

Single Index Models Regression : Some Recent Advances

Michel Delecroix
CREST-ENSAI
Campus de Ker Lann
Bruz, FRANCE
delecroi@ensai.fr

We address the problem of estimating the regression function in a Single Index Model: the observed data $(X_i, Y_i) \in \mathbb{R}^k \times \mathbb{R}$, for $i = 1, \dots, n$ are independent, and we assume that:

$$R(x) = E[Y_i | X_i = x] = g_{\mathbf{b}_0}(\mathbf{b}_0 x),$$

where $\mathbf{b}_0 x$ is the usual product of two vectors from \mathbb{R}^k , and $g_{\mathbf{b}}$ is defined by:

$$g_{\mathbf{b}}(z) = E[Y_i | \mathbf{b}X_i = z]$$

The Single Index Models have been extensively used in the literature in actuarial sciences, in biometrics or in econometrics, but with a *fixed* link function $g_{\mathbf{b}_0}(\cdot)$, in the framework of Generalized Linear Models (GLM, see McCullagh and Nelder, 1989). Here we focus on the problem of estimating *simultaneously* the link, the parameter \mathbf{b} , and finally the regression function R , without any particular assumption on the conditional law of Y_i given X_i .

One of the most attractive approaches for this purpose is based on M-estimation methods (see Härdle, Hall and Ichimura (1993), Sherman (1994), Delecroix, Hristache (1999)): a consistent estimator $\hat{\mathbf{b}}_n$ of \mathbf{b}_0 can be defined by maximizing with respect to \mathbf{b} the empirical mean of some objective function Ψ :

$$\hat{\mathbf{b}}_n = \arg \max_{\mathbf{b}} \frac{1}{n} \sum_{i=1}^n \Psi(Y_i, \hat{g}_{\mathbf{b}, h_n}(\mathbf{b}X_i)),$$

where $\hat{g}_{\mathbf{b}, h_n}^{(-i)}$, defined below, is a Nadaraya-Watson “leave-one-out” estimator of the function $g_{\mathbf{b}}$, and h_n is the series of corresponding bandwidths, which tends to zero at some appropriate rate as $n \rightarrow \infty$, that is:

$$\hat{g}_{\mathbf{b}, h_n}^{(-i)}(\mathbf{b}X_i) = \frac{\sum_{j \neq i} Y_j K_{h_n}(\mathbf{b}X_i - \mathbf{b}X_j)}{\sum_{j \neq i} K_{h_n}(\mathbf{b}X_i - \mathbf{b}X_j)},$$

where $K_{h_n}(x) = h_n^{-1} K(x/h_n)$ and K is a fixed kernel function (typically a symmetric probability function).

Once \mathbf{b}_0 has been consistently estimated, the regression function $R(x) = E(Y|X = x)$ can be estimated, in a second stage, from the nonparametric regression of Y_i on the estimated index $\hat{\mathbf{b}}_n X_i$, using the usual Nadaraya-Watson estimator based on another series of bandwidths h'_n . The above S.I.M. assumption avoids the so-called “curse of dimensionality” which would appear when

estimating directly R in a pure nonparametric way: one uses only nonparametric estimators of the regression of Y_i on real variables.

In practice, the choice of the bandwidths h_n and h'_n , is crucial. To solve this problem, we suggest here to define:

$$(\hat{\mathbf{b}}_n, \hat{h}_n) = \arg \max_{h, \mathbf{b}} \frac{1}{n} \sum_{i=1}^n \Psi[Y_i, \hat{g}_{\mathbf{b}, h}^{(-i)}(\mathbf{b}X_i)]$$

and

$$\hat{R}_n(x) = \hat{g}_{\hat{\mathbf{b}}_n, \hat{h}_n}(\hat{\mathbf{b}}_n x),$$

After proving the asymptotic consistency and \sqrt{n} -normality of $\hat{\mathbf{b}}_n$ and $\hat{R}_n(x)$, we will study the optimality of \hat{h}_n from a theoretical point of view, and then address the problem of the practical implementation of the method.

REFERENCES

Delecroix, M. and M.Hristache (1999). M-estimateurs semi-paramétriques dans les modèles à direction révélatrice unique. *Bull. Belg. Math. Soc.* **6**, 161-185.

Delecroix, M., Hristache, M. and V.Patilea (1999). Optimal smoothing in semiparametric index approximation of regression functions. *Cahiers du CREST* n°. 9952, INSEE, Paris.

Härdle, W., Hall, P. and H. Ichimura (1993). Optimal smoothing in single-index models. *Ann. Statist.* **21**, 157-178.

McCullagh, P. and J.A. Nelder (1989). *Generalized Linear Models*. London: Chapman and Hall.

Sherman, R. P. (1994). U-processes in the analysis of a generalized semiparametric regression estimator. *Econometric Theory*. **10**, 372-395.

RÉSUMÉ

On considère un modèle à direction révélatrice unique: les variables observées $(X_i, Y_i) \in \mathbb{R}^k \times \mathbb{R}$, sont indépendantes et de même loi, et on suppose l'existence d'un élément \mathbf{b}_0 de \mathbb{R}^k tel que $R(x) = E[Y_i | X_i = x] = g_{\mathbf{b}_0}(\mathbf{b}_0 x)$. On propose un estimateur de $R(x)$, basé sur une M-estimation préalable de \mathbf{b}_0 qui évite le problème délicat du choix a priori de paramètres de lissage (les «fenêtres»). On étudie le comportement asymptotique de cet estimateur et sa mise en œuvre pratique