C. R. Acad. Sci. Paris, Ser. I 349 (2011) 1281–1285

Contents lists available at SciVerse ScienceDirect

C. R. Acad. Sci. Paris, Ser. I

www.sciencedirect.com

Probability Theory/Statistics

Nonparametric estimation of the density of regression errors

Estimation nonparamétrique de la densité des erreurs de régression

Rawane Samb

ISBA, Université Catholique de Louvain, 20, voie du Roman pays, B-1348 Louvain-la-Neuve, Belgium

ARTICLE INFO

Article history:
Received 19 August 2011
Accepted after revision 19 October 2011
Available online 9 November 2011

Presented by Paul Deheuvels

ABSTRACT

Consider the nonparametric regression model $Y = m(X) + \varepsilon$, where the function m is smooth but unknown, and ε is independent of X. An estimator of the density of the error term ε is proposed and its weak consistency is obtained. The strategy used here is based on the kernel estimation of the residuals. Our contribution is twofold. First, we evaluate the impact of the estimation of the regression function m on the error density estimator. Secondly, the optimal choices of the first and second-step bandwidths used for estimating the regression function and the error density respectively, are proposed. Further, we investigate the asymptotic normality of the error density estimator and its rate-optimality.

RÉSUMÉ

Nous présentons un estimateur nonparamétrique de la densité de l'erreur dans le modèle de régression $Y = m(X) + \varepsilon$, où la fonction m est lisse mais inconnue, et le terme d'erreur ε est indépendant de X. L'estimateur proposé est basé sur une estimation nonparamétrique des résidus, et sa consistance faible est obtenue. Notre contribution se situe à deux niveaux. D'abord, nous évaluons l'impact de l'estimation de la fonction de régression sur l'estimateur final de la densité de l'erreur. Ensuite, nous proposons les choix optimaux des fenêtres de première et de deuxième étape utilisées respectivement pour les estimations de m et de la densité des résidus. Nous étudions également la normalié asymptotique de l'estimateur de la densité de l'erreur et sa vitesse de convergence.

Version française abrégée

Soit $(X_1, Y_1), \ldots, (X_n, Y_n)$ un échantillon de variables aléatoires indépendentes et identiquement distribuées (i.i.d.), de même loi que (X, Y). On suppose que Y est une variable univariée à valeurs dans $\mathbb{R}$, et que X désigne une variable explicative multivariée prenant ses valeurs dans $\mathbb{R}^d$, $d \geqslant 1$. Soit $m(x)$ l'espérance conditionnelle de Y sachant que $X = x$, de telle sorte que le modèle de régression relatif à X et Y s'écrit

$$Y_i = m(X_i) + \varepsilon_i, \quad i = 1, \ldots, n, \tag{1}$$

où les erreurs $\{\varepsilon_i\}$ sont supposées être des variables aléatoires i.i.d., indépendantes des $\{X_i\}$, de même loi que ε satisfaisant en particulier $\mathbb{E}[\varepsilon] = 0$. Dans cette note nous proposons un estimateur nonparamétrique de la densité f de ε. L'approche

E-mail address: rawane.samb@uclouvain.be.

1631-073X/$ – see front matter
doi:10.1016/j.crma.2011.10.017

proposée ici est basée sur une procédure à deux étapes qui, dans un premier temps, consiste à estimer nonparamétriquement chaque résidu ε_i par $\widehat{\varepsilon}_i = Y_i - \widehat{m}_{in}(X_i)$, où

$$\widehat{m}_{in}(X_i) = \frac{\sum_{j=1,\, j\neq i}^{n} Y_j K_0(\frac{X_j - X_i}{b_0})}{\sum_{j=1,\, j\neq i}^{n} K_0(\frac{X_j - X_i}{b_0})} \tag{2}$$

est l'estimateur «leave-one-out» de $m(X_i)$. Ici K_0 est une fonction noyau définie dans $\mathbb{R}^d$, et $b_0 = b_0(n)$ est une fenêtre dépendant de n. Dans un second temps, les résidus estimés $\widehat{\varepsilon}_i$ sont utilisés pour construire l'estimateur $\widehat{f}_n$ de f

$$\widehat{f}_n(e) = \frac{1}{b_1 \sum_{i=1}^{n} \mathbb{1}(X_i \in \mathcal{X}_0)} \sum_{i=1}^{n} \mathbb{1}(X_i \in \mathcal{X}_0) K_1\left(\frac{\widehat{\varepsilon}_i - e}{b_1}\right), \quad e \in \mathbb{R}, \tag{3}$$

où K_1 est une fonction noyau, $b_1 = b_1(n)$ est une fenêtre qui dépend de n, et $\mathcal{X}_0$ est un ensemble fixe d'intérieur non vide, dont la fermeture est à l'interieur du support $\mathcal{X}$ de X. Ce choix de $\mathcal{X}_0$ est motivé par le soin d'exclure les points X_i proches des bords de leur support, où les résidus peuvent avoir un biais suffisamment grand. Une des contributions majeures de cette note est de caractériser la façon optimale de choisir la fenêtre de première b_0. Sous les hypothèses (A_1)–(A_9) décrites dans la Section 3, et les notations des Sections 2 et 4 on obtient d'abord le résultat suivant :

$$\widehat{f}_n(e) - f(e) = O_{\mathbb{P}}\big(AMSE(b_1) + R_n(b_0, b_1)\big)^{1/2}.$$

De ce résultat, nous déduisons le choix optimal de la fenêtre de première étape b_0 utilisée pour l'estimation des résidus. Cette fenêtre optimale est définie par

$$b_0^* = b_0^*(b_1) = \arg\min_{b_0} R_n(b_0, b_1),$$

où la minimisation se fait sur l'ensemble des fenêtres b_0 satisfaisant (A_8). Sous (A_1)–(A_9), b_0^* vérifie

$$b_0^* \asymp \max\left\{\left(\frac{1}{n^2 b_1^3}\right)^{\frac{1}{d+4}}, \left(\frac{1}{n^3 b_1^7}\right)^{\frac{1}{2d+4}}\right\}$$

et on a

$$R_n\big(b_0^*, b_1\big) \asymp \max\left\{\left(\frac{1}{n^2 b_1^3}\right)^{\frac{4}{d+4}}, \left(\frac{1}{n^3 b_1^7}\right)^{\frac{4}{2d+4}}\right\}.$$

En utilisant ce résultat, on montre que pour $d \in \{1, 2\}$, la fenêtre b_1^* qui minimise $AMSE(b_1) + R_n(b_0^*, b_1)$ est d'ordre $n^{-1/5}$, conduisant à la vitesse de convergence $n^{-2/5}$ pour $\widehat{f}_n(e) - f(e)$.

1. Introduction

Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be a sample of independent replicates of the random vector (X, Y), where Y is the univariate dependent variable and X is the covariate of dimension d. Let $m(x)$ be the conditional expectation of Y given $X = x$, so that the related regression model is (1), where the errors $\{\varepsilon_i\}$ are assumed to have mean zero and to be independent of the $\{X_i\}$, and the function m is smooth but unknown. In this note, we investigate the nonparametric estimation of the probability density function (p.d.f.) f of the error terms $\{\varepsilon_i\}$. The difficulty of this study is the fact that the errors are not observed, since the function m is unknown and must be estimated. The approach used here for estimating $f(e)$ is based on a two-steps procedure, which, in the first step, replaces the unobserved residuals terms ε_i by some nonparametric estimators $\widehat{\varepsilon}_i = Y_i - \widehat{m}_{in}(X_i)$, where each $\widehat{m}_{in}(X_i)$ is a nonparametric estimator of $m(X_i)$. In a second step, the estimated residuals $\widehat{\varepsilon}_i$ are used to estimate $f(e)$, as if they were the true errors ε_i. Though proceeding so may remedy the curse of dimensionality for large sample sizes, a challenging issue is to evaluate the impact of the estimated residuals on the estimation of $f(e)$, and to find the order of the optimal first-step bandwidth b_0 used for estimating the error terms.

A great deal of effort has been devoted to the estimation of the errors density in regression models. Cheng in [2] establishes the asymptotic normality of an estimator of the error density based on the estimated residuals. This estimator is constructed by splitting the sample into two parts: the first part is used for the estimation of the residuals, while the second part of the sample is devoted to the construction of the density estimator. Efromovich [4] proposes an adaptive estimator of the error density, based on a density estimator proposed by Pinsker [8]. Plancade [9] presents an estimator of the density of the error in homoscedastic regression model, based on model selection methods, and proposes a bound for the quadratic integrated risk. Other approaches have also been proposed. Akritas and Van Keilegom [1] estimate the cumulative distribution function of the regression error in heteroscedastic model with univariate covariates. Recently, Müller, Schick and Wefelmeyer [5] and Neumeyer and Van Keilegom [7] investigated the estimation of the error distribution function in the nonparametric regression model with multivariate covariates. Both articles derived the uniform expansions for (suitably

chosen) residual-based empirical distribution functions, and in particular, obtained the parametric convergence rate $n^{-1/2}$, independently of the covariates dimension.

Although these authors used the estimated residuals for constructing an estimator of the error distribution, none of them investigated the impact of the covariates dimension on the estimation of $f(e)$, nor the influence of the first-step bandwidth used to estimate the regression function on the final estimator of the error distribution.

The contribution of this note is twofold. First, we evaluate the impact of the estimation of the regression function on the error density estimator. Secondly, the optimal choices of the first-step and second-step bandwidths used for estimating the residual terms and the error density respectively, are proposed. Further, we investigate the asymptotic normality of the error density estimator and its rate-optimality.

2. Construction of the estimators and notations

The approach used here for the nonparametric kernel estimation of $f(e)$ is based on a two-steps procedure, which builds, in a first step, the estimated residuals $\widehat{\varepsilon}_i = Y_i - \widehat{m}_{in}(X_i)$, where each $\widehat{m}_{in}(X_i)$ is the leave-one out version of the Nadaraya [6] and Watson [13] kernel estimator of $m(X_i)$ given by (2), in which K_0 is a kernel function defined on $\mathbb{R}^d$, and $b_0 = b_0(n)$ is a bandwidth sequence. It is tempting to use, in the second step, the estimates $\widehat{\varepsilon}_i$ as if they were the true residuals ε_i. This would ignore the fact that the $\widehat{m}_{in}(X_i)$'s can result in severely biased estimates of the $m(X_i)$'s for those X_i which are close to the boundaries of the support $\mathcal{X}$ of the covariate distribution. That is why we built an estimator of f which trims the observations X_i outside a set $\mathcal{X}_0$ with a nonempty interior, whose closure is in the interior of $\mathcal{X}$. Our proposed estimator for $f(e)$ is given by (3), where K_1 is a univariate kernel function and $b_1 = b_1(n)$ is a bandwidth sequence. This estimator is called the residual-based estimator of $f(e)$. In principle, it would be possible to assume that most of the X_i's fall in $\mathcal{X}_0$ when this latter is very close to $\mathcal{X}$. This would give an estimator close to the more natural kernel estimator $\sum_{i=1}^n K_1((\widehat{\varepsilon}_i - e)/b_1)/(nb_1)$. However, in the rest of the note, a fixed subset $\mathcal{X}_0$ will be considered for the sake of simplicity. Observe that the two-steps kernel estimator $\widehat{f}_n(e)$ is a feasible estimator in the sense that it does not depend on any unknown quantity, as desirable in practice. This contrasts with the unfeasible ideal kernel estimator

$$\widetilde{f}_n(e) = \frac{1}{b_1 \sum_{i=1}^n \mathbb{1}(X_i \in \mathcal{X}_0)} \sum_{i=1}^n \mathbb{1}(X_i \in \mathcal{X}_0) K_1\left(\frac{\varepsilon_i - e}{b_1}\right), \tag{4}$$

which depends in particular on the unknown regression error terms. It is however intuitively clear that a proportion of the estimated residuals (those with X_i not close to the boundary of $\mathcal{X}$) yield a density estimator rivaling the one based on the corresponding proportion of the true errors.

3. Assumptions

The assumptions we need for our results are listed below for convenient reference.

- (A_1) *The subset $\mathcal{X}_0$ of $\mathcal{X}$ has a nonempty interior and its closure is in the interior of $\mathcal{X}$.*
- (A_2) *The p.d.f. g of the i.i.d. covariates X_i is strictly positive over the closure of $\mathcal{X}_0$, and has continuous second order partial derivatives over $\mathcal{X}$.*
- (A_3) *The regression function m has continuous second order partial derivatives over $\mathcal{X}$.*
- (A_4) *The i.i.d. centered error regression terms ε_i have finite 6th moments and are independent of the covariates X_i.*
- (A_5) *The probability density function f of the ε_i's has bounded continuous second order derivatives over $\mathbb{R}$ and satisfies $\sup_{e\in\mathbb{R}} |h_p^{(k)}(e)| < \infty$, where $h_p(e) = e^p f(e)$, $p \in [0, 2]$ and $k \in \{0, 1, 2\}$.*
- (A_6) *The kernel K_0 is symmetric, continuous over $\mathbb{R}^d$ with support contained in $[-1/2, 1/2]^d$ and satisfies $\int K_0(z)\,\mathrm{d}z = 1$.*
- (A_7) *The kernel K_1 is symmetric, has a compact support, is three times continuously differentiable over $\mathbb{R}$, and satisfies $\int K_1(v)\,\mathrm{d}v = 1$, $\int K_1^{(\ell)}(v)\,\mathrm{d}v = 0$ for $\ell \in \{1, 2, 3\}$, and $\int vK_1^{(\ell)}(v)\,\mathrm{d}v = 0$ for $\ell \in \{2, 3\}$.*
- (A_8) *The bandwidth b_0 decreases to 0 when $n \to \infty$ and satisfies, for $d^* = \sup\{d + 2, 2d\}$, $nb_0^{d^*}/\ln n \to \infty$ and $\ln(1/b_0)/\ln(\ln n) \to \infty$ when $n \to \infty$.*
- (A_9) *The bandwidth b_1 decreases to 0 and satisfies $n^{(d+8)}b_1^{7(d+4)} \to \infty$ when $n \to \infty$.*

4. Main results

Our first main result evaluates the order of the difference between $\widehat{f}_n(e)$ and $f(e)$, for all $e \in \mathbb{R}$.

Theorem 1. *Under (A_1)–(A_9), we have, for all $e \in \mathbb{R}$, and for b_0 and b_1 going to 0,*

$$\widehat{f}_n(e) - f(e) = O_{\mathbb{P}}\big(AMSE(b_1) + R_n(b_0, b_1)\big)^{1/2},$$

where

$$AMSE(b_1) = \mathbb{E}_n\big[\big(\widetilde{f}_n(e) - f(e)\big)^2\big] = O_{\mathbb{P}}\left(b_1^4 + \frac{1}{nb_1}\right),$$

and

$$R_n(b_0,b_1)=b_0^4+\left[\frac{1}{(nb_1^5)^{1/2}}+\left(\frac{b_0^d}{b_1^3}\right)^{1/2}\right]^2\left(b_0^4+\frac{1}{nb_0^d}\right)^2+\left[\frac{1}{b_1}+\left(\frac{b_0^d}{b_1^7}\right)^{1/2}\right]^2\left(b_0^4+\frac{1}{nb_0^d}\right)^3.$$

The result of Theorem 1 is based on the evaluation of the difference between $\widehat{f}_n(e)$ and $\widetilde{f}_n(e)$. This evaluation gives an indication about the impact of the kernel estimation of the residuals on the nonparametric estimation of the error density. In fact, the remainder term $R_n(b_0,b_1)$ comes from the replacement of the unknown quantities $m(X_i)$ in ε_i by their estimates $\widehat{m}_{in}(X_i)$.

As stated in the next result, Theorem 2 gives some guidelines for the optimal choice of the bandwidth b_0 used in the kernel estimation of the regression errors. In what follows, $a_n \asymp b_n$ means that $a_n = O(b_n)$ and $b_n = O(a_n)$, i.e. that there is a constant $C>0$ such that $|a_n|/C \leqslant |b_n| \leqslant C|a_n|$ for n large enough.

Theorem 2. *Assume* (A_1)–(A_9) *and define*

$$b_0^* = b_0^*(b_1) = \arg\min_{b_0} R_n(b_0,b_1),$$

where the minimization is performed over bandwidth b_0 *fulfilling* (A_8). *Then,*

$$b_0^* \asymp \max\left\{\left(\frac{1}{n^2b_1^3}\right)^{\frac{1}{d+4}},\left(\frac{1}{n^3b_1^7}\right)^{\frac{1}{2d+4}}\right\},$$

and

$$R_n\big(b_0^*,b_1\big) \asymp \max\left\{\left(\frac{1}{n^2b_1^3}\right)^{\frac{4}{d+4}},\left(\frac{1}{n^3b_1^7}\right)^{\frac{4}{2d+4}}\right\}.$$

Our next theorem gives the conditions for which the estimator $\widehat{f}_n(e)$ reaches the optimal rate $n^{-2/5}$ when $b_0 = b_0^*$. We show that for $d \leqslant 2$, the bandwidth that minimizes the term $AMSE(b_1) + R_n(b_0^*, b_1)$ has the same order as $n^{-1/5}$, yielding the optimal order $n^{-2/5}$ for $(AMSE(b_1) + R_n(b_0^*, b_1))^{1/2}$. Note that the order $n^{-2/5}$ is the optimal rate achieved by the kernel estimator of a univariate density. See, for instance, Deheuvels and Mason [3], Scott [10] or Wand and Jones [12].

Theorem 3. *Assume* (A_1)–(A_9) *and let*

$$b_1^* = \arg\min_{b_1}\big(AMSE(b_1) + R_n\big(b_0^*,b_1\big)\big),$$

where $b_0^* = b_0^*(b_1)$ *is defined as in Theorem* 2. *Then,*

(i) *For* $d \in \{1,2\}$, *we have*

$$b_1^* \asymp \left(\frac{1}{n}\right)^{\frac{1}{5}}$$

and

$$\big(AMSE\big(b_1^*\big) + R_n\big(b_0^*,b_1^*\big)\big)^{\frac{1}{2}} \asymp \left(\frac{1}{n}\right)^{\frac{2}{5}}.$$

(ii) *For* $d \geqslant 3$, *we have*

$$b_1^* \asymp \left(\frac{1}{n}\right)^{\frac{3}{2d+11}}$$

and

$$\big(AMSE\big(b_1^*\big) + R_n\big(b_0^*,b_1^*\big)\big)^{\frac{1}{2}} \asymp \left(\frac{1}{n}\right)^{\frac{6}{2d+11}}.$$

The results of this theorem show for $d \geqslant 3$, we do not achieve the convergence rate $n^{-2/5}$ for our proposed estimator $\widehat{f}_n(e)$. However, we observe that the order of b_1^* goes to 0 faster than $n^{-1/(d+4)}$, which corresponds to the optimal bandwidth obtained in the case of the classical kernel estimator of a multivariate density. See, for example, Silverman [11] (pp. 84–86).

Our last result concerns the asymptotic normality of the estimator $\widehat{f}_n(e)$.

Theorem 4. *Assume* (A_1)–(A_9) *and*

$$(A_{10}): \quad nb_0^{d+4} = O(1), \qquad nb_0^4 b_1 = o(1), \qquad nb_0^d b_1^3 \to \infty,$$

when $n \to \infty$. *Then,*

$$\sqrt{nb_1}\big(\widehat{f}_n(e) - \overline{f}_n(e)\big) \xrightarrow{d} N\left(0, \frac{f(e)}{\mathbb{P}(X \in \mathcal{X}_0)} \int K_1^2(v)\,\mathrm{d}v\right),$$

where

$$\overline{f}_n(e) = f(e) + \frac{b_1^2}{2} f^{(2)}(e) \int v^2 K_1(v)\,\mathrm{d}v + o\big(b_1^2\big).$$

Acknowledgements

This study was realized when I was in doctoral thesis at Laboratoire de Statistique Théorique et Appliquée (LSTA), Université Pierre et Marie Curie (Paris 6), which support is really acknowledged. I am particularly grateful to Emmanuel Guerre (Queen Mary, University of London) and I acknowledge him for many helpful comments and suggestions. All errors are mine and under my own responsibility.

References

[1] M.G. Akritas, I. Van Keilegom, Non-parametric estimation of the residual distribution, Scand. J. Stat. 28 (2001) 549–567.
[2] F. Cheng, Asymptotic distributions of error density and distribution function estimators in nonparametric regression, J. Statist. Plann. Inference 128 (2005) 327–349.
[3] P. Deheuvels, D.M. Mason, General asymptotic confidence bands based on kernel-type function estimators, Stat. Inference Stoch. Process. 7 (2004) 225–277.
[4] S. Efromovich, Estimation of the density of the regression errors, Ann. Statist. 33 (2005) 2194–2227.
[5] U.U. Müller, A. Schick, W. Wefelmeyer, Estimating the error distribution in nonparametric regression with multivariate covariates, Statist. Probab. Lett. 79 (2009) 957–964.
[6] E.A. Nadaraya, On a regression estimate, Teor. Verojatnost. i Primenen. 9 (1964) 157–159.
[7] N. Neumeyer, I. Van Keilegom, Estimating the error distribution in nonparametric multiple regression with applications to model testing, J. Multivariate Anal. 101 (2010) 1067–1078.
[8] M.S. Pinsker, Optimal filtering of a square integrable signal in Gaussian white noise, Probl. Inf. Transm. 16 (1980) 52–68.
[9] S. Plancade, Nonparametric estimation of the density of the regression noise, C. R. Acad. Sci. Paris, Ser. I 346 (2008) 461–466.
[10] W.S. Scott, Multivariate Density Estimation, Wiley, 1992.
[11] B.W. Silverman, Density Estimation for Statistics and Data Analysis, Monographs on Statistics and Applied Probability, vol. 26, Chapman and Hall, London, 1986.
[12] M.P. Wand, M.C. Jones, Kernel Smoothing, Chapman & Hall/CRC, 1995.
[13] G.S. Watson, Smooth regression analysis, Sankhyā Ser. A 26 (1964) 359–372.