\theta_0^t + \mathbb{E}\left(g_{ij}(\theta_0^t X)(1 - g_{ij}(\theta_0^t X))\nabla f(X)\nabla f(X)^t\right). \end{aligned}$$

Now, as $\delta_{ij}\delta_{ij}^t = \left(\mathbb{E}(f(X)g'_{ij}(\theta_0^t X))\right)^2\theta_0\theta_0^t$, (4.23) directly follows, by definition (4.22) of Σ_{ij}. □

Now, remind definition (4.15) of Δ, and let Δ^* be the vectorized version of it, in the following sense:

$$\Delta^* = (\delta_{11}^t, \delta_{12}^t, \ldots, \delta_{rs}^t)^t,$$

that is Δ^* is a (prs)-vector. Denote also $\widehat{\Delta}^*$ the estimated version of Δ^*, from estimators (4.20). From (4.21), we have

$$\sqrt{n}(\widehat{\Delta}^* - \Delta^*) \xrightarrow{\mathcal{L}} \mathcal{N}_{prs}(0, \Sigma_\Delta),$$

with

$$\Sigma_\Delta = \begin{pmatrix} \Sigma_{11} & \Sigma_{11,12} & & \cdots & & \Sigma_{11,rs} \\ \Sigma_{12,11} & \Sigma_{12} & & & & \\ & & \ddots & & \Sigma_{i_2j_2,i_1j_1} & \\ \vdots & & & \Sigma_{ij} & & \\ & & \Sigma_{i_1j_1,i_2j_2} & & \ddots & \\ \Sigma_{rs,11} & & & & & \Sigma_{rs} \end{pmatrix},$$

where matrices of type Σ_{ij} are defined by (4.23), and matrices of type $\Sigma_{i_1j_1,i_2,j_2}$ are given by†

$$\begin{aligned} \Sigma_{i_1j_1,i_2j_2} = 4\,\mathbb{C}\mathrm{ov}\left(f(X)g'_{i_1j_1}(\theta_0^t X), f(X)g'_{i_2j_2}(\theta_0^t X)\right)\theta_0\theta_0^t \\ - 4\mathbb{E}\left(g_{i_1j_1}(\theta_0^t X)g_{i_2j_2}(\theta_0^t X)\nabla f(X)\nabla f(X)^t\right). \quad (4.24) \end{aligned}$$

†This result can be shown exactly the same way as the proof of Lemma 4.3.1, seeing that $\Sigma_{i_1j_1,i_2,j_2}$ is the covariance matrix between $2r_{i_1j_1}(X,Z)$ and $2r_{i_2j_2}(X,Z)$.

Then, define the transformation $\phi : \mathbb{R}^{prs} \to \mathbb{R}^p$, given by

$$\phi(\Delta^*) = \frac{1}{\sum_{ij=11}^{rs} \delta_{ij}^{(1)^2}} \sum_{ij=11}^{rs} \delta_{ij}^{(1)} \delta_{ij}, \tag{4.25}$$

and see that $\theta_0 = \phi(\Delta^*)$, while from (4.17) we have also $\hat{\theta} = \phi(\widehat{\Delta}^*)$. By a usual Delta method argument, it directly follows that

$$\sqrt{n}\left(\hat{\theta} - \theta_0\right) \xrightarrow{\mathcal{L}} \mathcal{N}_p(0, \Sigma_{ADE}),$$

with $\Sigma_{ADE} \doteq \phi' \Sigma_\Delta \phi'^t$ and ϕ' being the $(p \times prs)$-matrix of partial derivatives of the transformation ϕ at Δ^*, that is $\phi'_{q_1,(q_2,ij)} = \frac{\partial \phi^{(q_1)}}{\partial \delta_{ij}^{(q_2)}}(\Delta^*)$. Differentiation of (4.25) and a bit algebra lead to

$$\phi'_{q_1,(q_2,ij)} = \begin{cases} \frac{\delta_{ij}^{(1)}}{\sum_{ij=11}^{rs} \delta_{ij}^{(1)^2}} & \text{if } q_1 \neq 1, q_2 = q_1 \\ -\theta_0^{(q_1)} \frac{\delta_{ij}^{(1)}}{\sum_{ij=11}^{rs} \delta_{ij}^{(1)^2}} & \text{if } q_1 \neq 1, q_2 = 1 \\ 0 & \text{otherwise} \end{cases}.$$

Call β_{ij} the quantity $\frac{\delta_{ij}^{(1)}}{\sum_{ij=11}^{rs} \delta_{ij}^{(1)^2}}$, and see that ϕ' can be written as a block matrix $\phi' = \begin{pmatrix} \beta_{11}\tilde{\phi'} & \beta_{12}\tilde{\phi'} & \dots & \beta_{rs}\tilde{\phi'} \end{pmatrix}$, with

$$\tilde{\phi'} = \begin{pmatrix} 0 & 0 & \cdots & 0 \\ -\theta_0^{(2)} & 1 & & 0 \\ \vdots & & \ddots & \\ -\theta_0^{(p)} & 0 & & 1 \end{pmatrix}. \tag{4.26}$$

As it can easily be checked that $\tilde{\phi'}\theta_0 = 0$, any contribution of the first term of the right-hand sides of (4.23) and (4.24), for all (i, j), vanishes in Σ_{ADE}, so that it remains, after a little bit more algebraic work,

$$\Sigma_{ADE} = \frac{4}{\left(\sum_{ij=11}^{rs} \delta_{ij}^{(1)^2}\right)^2} \mathbb{E}\left(\left[\sum_{ij=11}^{rs} \delta_{ij}^{(1)^2} g_{ij}(\theta_0^t X)(1 - g_{ij}(\theta_0^t X)) \right.\right.$$
$$\left.\left. - \sum_{i_1 j_1} \sum_{i_2 j_2 \neq i_1 j_1} \delta_{i_1 j_1}^{(1)} \delta_{i_2 j_2}^{(1)} g_{i_1 j_1}(\theta_0^t X) g_{i_2 j_2}(\theta_0^t X) \right] \tilde{\phi'} \nabla f(X) \nabla f(X)^t \tilde{\phi'}^t \right),$$

that is

$$\Sigma_{ADE} = 4\mathbb{E}\left(\frac{\alpha^t \Omega(\theta_0^t X)\alpha}{\alpha^t \alpha} \tilde{\phi'} \nabla f(X) \nabla f(X)^t \tilde{\phi'}^t \right), \tag{4.27}$$

with the matrix $\Omega(\theta_0^t x)$ already defined in (4.12).

Hence, we finally state the following result:

Theorem 4.3. *Under assumptions 1-4 and (ADE1)-(ADE5), the Average Derivative estimator $\hat{\theta}$ defined by (4.17) satisfies*

$$\sqrt{n}\left(\hat{\theta} - \theta_0\right) \xrightarrow{\mathcal{L}} \mathcal{N}_p(0, \Sigma_{ADE}),$$

with variance-covariance matrix Σ_{ADE} given by (4.27).

4.4 Sliced Inverse Regression estimator (SIR)

Sliced Inverse Regression, or Slicing Regression, is another direct estimation scheme, introduced by Duan and Li (1991). As it will be seen, this procedure does not need any kernel estimates of the link function or other, what makes it very interesting in practice, since a.o. the delicate problem of bandwidth choice is avoided. However, it rests on an extra assumption on the design:

Assumption (SIR1). *The vector of regressors X follows a nondegenerate elliptically symmetric distribution, i.e. from its mean vector μ_X and its variance-covariance matrix Σ_X, its density f can be written*

$$f(x) = k_p |\Sigma_X|^{-1/2} v\left((x-\mu_X)^t \Sigma_X^{-1}(x-\mu_X)\right), \tag{4.28}$$

where v is a one-dimensional real-valued function independent of p, and k_p is a scalar proportionality constant.

For example, the multivariate normal distribution is elliptically symmetric, with $v(\cdot) = \exp(-\cdot/2)$ and $k_p = (2\pi)^{-p/2}$, so that this assumption is often admissible. Besides, Duan and Li (1991) provide a bound on the bias appearing when this design assumption is violated.

In the classical SIM setting (2.1)-(2.2), the method takes advantage of the following two facts. First, contrary to the usual regression function $m(x) = E(Y|X=x)$ whose estimation hardly suffers from the curse of dimensionality for large dimensional X, the inverse regression function $\xi(y) \doteq E(X|Y=y)$ can be component by component safely estimated, as Y is one-dimensional. Second, the assumed elliptically symmetric distribution of X permits to draw an interesting relationship between this inverse regression function and the vector θ_0, what will be the base of the estimation procedure. Moreover, it appears that the function $\xi(y)$ can be estimated very crudely without affecting the performance of the estimator of θ_0, what is of interest since no kernel-type estimation has to be considered. In fact, a step function is used, after having partitionned the range of Y into slices.

The concept of slices in the response range is especially well adapted to the context defined by Assumption 1. Indeed, the vector of responses Z essentially defines rs groups of individuals,

one for each cell of the contingency table. There are therefore a priori existing "slices" in the observations, and the "crude" sliced inverse regression estimation is therefore the best possible. It becomes $\xi(z) = E(X|Z = z)$, which actually takes only on rs values

$$\xi_{ij} = \mathbb{E}(X|Z^{(ij)} = 1).$$

Then, from assumption (SIR1), it can be shown (see results of Duan and Li (1991)) that

$$\xi_{ij} = \mu_X + \gamma_{ij}\Sigma_X\theta_0 \tag{4.29}$$

with

$$\gamma_{ij} = \frac{\mathbb{E}\left(\theta_0^t(X - \mu_X)|Z^{(ij)} = 1\right)}{\theta_0^t\Sigma_X\theta_0}.$$

It directly follows from (4.29) that

$$\theta_0 = \frac{1}{\gamma_{ij}}\Sigma_X^{-1}(\xi_{ij} - \mu_X)$$

for any cell (i, j) such that $\gamma_{ij} \neq 0$. Therefore, due to the identifiability constraint $\theta_0^{(1)} = 1$, estimating one ξ_{ij}, μ_X and Σ_X is sufficient for estimating θ_0. In addition, in order to combine information from all cells, define $\Sigma_\xi = \mathbb{C}\text{ov}(\xi(Z))$. Corollary 2.2 of Duan and Li (1991) states that

$$\theta_0 = \arg\max_{\theta\in\Theta}\frac{\theta^t\Sigma_\xi\theta}{\theta^t\Sigma_X\theta}, \tag{4.30}$$

and that this maximizer is unique if and only if there exists at least one $\gamma_{ij} \neq 0$. Consistency of the procedure therefore requires the following assumption:

Assumption (SIR2). *There exists at least one cell of the table, say (i, j), such that $\mathbb{E}\left(\theta_0^t(X - \mu_X)|Z^{(ij)} = 1\right) \neq 0$.*

This condition simply requires that in at least one cell, the expectation of the index is different to the global expectation of it. Note also that (4.30) amounts to say that θ_0 is the principal eigenvector of $\Sigma_X^{-1}\Sigma_\xi$, belonging to Θ. It is furthermore the only eigenvector associated to a non-zero eigenvalue, as Σ_ξ has rank one, from (4.29).

Now, Σ_X will be estimated by the usual sample covariance matrix $\hat{\Sigma}_X$ for the observed $\{X_k\}$, while Σ_ξ will be estimated the following way. First, take

$$\hat{\xi}_{ij} = \frac{\sum_{k=1}^n X_k Z_k^{(ij)}}{\sum_{k=1}^n Z_k^{(ij)}}$$

as estimate of the inverse regression function in the (ij)th cell of the table, that is the (ij)th slice. Then, introducing the following notations

$$\gamma = (\gamma_{11}, \gamma_{12}, \ldots, \gamma_{rs})^t \qquad \Xi = (\xi_{11}, \xi_{12}, \ldots, \xi_{rs}) \qquad \hat{\Xi} = (\hat{\xi}_{11}, \hat{\xi}_{12}, \ldots, \hat{\xi}_{rs})$$
$$\hat{p}_{ij} = \frac{\sum_{k=1}^{n} Z_k^{(ij)}}{n} \qquad \hat{p} = (\hat{p}_{11}, \hat{p}_{12}, \ldots, \hat{p}_{rs})^t \qquad \hat{\Omega} = \mathrm{diag}(\hat{p}) - \hat{p}\hat{p}^t \qquad \Omega = \mathrm{diag}(\pi) - \pi\pi^t,$$

we take

$$\hat{\Sigma}_\xi = \hat{\Xi}\hat{\Omega}\hat{\Xi}^t.$$

Finally, we define the SIR estimator as

$$\hat{\theta} = \arg\max_{\theta \in \Theta} \frac{\theta^t \hat{\Sigma}_\xi \theta}{\theta^t \hat{\Sigma}_X \theta}, \tag{4.31}$$

or as the principal eigenvector of $\hat{\Sigma}_X^{-1}\hat{\Sigma}_\xi$ belonging to Θ.

Remark 4.6. *$\hat{\Sigma}_\xi$ can be interpreted as a weighted covariance matrix for the matrix $\hat{\Xi}$, with weight matrix $\hat{\Omega}$, an estimated version of Ω, the matrix defining the distribution of the individuals through the different cells of the table. Another weighting could be considered, but Duan and Li (1991, section 5) show that this weighting is optimal when the design distribution is normal, so that we only consider it here.*

Main results of Duan and Li (1991) focus on β_0, the vector collinear to θ_0 normalized as $\beta_0^t \Sigma_X \beta_0 = 1$, and $\hat{\beta}$ the vector collinear to $\hat{\theta}$ such that $\hat{\beta}^t \hat{\Sigma}_X \hat{\beta} = 1$. They show that

$$\sqrt{n}(\hat{\beta} - \beta_0) \xrightarrow{\mathcal{L}} \mathcal{N}_p(0, V),$$

with $V = S(\Sigma_X^{-1} - \beta_0\beta_0^t) + T\beta_0\beta_0^t$, for some defined constants S and T.

In order to adapt those results to our estimator $\hat{\theta}$, define the transformation $\phi : \mathbb{R}^p \to \mathbb{R}^p$, such that

$$\phi(\beta) = \frac{\beta}{\beta^{(1)}},$$

and see that $\theta_0 = \phi(\beta_0)$ and $\hat{\theta} = \phi(\hat{\beta})$. Now use Delta method arguments to derive

$$\sqrt{n}\left(\hat{\theta} - \theta_0\right) \xrightarrow{\mathcal{L}} \mathcal{N}_p(0, \Sigma_{SIR}),$$

with $\Sigma_{SIR} \doteq \phi' V \phi'^t$ and ϕ' the $p \times p$ matrix of partial derivatives of ϕ, taken at β_0, which can easily be shown to be

$$\phi' = \frac{1}{\beta_0^{(1)}} \begin{pmatrix} 0 & 0 & \cdots & 0 \\ -\theta_0^{(2)} & 1 & & 0 \\ \vdots & & \ddots & \\ -\theta_0^{(p)} & 0 & & 1 \end{pmatrix}.$$

Now, remark first that, as $\beta_0^t \Sigma_X \beta_0 = 1$, we have that $\beta_0^{(1)} = \frac{1}{\sqrt{\theta_0^t \Sigma_X \theta_0}}$. Second, as $\phi' \beta_0 = 0$, it follows

$$\Sigma_{SIR} = S(\theta_0^t \Sigma_X \theta_0) \tilde{\phi}' \Sigma_X^{-1} \tilde{\phi}'^t, \tag{4.32}$$

where $\tilde{\phi}'$ is the matrix defined in (4.26). Third, from the design assumption (SIR1), it can be shown that

$$\mathbb{V}\mathrm{ar}(X|\theta_0^t X) = w(\theta_0^t X) \left(\Sigma_X - \frac{1}{\theta_0^t \Sigma_X \theta_0} \Sigma_X \theta_0 \theta_0^t \Sigma_X \right),$$

where w is a scalar function such that $\mathbb{E}(w(\theta_0^t X)) = 1$. Note that if the design is normal, then $w \equiv 1$. Denote

$$w_{ij} = \mathbb{E}(w(\theta_0^t X)|Z^{(ij)} = 1)$$

and define also

$$c_{ij} = \frac{\mathbb{E}(w(\theta_0^t X)\theta_0^t (X - \mu_X)|Z^{(ij)}) = 1}{\theta_0^t \Sigma_X \theta_0}, \; c = (c_{11}, c_{12}, \ldots, c_{rs})^t \text{ and } \eta = \Omega\gamma.$$

From results of Duan and Li (1991), we find that the constant S can be written as $S = A + B - 2C$, with

$$A = \frac{\sum_{ij=11}^{rs} w_{ij} \eta_{ij}^2 / \pi_{ij}}{(\eta^t \gamma)^2}, \quad B = \frac{\mathbb{E}(w(\theta_0^t X)(\theta_0^t (X - \mu_X))^2)}{\theta_0^t \Sigma_X \theta_0}, \quad C = \frac{\eta^t c}{\eta^t \gamma}.$$

We can now state :

Theorem 4.4. *Under Assumptions 1-2 and (SIR1)-(SIR2), the Sliced Inverse Regression estimator defined by (4.31) satisfies*

$$\sqrt{n} \left(\hat{\theta} - \theta_0 \right) \xrightarrow{\mathcal{L}} \mathcal{N}_p(0, \Sigma_{SIR}),$$

with Σ_{SIR} given in (4.32).

Remark 4.7. *See that the link functions $\{g_{ij}\}$ do not arise in the asymptotic distribution of the SIR estimator, what was expected as those quantities are nowhere used in the procedure. Only the difference of behaviour of X in the different cells plays a role.*

Remark 4.8. *As already mentionned, in case of normal design, we have $w \equiv 1$. Then, great simplifications in the result occur. Indeed, we would have*

$$w_{ij} = 1 \quad \forall (i,j), \qquad B = 1, \qquad C = 1,$$

so that

$$S = \frac{\sum_{ij=11}^{rs} \eta_{ij}^2/\pi_{ij}}{(\eta^t\gamma)^2} - 1 = \frac{\gamma^t\Omega(\mathrm{diag}(\pi))^{-1}\Omega\gamma}{(\gamma^t\Omega\gamma)^2} - 1,$$

that can still be simplified to

$$S = \frac{1}{\gamma^t\Omega\gamma} - 1.$$

Remark 4.9. *A natural question is to ask whether the estimator still behaves quite well when the design assumption (SIR1) is not fulfilled. Duan and Li (1991) answer this question by providing a bound for the noncollinearity*[†] *between* θ_0 *and* $\hat{\theta}$*:*

$$\sin^2(\hat{\theta}, \theta_0) \leq \frac{\tau(1-\lambda)}{\lambda(1-\tau)},$$

where τ *is the second eigenvalue of* $\Sigma_X^{-1}\Lambda$*, with* $\Lambda = \mathbb{V}\mathrm{ar}(\mathbb{E}(X|\theta_0^t X))$*, and* λ *the maximum value attained by the ratio (4.30). If Assumption (SIR1) holds, then* Λ *is of rank one and* $\tau = 0$*, so that* $\hat{\theta}$ *and* θ_0 *are collinear (what obviously leads to consistency). If not, but the design is nearly elliptically symmetric, then* $\tau \simeq 0$ *and the estimator remains a good approximation of the true direction. Estimating* τ *permits to control the deviation between the estimated and the true directions, when the elliptical symmetry of the distribution of* X *is not ensured. Anyway, the usual good behavior of the SIR estimator is discussed in the rejoinder to Li (1991), even under mild to severe violation of Assumption (SIR1).*

4.5 Discussion

In this section are discussed some advantages and drawbacks of the four estimation schemes presented in the previous sections, from a theoretical point of view. First of all, Theorems 4.1, 4.2, 4.3 and 4.4 all state the root-n consistency and the asymptotic normality of the analyzed four estimators $\hat{\theta}$. Besides, the particular forms of matrices Σ_{SML}, Σ_{SLS}, Σ_{ADE} and Σ_{SIR}, with first row and first column equal to 0[‡], obviously render the fact the the first component of θ_0 is fixed to one. More interesting is the (almost) efficient character of the M-estimators (SML and SLS with optimal weighting), while the direct estimators (ADE and SIR) fail to reach the semiparametric efficiency bound (4.6).

Concerning the ease of computation of the estimator, it is clear that M-estimators are computed in a much more tricky way than direct ones, as the formers are the solutions of

[†]The sine function is here defined with respect to the inner product $(\theta_1, \theta_2) = \theta_1^t\Sigma_X\theta_2$, by analogy with results of Duan and Li (1991), i.e. $\sin^2(\hat{\theta}, \theta_0) = 1 - \frac{(\hat{\theta}^t\Sigma_X\theta_0)^2}{(\hat{\theta}^t\Sigma_X\hat{\theta})(\theta_0^t\Sigma_X\theta_0)}$.

[‡]Explicit for Σ_{SML} and Σ_{SLS}, directly induced by the structure of $\tilde{\phi}$ for Σ_{ADE} and Σ_{SIR}.

complicated optimization problems (4.2) and (4.9), since the latters are given by analytical expressions like (4.17)-(4.20) and (4.31). Moreover, each iteration in the optimization processes requires the evaluation of Nadaraya-Watson estimators $\{\hat{g}_{ij}^{\theta}\}$ at observations, what leads to a still more computing-intensive procedure. By the way, the required conditions on the kernel and the bandwidth for this estimation are more restricting for the SML than for the SLS estimator. Indeed, in theory, the former needs a higher-order kernel and a bandwidth $h \sim n^{-a}$, with $a \in]1/5, 1/4[$, while the latter allows a second order kernel, and a bandwidth $h \sim n^{-b}$, with $b \in]1/8, 1/3[$. In particular, the usual optimal order of the bandwidth, i.e. $h \sim n^{-1/5}$, is acceptable for SLS, not for SML. Finally, note that the ADE estimator also relies on a (multivariate) kernel estimation of the density of X, with higher-order kernel, but which has to be computed only one time. On the other hand, SIR estimator is not built on any kernel estimator, what makes its computation very fast and simple. For example, no bandwidth has to be selected, contrary to ADE.

Besides their lack of efficiency, direct estimators also suffers for their need of strong assumptions on the design of the covariates. First of all, ADE and SIR basically adapt to continuously distributed vector of regressors only. Further, the SIR methodology requires vector X to have an elliptical symmetric distribution, which is not trivially the case in applications, and that $\mathbb{E}\left(\theta_0^t(X - \mu_X)|Z^{(ij)}\right) \neq 0$ for at least one cell (i, j), which excludes a.o. situations where the distribution of X is symmetric around μ_X in each subpopulation defined by the cells of the contingency table. The ADE procedure requires the same kind of condition, namely $\mathbb{E}\left(f(X)g'_{ij}(\theta_0^t X)\right) \neq 0$ for at least one cell (i, j), which excludes for example spherically symmetric (around 0) distributions of X with even link functions. Also, the condition $f(x) = 0\ \forall x \in \partial S_X$ in Assumption (ADE2) excludes uniform designs, for example. None of such structural assumptions are required for SML and SLS, except the identification Assumptions 1-4.

Finally, having a look at the technical conditions assumed by the four theorems, it appears that SLS requires smoothness of f_θ and g^θ (3 times continuously differentiable, with a third derivative Lipschitz), that SML needs $f_1(x^{(1)}|X^{(-1)} = x^{(-1)})$ positive and 4 times differentiable w.r.t. $x^{(1)}$ (what implies f_θ continuous and uniformly bounded, see comment following Assumption 2 in Lee (1995)), that a strong assumption on the density of X, namely assumption (ADE2), is necessary for ADE, but nothing about the behaviour of the index (and the related functions), and that SIR is free of that type of conditions, except the identification conditions.

As a conclusion, it seems that the M-estimators (SML and SLS) provide the most interesting

procedures, in terms of efficiency and mildness of the required assumptions, but at the expense of solving a possibly intricate optimization problem. When the distribution of X is continuous and can be considered as ellipitical symmetric[†], SIR could be a serious competitor, while ADE does not seem to present many advantages, from a purely theoretical point-of-view, with respect to other methods. The small sample performances of these four estimators are analyzed, through a simulation study, in the next section.

5 A simulation study

In this section a simulation study is performed in order to compare the methods described in the previous sections from a practical point-of-view. Three simulated models were analyzed. For each, $r = s = 2$ and $p = 2$, and the assumed conditional probabilities satisfy Assumption 2 (Sinlge-Index assumption), with $\theta_0 = (1, 2)^t$. Note that, as $\theta_0^{(1)}$ is fixed to 1 for identification, the only unknown to be estimated is the second component $\theta_0^{(2)} = 2$. For each model, three sample sizes were considered, $n = 50$, $n = 200$ and $n = 500$, for which 500 Monte-Carlo replications were drawn. We computed 8 estimators, namely the Semiparametric Maximum Likelihood estimator with a second order kernel (SML2), the Semiparametric Maximum Likelihood with a fourth order kernel (SML4)[‡], the unweighted Semiparametric Least Squares estimator (SLS), the weighted Semiparametric Least Squares estimator (WSLS), the Average Derivatives Estimator with 3 bandwidths set to $Cn^{-1/7}$, with $C = 1$, 2 and 3 (ADE1, ADE2 and ADE3), and the Sliced Inverse Regression estimator. For the M-estimators, the optimization problem was solved via a grid search, with a bandwidth determined by a plug-in method.

For the first scenario, we took

$$X = (X_1, X_2)^t \sim \mathcal{N}\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & 0.5\\0.5 & 1\end{pmatrix}\right)$$

and conditional probabilities as

$$\begin{aligned}
\pi_{1\cdot}(x) = 0.95\exp(-(\theta_0^t x)^2), \quad \pi_{2\cdot}(x) = 1 - \pi_{1\cdot}(x),\\
\pi_{\cdot 1}(x) = 0.95\frac{\exp(-(\theta_0^t x))}{1+\exp(-(\theta_0^t x))}, \quad \pi_{\cdot 2}(x) = 1 - \pi_{\cdot 1}(x)\\
\pi_{ij}(x) = \pi_{i\cdot}(x)\pi_{\cdot j}(x) \qquad \forall (i,j).
\end{aligned}$$

[†]Tests for elliptical symmetry are described in Huffer and Park (2007), Manzotti et al (2002) or Schott (2002). Other more classical references, as Mardia (1970) and Baringhaus and Henze (1988), deal with testing for multivariate normality.

[‡]See Remark 4.2.

The mean and the MSE of each estimators, computed from the Monte-Carlo replications, are shown in table 1.

	$n = 50$		$n = 200$		$n = 500$	
	mean($\hat{\theta}^{(2)}$)	MSE($\hat{\theta}^{(2)}$)	mean($\hat{\theta}^{(2)}$)	MSE($\hat{\theta}^{(2)}$)	mean($\hat{\theta}^{(2)}$)	MSE($\hat{\theta}^{(2)}$)
SML2	2.209	0.080	2.061	0.019	2.007	0.009
SML4	2.449	0.258	2.213	0.070	2.082	0.022
SLS	2.142	0.052	2.060	0.016	2.009	0.015
WSLS	2.108	0.042	2.054	0.015	2.010	0.007
ADE1	0.040	3.871	0.234	3.145	0.483	2.330
ADE2	0.547	2.145	1.302	0.515	1.794	0.065
ADE3	1.065	0.904	1.729	0.095	1.958	0.018
SIR	2.565	0.396	2.121	0.041	2.014	0.013

Table 1: Results for scenario 1.

For scenario 2, we took

$$(X_1 + 1)/2 \sim Beta(2, 2), \quad (X_2 + 1)/2 \sim Beta(2, 2),$$

X_1 independent to X_2, that is a non elliptical symmetric distribution for vector X, as it can be written

$$f(x_1, x_2) = \frac{9}{4}(1 - x_1^2)(1 - x_2^2)$$

on $[-1, 1] \times [-1, 1]$, which fails to be written as (4.28). Nevertheless, the bound given by Remark 4.9 is found to be close to zero. The conditional probabilities are the same as in scenario 1. Table 2 shows the results for the 8 estimation schemes.

	$n = 50$		$n = 200$		$n = 500$	
	mean($\hat{\theta}^{(2)}$)	MSE($\hat{\theta}^{(2)}$)	mean($\hat{\theta}^{(2)}$)	MSE($\hat{\theta}^{(2)}$)	mean($\hat{\theta}^{(2)}$)	MSE($\hat{\theta}^{(2)}$)
SML2	2.017	0.052	2.158	0.048	2.056	0.015
SML4	2.545	0.341	2.148	0.103	2.268	0.093
SLS	2.047	0.059	2.120	0.035	2.020	0.011
WSLS	2.042	0.058	2.110	0.031	2.013	0.010
ADE1	0.307	2.897	0.635	1.894	1.191	0.678
ADE2	0.846	1.364	1.620	0.167	1.852	0.038
ADE3	0.851	1.347	1.614	0.168	1.855	0.035
SIR	1.758	0.126	2.093	0.036	2.000	0.015

Table 2: Results for scenario 2.

The third scenario was the following. The distribution of X was taken to be

$$X = (X_1, X_2)^t \sim \mathcal{N}\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & 0.5\\0.5 & 1\end{pmatrix}\right)$$

as in scenario 1, while the conditional probabilities were

$$\pi_{1\cdot}(x) = 0.5 + (\sin(\theta_0^t x) + \cos(\theta_0^t x))/3, \quad \pi_{2\cdot}(x) = 1 - \pi_{1\cdot}(x),$$
$$\pi_{\cdot 1}(x) = 0.95\frac{\exp(-(\theta_0^t x))}{1+\exp(-(\theta_0^t x))}, \quad \pi_{\cdot 2}(x) = 1 - \pi_{\cdot 1}(x)$$
$$\pi_{ij}(x) = \pi_{i\cdot}(x)\pi_{\cdot j}(x) \qquad \forall (i,j).$$

One could think that the unusual form of $\pi_{1\cdot}(x)$ (periodic function) leads to a more challenging situation with respect to the estimation of θ_0. The results are given in table 3.

	$n = 50$		$n = 200$		$n = 500$	
	mean($\hat{\theta}^{(2)}$)	MSE($\hat{\theta}^{(2)}$)	mean($\hat{\theta}^{(2)}$)	MSE($\hat{\theta}^{(2)}$)	mean($\hat{\theta}^{(2)}$)	MSE($\hat{\theta}^{(2)}$)
SML2	2.191	0.086	2.090	0.030	2.046	0.016
SML4	2.456	0.277	2.409	0.207	2.217	0.069
SLS	2.241	0.110	2.105	0.032	2.019	0.011
WSLS	2.166	0.081	2.096	0.030	2.023	0.011
ADE1	0.091	3.670	0.137	3.501	0.348	2.760
ADE2	0.349	2.755	1.035	0.961	1.552	0.223
ADE3	0.808	1.448	1.599	0.185	1.827	0.046
SIR	2.282	0.146	2.170	0.059	2.024	0.016

Table 3: Results for scenario 3.

It clearly appears from the results of Tables 1, 2 and 3 that the Weighted Semiparametric Least Squares is the best performer in practice, as it attains the minimum Mean Squared Error among the considered 8 estimators, for all scenarii and all sample sizes, except (scenario 2, $n = 50$). With respect to the unweighted SLS, the efficiency gained by the weighting appears, but is quite fair. It is also seen that the Semiparametric Maximum Likelihood estimator is much more stable when using a second order kernel rather than a fourth order kernel, what makes the SML2 estimator another good competitor. Results related to the ADE estimators seems to indicate that the method is very sensitive to the bandwidth choice. Estimator ADE1 was based on a clearly not appropriate bandwidth, while estimator ADE3 seems to be the best among those 3, but still far after the other estimation schemes. Finally, the SIR estimator leads to good results when the sample size is important enough, even when the elliptical symmetry assumption on the distribution of X is not fulfilled. As it is given by an analytic form, so fast and easy computed, this estimator could be the preliminary

estimator of θ_0 needed in the weighting procedure of the WSLS estimator, or could be of use as initial value in the optimization process of the M-estimators.

6 Conclusion

When analyzing a contingency table, built on the cross-classification of a sample of individuals with respect to the levels of two categorical variable R and S, it is often worth considering the conditional joint distribution of (R, S) given an eventual set of explanatory variables, say X. First, this allows to check a possible effect of X and the distribution of (R, S), and second, this allows to take this effect into account when performing the usual analyzes of such tables. In this paper, a semiparametric model for this conditional distribution is proposed, in order to avoid the rash maintained hypotheses of parametric approaches, as well as the well known curse of dimensionality problem of nonparametric procedures. Essentially, it is assumed that the effect of the vector X on (R, S) can be captured by a single index $\theta_0^t X$, where θ_0 is an unknown vector. The link between this index and the related conditional probabilities is also kept free, which grants the model an important flexibility, while the univariate character of the index permits to avoid the curse of dimensionality. Inspired by the usual estimation schemes already proposed for Single-Index Models in classical regression problems, four estimators for θ_0 are proposed, namely a Semiparametric Maximum Likelihood estimator, a Semiparametric Least Squares estimator, an Average Derivatives estimator and a Sliced Inverse Regression estimator. These are all root-n consistent, with asymptotic normal distribution. The former two asymptotically reach the semiparametric efficiency bound (up to a technical trimming), but are defined as the solution of a possibly tricky optimization problem. The latter two, directly given by an analytical expression, are fast and easy to compute, but are not asymptotically efficient and are based on stronger structural assumptions on the covariates. The practical performances of the estimators are also compared through a simulation study. The Semiparametric Least Squares estimator, with a suitable weighting scheme, gives the best results, while the Semiparametric Maximum Likelihood estimator and the Sliced Inverse Regression also lead to good results. On the other hand, the Average Derivatives estimator seems to be a cut below. At a second step, the conditional probabilities are estimated via standard univariate nonparametric regression techniques, without being affected by the estimation of θ_0. Obviously, the study developped in this work would be of greater interest if the Single-Index assumption could be tested from a sample. Future work will be devoted to implement such a test.

References

[1] Ai, C. (1997). A Semiparametric Maximum Likelihood Estimator. Econometrica, 65, 933-963.

[2] Azzalini, A., Bowman, A.W. and Hardle, W. (1989). On the use of nonparametric regression for model checking. Biometrika, 76, 1-11.

[3] Baringhaus, L. and Henze, N. (1988). A consistent test for multivariate normality based on the empirical characteristic function. Metrika, 35, 339-348.

[4] Chu, C.K. and Cheng, K.F. (1995). Nonparametric regression estimates using misclassified binary responses. Biometrika, 82, 315-325.

[5] Copas, J.B. (1983). Plotting p against x. Appl. Statist., 32, 25-31.

[6] Delecroix, M., Härdle, W. and Hristache, M. (2003). Efficient estimation in conditional single-index regression. J. Mult. Anal., 86, 213-226.

[7] Duan, N. and Li, K.-C. (1991). Slicing regression : a link-free regression method. Ann. Stat., 19, 505-530.

[8] Everitt, B.S. (1992). The Analysis of Contingency Tables. Chapman and Hall, London. 2nd Edition.

[9] Geenens, G. and Delecroix, M. (2006). A survey about single-index theory, International Journal of Statistics and Systems, 1, 213-242.

[10] Geenens, G. and Simar, L. (2008). Nonparametric test for conditional independence in two-way contingency tables. Discussion paper no 0801, Institut de Statistique, Université catholique de Louvain. http://www.stat.ucl.ac.be/ISpub/dp/2008/DP0801.pdf

[11] Glonek, G.F.V. and McCulagh, P. (1995). Multivariate Logistic Models. J. Roy. Statist. Soc. B, 57, 533-546.

[12] Glonek, G.F.V. (1996). A class of regression models for multivariate categorical responses. Biometrika, 83, 15-28.

[13] Härdle, W. (1990). Applied Nonparametric Regression. Cambridge University Press.

[14] Härdle, W., Müller, M., Sperlich, S. and Werwatz, A. (2004). Nonparametric and Semiparametric Models. An Introduction. Springer-Verlag, New-York.

[15] Härdle, W. and Stoker, T,M. (1989). Investigating smooth multiple regression by the method of averages derivatives. J. Amer. Stat. Assoc., 84, 986-995.

[16] Horowitz, J.L. and Härdle, W. (1996). Direct Semiparametric Estimation of Single-Index Models with Discrete Covariates. J. Amer. Statist. Assoc., 91, 1632-1640.

[17] Horowitz, J.L. (1998). Semiparametric methods in econometrics. Springer, New-York.

[18] Huffer, F.W. and Park, C. (2007). A test for elliptical symmetry. J. Mult. Anal., 98, 256-281.

[19] Ichimura, H. (1987). Estimation of single index models. Ph.D. thesis, Department of Economics, MIT, Cambridge, MA.

[20] Ichimura, H. (1993). Semiparametric Least Squares (SLS) and weighted SLS estimation of single-index models. J. Econometrics, 58, 71-120.

[21] Klein, R.L. and Spady, R.H. (1993). An efficient semiparametric estimator for binary response models. Econometrica, 61, 387-421.

[22] Lee, L.-F. (1995). Semiparametric maximum likelihood estimation of polychotomous and sequential choice models. J. Econometrics, 65, 381-428.

[23] Li, K.C. (1991). Sliced inverse regression for dimension reduction (with discussion). J. Amer. Statist. Assoc., 86, 316-342.

[24] Manski, C.F. (1985). Semiparametric analysis of discrete response: Asymptotic properties of the maximum score estimator. J. Econometrics, 27, 313-334.

[25] Manzotti, A., Pérez, F.J. and Quiroz, A.J. (2002). A statistic for testing the null hypothesis of elliptical symmetry. J. Mult. Anal., 81, 274-285.

[26] Mardia, K.V. (1970). Measures of multivariate skewness and kurtosis with applications. Biometrika, 57, 519-530.

[27] McCullagh, P. and Nelder, J.A. (1989). Generalized Linear Models. Chapman and Hall, London.

[28] Newey, W.K. (1990). Semiparametric Efficiency Bounds. J. Appl. Econom., 5, 99-135.

[29] Newey, W.K. and Stoker, T.M. (1993). Efficiency of weighted averages derivative estimators and index models. Econometrica, 61, 1199-1223.

[30] Pagan, A. and Ullah, A. (1999). Nonparametric Econometrics. Cambridge University press.

[31] Powell, J.L., Stock, J.H. and Stoker, T.M. (1989). Semiparametric Estimation of Index Coefficients. Econometrica, 51, 1403-1430.

[32] Rodriguez-Campos, M.C. and Cao-Abad, R. (1993). Nonparametric bootstrap confidence intervals for discrete regression functions. J. Econometrics, 58 (1-2), 207-222.

[33] Schott, J.R. (2002). Testing for elliptical symmetry in covariance-matrix-based analyses. Statist. Probab. Lett., 60, 395-404.

[34] Stone, C.J. (1980). Optimal Rates of Convergence for Nonparametric Estimators. Ann. Stat., 8, 1348-1360.

[35] Thompson, T.S. (1993). Some efficiency bounds for semiparametric discrete choice models. J. Econometrics, 58, 257-274.

[36] Wand, M.P. and Jones, M.C. (1995). Kernel Smoothing. Chapman and Hall, London.