Tests non paramétriques d'adéquation pour les modèles de régression de type single-index

Samuel Maistre[1] & Valentin Patilea[1]

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

Résumé. Les modèles semi-paramétriques à direction révélatrice unique (single-index model ou SIM) sont de plus en plus utilisés. En régression, l'hypothèse SIM signifie que l'espérance conditionnelle de la variable réponse sachant le vecteur de covariables est la même que celle sachant uniquement une projection linéaire de ce vecteur. On peut aussi faire cette hypothèse sur la loi conditionnelle de la réponse sachant les covariables. Cette manière de réduire la dimension est un compromis convenable entre les approches paramétrique et purement non paramétrique dans les deux cas sus-mentionnés.

Plusieurs techniques sont disponibles pour estimer le modèle de régression SIM. Néanmoins, le problème du test d'adéquation à ce modèle a été moins étudié, et les propositions existantes ont encore d'importants défauts. Dans cette présentation, nous introduisons une nouvelle manière d'effectuer ce test. Le vecteur de covariables n'a pas besoin d'avoir une densité et seule la projection estimée est utilisée dans le lissage à noyau. On évite ainsi l'effet d'un lissage en grande dimension tout en obtenant la normalité asymptotique de la statistique de test. Ce nouveau test détecte des alternatives locales approchant l'hypothèse nulle à une vitesse moins grande que $n^{-1/2}h^{-1/4}$, et ce indépendamment de la dimension du vecteur de covariable. Nous proposons également une procédure de bootstrap pour obtenir les valeurs critiques et nous comparons notre procédure à l'existant.

Mots-clés. Régression de type single-index, test d'adéquation, lissage à noyau, $U-$statistiques

Abstract. Semi-parametric single index models (SIM) are widely used tools for statistical modeling. In a regression setup, 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 scalar projection of the covariate vector. In a conditional distribution modeling, under the SIM assumption the conditional law of a response given the covariate vector coincides with the conditional law given a linear combination of the covariates. This convenient dimension-reduction approach is a 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. However, the problem of testing the goodness-of-fit seems less explored and the existing proposals still have some major drawbacks. In this paper, a novel kernel-based approach for testing the SIM assumption is introduced. The covariate vector needs not

have a density and only the index estimated under the SIM assumption is used in kernel smoothing. Hence the effect of high-dimensional smoothing is mitigated while asymptotic normality of the test statistic is obtained. Irrespective of the fixed dimension of the covariate vector, the new test detects local alternatives approaching the null hypothesis slower than $n^{-1/2}h^{-1/4}$, where h is the bandwidth used to build the test statistic and n is the sample size. We show the validity of wild bootstrap critical values and we compare the small sample performances of our test to existing procedures.

Keywords. Single-index regression, lack-of-fit test, kernel smoothing, $U-$statistics

1 Introduction

Les modèles semi-paramétriques à direction révélatrice unique (*single-index model* ou *SIM* en anglais) sont de plus en plus utilisés en modélisation statistique. En régression, l'hypothèse SIM signifie que l'espérance conditionnelle de la variable expliquée Y sachant le vecteur de covariables $X \in \mathbb{R}^p$ est la même que celle sachant uniquement une projection linéaire de ce vecteur. Autrement dit, il existe β_0 de norme unité tel que

$$\mathbb{E}[Y \mid X] = \mathbb{E}[Y \mid X'\beta_0]. \tag{1}$$

Le vecteur β_0 est appelé l'indice ou la direction révélatrice. Dans un tel modèle, la direction donnée par β_0 et la fonction de régression univariée $\mathbb{E}[Y \mid X'\beta_0 = \cdot]$ sont à estimer. Voir par exemple Hristache *et al.* (2001) et Delecroix *et al.* (2006).

Une autre hypothèse de ce type consiste à supposer que c'est la loi conditionnelle de la réponse Y sachant les covariables X qui ne dépend que d'une projection $X'\beta_0$, c'est-à-dire

$$Y \perp X \mid X'\beta_0. \tag{2}$$

Dans ce cas, la direction β_0 et la loi conditionnelle de Y sachant $X'\beta_0$ sont à estimer. Voir par exemple Delecroix *et al.* (2003) et Hall et Yao (2005).

Cette manière de réduire la dimension est un compromis naturel entre les approches purement paramétrique et purement non paramétrique dans les deux situations mentionnées. Toutefois, le statisticien devrait disposer d'un outil pour décider à partir des données si une direction révélatrice suffit ou si le vecteur de covariable continent d'information pertinente en dehors de la combinaison linéaire $X'\beta_0$. Un certain nombre de tests de l'hypothèse (1) ont été proposés dans la littérature, mais les conditions techniques les acompagnant sont souvent trop restrictives. Voir Fan et Li (1996), Xia *et al.* (2004), Stute et Zhu (2005), Chen et Van Keilegom (2009). A notre conaissance, il n'existe aucune procédure de test de l'hypothèse (2).

Dans ce travail nous proposons une nouvelle approche de test qui permet de tester (1) ou (2). Elle est inspirée d'une approche générale de test de significativité en régression non paramétrique qu'on présente dans la section suivante.

2 Une approche générale pour tester la significativité de variables

Soit $(\mathcal{H}, \langle\cdot,\cdot\rangle_{\mathcal{H}})$ un espace de Hilbert. Dans notre contexte, $\mathcal{H} = \mathbb{R}$ ou $\mathcal{H} = L^2[0,1]$. Considérons $U \in \mathcal{H}$, $Z \in \mathbb{R}^q$ et $W \in \mathbb{R}^p$ et un échantillon i.i.d. (U_i, Z_i, W_i), $1 \leq i \leq n$. Considérons le problème de test de l'égalité

$$\mathbb{E}[U \mid Z, W] = 0 \qquad \text{p.s.} \tag{3}$$

Plusieurs problèmes statistiques mènent à vérifier ce type d'égalité, en particulier les tests d'adéquation des SIM auxquels nous nous intéressons dans ce travail.

Pour une fonction l, soit $\mathcal{F}[l]$ la transformée Fourier de l. Soit K un noyau multivarié défini sur $\mathbb{R}^q$ tel que $\mathcal{F}[K] > 0$ et $\phi(s) = \exp(-s^2/2)$, $\forall s \in \mathbb{R}$. L'ensemble des noyaux K avec transformée Fourier positive est assez riche (gaussien, triangle, Student, logistique,...).

Notre procédure se base sur la remarque suivante, voir aussi Lavergne, Maistre & Patilea (2014). D'une part, si $\omega(\cdot)$ est une fonction de poids positive, alors pour tout $h > 0$

$$\begin{aligned} I(h) &= \mathbb{E}\left[\langle U_1, U_2\rangle_{\mathcal{H}}\, \omega(Z_1)\omega(Z_2) h^{-q} K((Z_1 - Z_2)/h)\phi(\|W_1 - W_2\|)\right] \\ &= \mathbb{E}\left[\langle U_1, U_2\rangle_{\mathcal{H}}\, \omega(Z_1)\omega(Z_2) \int_{\mathbb{R}^q} e^{2\pi i t'(Z_1 - Z_2)} \mathcal{F}[K](th)dt \int_{\mathbb{R}^p} e^{2\pi i s'(W_1 - W_2)} \mathcal{F}[\phi](s)ds\right] \\ &= \int_{\mathbb{R}^q}\int_{\mathbb{R}^p} \left\| \mathbb{E}\left[\mathbb{E}[U \mid Z, W]\omega(Z) e^{-2\pi i\{t'Z + s'W\}}\right]\right\|_{\mathcal{H}}^2 \mathcal{F}[K](th)\mathcal{F}[\phi](s)dtds. \end{aligned}$$

Etant donné que $\mathcal{F}[\phi], \mathcal{F}[K] > 0$, si $\omega(\cdot) > 0$, pour tout p nous avons l'équivalence

$$\mathbb{E}[U \mid Z, W] = 0 \;\; p.s. \;\; \Leftrightarrow \;\; I(h) = 0, \quad \forall h > 0.$$

Ensuite l'idée de la nouvelle approche est de construire une statistique de test basée sur une approximation de $I(h)$. Pour cela la fonction ω sera choisie convenablement afin d'éviter des dénominateurs.

Comme estimation de $I(h)$, on peut utiliser la $U-$statistique

$$I_n(h) = \frac{1}{n(n-1)h^q} \sum_{1 \leq i \neq j \leq n} \left\langle \widehat{U_i\omega(Z_i)},\, \widehat{U_j\omega(Z_j)} \right\rangle_{\mathcal{H}} K_{ij}(h)\, \phi_{ij},$$

où

$$K_{ij}(h) = K((Z_i - Z_j)/h), \qquad \phi_{ij} = \phi(\|W_i - W_j\|) = \exp(-\|W_i - W_j\|^2/2).$$

La variance de $I_n(h)$ peut s'estimer par

$$v_n^2(h) = \frac{2}{n^2(n-1)^2h^{2q}} \sum_{1 \leq i \neq j \leq n} \left\langle \widehat{U_i\omega(Z_i)},\, \widehat{U_j\omega(Z_j)} \right\rangle_{\mathcal{H}}^2 K_{ij}^2(h)\, \phi_{ij}^2.$$

La statistique de test pour éprouver (3) est donc

$$T_n = \frac{I_n(h)}{v_n(h)}.$$

Sous de conditions techniques raisonnables, on peut montrer que la statistique de test a une loi normale standard si l'hypothèse nulle $\mathbb{E}[U \mid Z, W] = 0$ p.s. est vérifiée. Voir aussi Lavergne, Maistre & Patilea (2014). On peut montrer également que le test est consistant contre toute alternative fixe. De plus, il détecte les alternatives à la Pitman

$$H_{1n} : \ \mathbb{E}(U \mid Z, W) = r_n \delta(Z, W), \quad n \geq 1,$$

avec une probabilité tendant vers 1 dès que $r_n^2 n h^{q/2} \to \infty$.

3 Tests pour les modèles de type single-index

Considérons maintenant le problème de test de l'hypothèse (1). Dans ce cas $q = 1$, $Z = Z(\beta_0)$ et $W = W(\beta_0)$ où, pour β dans un voisinage de β_0,

$$Z(\beta) = X'\beta \qquad \text{et} \qquad W(\beta) = X'\mathbf{A}(\beta)$$

où $\mathbf{A}(\beta) \in \mathbb{R}^{p\times(p-1)}$ est une matrice telle que

$$\begin{pmatrix} \beta & \mathbf{A}(\beta) \end{pmatrix} \in \mathbb{R}^{p\times p}$$

est une matrice orthogonale. De plus, $\mathcal{H} = \mathbb{R}$, $U\omega(Z) = U(\beta_0)\omega(Z;\beta_0)$ où

$$U(\beta)\omega(Z;\beta) = \{Y - \mathbb{E}[Y \mid X'\beta]\} f_\beta(X'\beta),$$

avec f_β la densité de la variable $X'\beta$ supposée continue. Dans le cas du problème de test de l'hypothèse (2) on considère $\mathcal{H} = L^2[0,1]$ et

$$U(t;\beta)\omega(Z;\beta) = \left\{\mathbf{1}\{Y \leq \Phi^{-1}(t)\} - \mathbb{P}[Y \leq \Phi^{-1}(t) \mid X'\beta]\right\} f_\beta(X'\beta), \qquad t \in [0,1],$$

où Φ est la fonction de repartition de la loi normale standard.

Par la suite nous allons nous concentrer sur le test de de l'hypothèse (1). Soit $\hat{\beta}$ un estimateur de β_0 convergent à la vitesse $O_p(n^{-1/2})$ si l'hypothèse (1) est vérifiée. On définit

$$\widehat{U_i\omega(Z_i)} = \frac{1}{n(n-1)} \sum_{k\neq i} (Y_i - Y_k) \frac{1}{g^q} L_{n,ik}(\hat{\beta}),$$

où L est un noyau, $L_{n,ik}(\hat{\beta}) = L((Z_i(\hat{\beta}) - Z_k(\hat{\beta}))/g)$ et g une fenêtre qui converge vers zéro à une certaine vitesse.

On calcule la quantité

$$I_n(\hat{\beta}) \;=\; \frac{1}{n\,(n-1)\,h}\sum_{1\le i\neq j\le n}\widehat{U_i\omega(Z_i)}\widehat{U_j\omega(Z_j)}K_{n,ij}(\hat{\beta})\phi(\|W_i(\hat{\beta})-W_j(\hat{\beta})\|)$$

où K est un noyau, $K_{n,ij}(\hat{\beta}) = K((Z_i(\hat{\beta}) - Z_j(\hat{\beta}))/g)$ et h une autre fenêtre qui converge vers zéro. On montre que $I_n(\hat{\beta}) - I_n(\beta_0)$ est négligeable comparé à $I_n(\beta_0)$. Ensuite, on estime la variance de $I_n(\hat{\beta})$ par

$$\hat{\omega}_n^2(\hat{\beta}) = \frac{2}{n^2\,(n-1)^2\,h^2}\sum_{1\le i\neq j\le n}\widehat{U_i\omega(Z_i)}^2\widehat{U_j\omega(Z_j)}^2K_{n,ij}^2(\hat{\beta})\phi^2(\|W_i(\hat{\beta})-W_j(\hat{\beta})\|).$$

La statistique de test est alors

$$T_n(\hat{\beta}) = \frac{I_n(\hat{\beta})}{\hat{\omega}_n(\hat{\beta})}.$$

Cette statistique conserve la normalité asymptotique sous H_0 et la détection d'alternatives à la Pitman à vitesse $n^{-1/2}h^{-1/4}$ lorsqu'elle est définie avec $\hat{\beta}$, et a fortiori avec β_0, sous des hypothèses standards sur g et h.

Bibliographie

[1] Chen, S.X. & Van Keilegom, I. (2009). A goodness-of-fit test for parametric and semiparametric models in multiresponse regression. *Bernoulli* **15**, 955–976.

[2] Delecroix, M., Härdle, W. & Hristache, M. (2003). Efficient estimation in conditional single-index regression. *J. Multivariate Anal.* **86**, 213–226.

[3] Delecroix, M., Hristache, M., & Patilea, V. (2006). On semiparametric $M-$estimation in single-index regression. *J. Statist Plan. Inference* **136**, 730–769.

[4] Fan, Y. & Li, Q (1996). Consistent model specification tests: Omitted variables and semiparametric functional forms. *Econometrica* **64**, 865–890.

[5] Hall, P., & Yao, Q. (2005). Approximating conditional distribution functions using dimension reduction. *Ann. Statist.* **33**, 977–1454.

[6] Hristache, M., Juditsky, A., & Spokoiny, V. (2001). Direct estimation of the index coefficient in a single-index model. *Ann. Statist.* **29**, 595–917.

[7] Lavergne, P. & Vuong, Q. (2000). Nonparametric significance testing. *Econometric Theory* **16**, 576–601.

[8] Lavergne, P., Maistre, S. & Patilea, V. (2014). A significance test for covariates in nonparametric regression. Working Paper CREST.

[9] Stute, W. & Zhu, L.-X. (2005). Nonparametric checks for single-index models. *Ann. Statist.* **33**, 1048–1083.

[10] Xia, Y., Li, W.K., Tong, H. & Zhang, D. (2004). A goodness-of-fit test for single-index models. *Statist. Sinica* **14**, 1–39.