Une nouvelle méthode d'estimation semi paramétrique pour les modèles de régression de type single-index

Weiyu Li[1] & Valentin Patilea[1]

[1] *CREST-Ensai & IRMAR, Campus de Ker Lann, rue Blaise Pascal, BP 37203, 35172 Bruz cedex, France. Emails: Weiyu.LI@ensai.fr, patilea@ensai.fr*

Résumé. Actuellement, les modèles single index (SIM) semi-paramétrique sont largement utilisés par les statisticiens. Dans le cas de régression moyenne, l'hypothèse de SIM indique que l'espérance conditionnelle de la variable réponse sachant le vecteur des variables explicatives est identique à celle sachant une combinaison linéaire des variables explicatives. Récemment, cette idée a été généralisée à la régression quantile. Lorsque l'on s'intéresse à la distribution conditionnelle d'une variable observée sachant les variables explicatives, la variable réponse est conditionnellement indépendante des variables explicatives sachant une combinaison linéaire de ces variables, sous l'hypothèse de SIM.
Pour les modéles de régression et les modèles à distribution conditionnelle, il s'agit d'une idée naturelle de réduction de la dimension en faisant un compromis entre la régression paramétrique et la régression non-paramétrique.
Bien qu'il existe aujourd'hui déjà certaines techniques employées souvent en pratique pour estimer un modéle SIM, la plupart de ces méthodes permet uniquement d'estimer des modèles de régression et peu de méthodes existent pour des modèles à distribution conditionnelle.
Dans ce rapport, on propose ainsi une nouvelle approche semi-paramétrique basée sur une technique de noyau par une minimisation d'une forme quadratique. Cette nouvelle approche est avantageuse pour une approche d'index unique dans le cas de régression moyenne et de modèle à distribution conditionnelle. Nous comparerons la performance de notre nouvelle méthode d'estimation avec les méthodes existantes.
Mots clés. single-index modèle, régression linéaire locale, semi-paramétrique, estimation par noyau
Abstract. Semi-parametric single index models (SIM) are now widely used by the statisticians. For mean regressions, the SIM assumption means that the conditional expectation of the response given the vector of covariates is the same as the conditional expectation of the response given a linear combination of the covariates. Recently, this idea was extended to quantile regression. When modeling the conditional distribution of an observed variable given the covariates, under the SIM assumption the response is conditionally independent of the covariate vector given a suitable linear combination of the covariates. This convenient dimension-reduction approach is a natural compromise between the parametric and fully nonparametric regressions or models for conditional laws.

Several estimation techniques for single-index regression are available and commonly used in applications. Most of them concern the mean or quantile regression, only few methods were proposed for conditional law modeling.
In this paper, we propose a novel kernel-based semiparametric estimation approach based on the minimization of a quadratic form. The new approach is convenient for the single-index approach in mean regression and conditional law modeling. We compare the performances of our new estimation method to existing procedures.

Keywords. single-index model, local linear regression, semi-parametric, kernel estimate

1 Our model

Denote Y being the response variable and X as the d dimensional explanatory variables. Consider T_u, $u \in [0,1]$ a family of transformation. Our model is, for $\forall u$

$$E[T_u(Y)|X] = E[T_u(Y)|X^T\beta],$$

where β is an unknown index vector which belongs to the parameter space $\beta = \{\beta = (\beta_1, \dots \beta_d)^T : \|\beta\| = 1, \beta_1 > 0, \beta \in \mathbb{R}^d\}$.

Examples

(a) Let $T_u(y) = y$, for $\forall u$ and y, then we have $E[Y|X] = E[Y|X^T\beta]$, this is the mean single-index model.

(b) Let $T_u(y) = \mathbf{1}_{\{y \le \Phi^{-1}(u)\}} = \mathbf{1}_{\{\Phi(y) \le u\}}$, where for instance $\Phi(t) = \int_{-\infty}^{t} \frac{1}{\sqrt{2\sigma}} e^{-\frac{s^2}{2}} ds$. We get $Y|X \sim Y|X^T\beta$, that the conditional law of Y given X is identical with the conditional law of Y given $X^T\beta$. This is the single-index distribution model.

2 Estimate method

For our models, we have $E[T_u(Y_i)|X_i] = E[T_u(Y_i)|X_i^T\beta]$, then denote $g(Y_i, X_i, \beta) = T_u(Y_i) - E[T_u(Y_i)|X_i^T\beta]$, we can get $E[g(Y_i, X_i, \beta)|X_i] = 0$. So we try to use the Smooth Minimum Distance Estimation method [Lavergne and Patilea, 2013]

$$\hat{\beta} = \arg\min \sum_{i,j=1}^{n} g(Y_i, X_i, \beta)^T g(Y_j, X_j, \beta) K_{i,j}$$

to estimator the unknown parameter β.
Suppose $E[T_u(Y_i)|X_i^T\beta] = \mu\{m(X_i^T\beta)\}$, where μ is a known function and m is an unknown univariate link function. We use local linear regression [Fan and Gijbels, 1996] to get the

estimators of $m(X_i^T\beta)$ and $m'(X_i^T\beta)$. So the estimator of $g(Y_i, X_i, \beta)$ is $\hat{g}(Y_i, X_i, \beta) = T_u(Y_i) - \hat{E}[T_u(Y_i)|X_i^T\beta] = T_u(Y_i) - \mu\{\hat{m}(X_i^T\beta)\}$ and

$$\hat{\beta} = \arg\min \sum_{i,j=1}^{n} \hat{g}(Y_i, X_i, \beta)^T \hat{g}(Y_j, X_j, \beta) K_{i,j} \tag{1}$$

For getting a more efficient estimator, we transform the equation (1) which is restricted in $\|\beta\| = 1$ to an unrestricted estimation equation [Wang et al., 2010]:

$$\sum_{i,j=1}^{n} [\partial \hat{g}(Y_i, X_i, \beta)/\partial \beta^{(1)}]^T \hat{g}(Y_j, X_j, \beta) K_{i,j} = 0 \tag{2}$$

where $\beta^{(1)} = (\beta_2, \dots \beta_d)^T$ and $\beta_1 = \sqrt{1 - \|\beta^{(1)}\|^2}$. And the equation (2) can be approximated by

$$Q(\beta) = \sum_{i,j=1}^{n} J^T [X_i - \hat{h}(X_i^T\beta)] \hat{g}'(Y_i, X_i, \beta)^T \hat{g}(Y_j, X_j, \beta) K_{i,j} = 0$$

with $J = \partial\beta/\partial\beta^{(1)}$ and $\hat{h}(t)$ is the local linear estimator of $h(t) = E[X|X^T\beta = t] = (h_1(t), \dots, h_d(t))^T$.

3 Algorithm

Let $Q(\beta) = J^T \hat{F}(\beta)$ with $\hat{F(\beta)} = (\hat{F}_1(\beta), \dots, \hat{F}_d(\beta))^T$ and

$$\hat{F}_s(\beta) = \sum_{i,j=1}^{n} [X_{si} - \hat{h}_s(X_i^T\beta)] \hat{g}'(Y_i, X_i, \beta)^T \hat{g}(Y_j, X_j, \beta) K_{i,j}$$

Setting $Q(\beta) = J^T \hat{F}(\beta) = 0$, we have

$$\beta \frac{\hat{F}_1(\beta)}{\|\hat{F}(\beta)\|} = \frac{|\hat{F}_1(\beta)|}{\|\hat{F}(\beta)\|} \times \frac{\hat{F}(\beta)}{\|\hat{F}(\beta)\|^2}$$

Because:

1 The value of $\|\hat{F}(\beta)\|$ sometimes is small.

2 The convergence rate of the algorithm depends on $\|\frac{\partial\{\hat{F}_1(\beta)/\|\hat{F}(\beta)\|\}}{\partial\beta}\| \leq L$.

So a constant M is introduced, adding $M\beta$ on both sides of the equation and dividing by $\hat{F}_1(\beta)/\|\hat{F}(\beta)\| + M$:

$$\beta = \frac{M}{\hat{F}_1(\beta)/\|\hat{F}(\beta)\| + M}\beta + \frac{\hat{F}_1(\beta)/\|\hat{F}(\beta)\|^2}{\hat{F}_1(\beta)/\|\hat{F}(\beta)\| + M}\hat{F}(\beta) \tag{3}$$

EFM	Our method
0.2602	0.2520

Table 1: Average estimation errors $\sum_{s=1}^{d} |\hat{\beta}_s - \beta_s|$

Step 1 choose initial values β_{old} for β.

Step 2 get the $\hat{g}(Y_i, X_i, \beta_{old})$ and $\hat{g}'(Y_i, X_i, \beta_{old})$.

Step 3 from equation (3) we get a new value β_{new} of β, and update β_{old} with $\beta_{old} = \beta_{new}/\|\beta_{new}\|$

Step 4 repeat step 2 and step 3, until β_{new} converge

Empirically $(1, \ldots, 1)^T/\sqrt{(d)}$ can be used to be the initial value. For the local linear regression we use the Epanechnikov kernel and the bandwidth can be chosen by any standard bandwidth selection methods, such as cross-validation. The weight used in function (1) is $K_{i,j} = K_h(X_i - X_j)/\sum_{j=1}^{n} K_h(X_i - X_j)$ where $K_h(\cdot)$ is a kernel function. We choose the bandwidth h by minimize the value of function (1). The constant M can be chosen by K-fold cross-validation method.

4 Property

Under some regularity conditions, the estimator is

1 asymptotically normal estimator

2 asymptotic efficient estimator

5 Simulation

The example we used here is

$$Y = (X^T\beta)^2 + \epsilon$$

The true parameter is $\beta = (2, 1, 0, \ldots, 0)^T/\sqrt{5}$, X is generated from $N_{10}(2, \mathbf{I})$ and $\epsilon = \exp(\sqrt{5}X^T\beta/14)N(0, 1)$. We used the sample of $n = 100$ observations and the replication time is $r = 250$. We compare our method with the EFM [Cui et al., 2011]. The result is shown in table 1.

References

[Cui et al., 2011] Cui, X., Härdle, W. K., and Zhu, L. (2011). The efm approach for single-index models. *The Annals of Statistics*, 39(3):1658–1688.

[Fan and Gijbels, 1996] Fan, J. and Gijbels, I. (1996). *Local polynomial modelling and its applications*. Number 66 in Monographs on statistics and applied probability series. Chapman and Hall, London [u.a.].

[Lavergne and Patilea, 2013] Lavergne, P. and Patilea, V. (2013). Smooth minimum distance estimation and testing with conditional estimating equations: Uniform in bandwidth theory. *Journal of Econometrics*, 177(1):47 – 59.

[Wang et al., 2010] Wang, J.-L., Xue, L., Zhu, L., and Chong, Y. S. (2010). Estimation for a partial-linear single-index model. *The Annals of Statistics*, 38(1):246 – 274.