Available online at www.sciencedirect.com
ScienceDirect
C. R. Acad. Sci. Paris, Ser. I 344 (2007) 331–335
COMPTES RENDUS MATHEMATIQUE
http://france.elsevier.com/direct/CRASS1/

Statistics

Merging information for a semiparametric projection density estimation

Jean-Baptiste Aubin, Samuela Leoni-Aubin

Université de technologie de Compiègne, centre de recherches de Royallieu, rue Personne de Roberval, BP 20529, 60205 Compiègne, France

Received 10 October 2006; accepted after revision 10 January 2007

Available online 12 February 2007

Presented by Paul Deheuvels

Abstract

A semiparametric density estimator is proposed under a m-sample density ratio model, which specifies that the ratio of $m-1$ probability density functions with respect to the mth is of a known parametric form without reference to any parametric model. This model arises naturally from retrospective studies and multinomial logistic regression model. A projection density estimator is constructed by smoothing the increments of the maximum semiparametric likelihood estimator of the underlying distribution function, using the combined data from all the samples. We also establish some asymptotic results on the proposed projection density estimator. ***To cite this article: J.-B. Aubin, S. Leoni-Aubin, C. R. Acad. Sci. Paris, Ser. I 344 (2007).***

Résumé

Fusion d'informations pour un estimateur de la densité semiparamétrique par projection. On propose un estimateur semiparamétrique de la densité dans le contexte d'un modèle à rapport de densités à m échantillons. Ce dernier spécifie que les rapports des $m-1$ premières densités par rapport à la dernière sont de forme paramétrique connue. Ce modèle est adapté à des études de type rétrospectif et inclut le cas de la régression logistique multinomial. Nous étudions le problème d'inférence semiparamétrique lié au modèle à rapport de densités en utilisant la méthode de la vraisemblance empirique. Nous utilisons les données combinées des m échantillons afin de calculer un nouvel estimateur adaptatif par projection des densités inconnues. ***Pour citer cet article : J.-B. Aubin, S. Leoni-Aubin, C. R. Acad. Sci. Paris, Ser. I 344 (2007).***

Version française abrégée

On propose un estimateur semiparamétrique de la densité dans le contexte d'un modèle à rapport de densités à m échantillons. Ce dernier spécifie que les rapports des $m-1$ premières densités par rapport à la dernière sont de forme paramétrique connue (voir (1)).

Ce modèle est adapté à des études de type rétrospectif et inclut le cas de la régression logistique multinomial. Nous étudions le problème d'inférence semiparamétrique lié au modèle à rapport de densités en utilisant la méthode de la vraisemblance empirique.

E-mail addresses: jean-baptiste.aubin@utc.fr, aubinjea@dma.utc.fr (J.-B. Aubin), samuela.leoni-aubin@utc.fr (S. Leoni-Aubin).

1631-073X/$ – see front matter
doi:10.1016/j.crma.2007.01.007

Nous utilisons les données combinées des m échantillons afin de calculer un nouvel estimateur adaptatif par projection des densités inconnues (voir (5)). Puis, nous établissons que l'estimateur atteint une vitesse déterminée pour l'erreur quadratique intégrée dans le cas général (voir Théorème 4.1) ainsi que pour différentes vitesses de décroissance des coefficients de Fourier (voir Corollaire 4.2).

1. Introduction

We dispose of m independent random samples $x_{i1}, \dots, x_{in_i}$, $i = 1, \dots, m$, with probability densities $g_i(x) = \mathrm{d}G_i(x)$, $i = 1, \dots, m$, respectively. We consider the following semiparametric density ratio model

$$g_i(x) = w(x, \theta_i) g_m(x), \quad i = 1, \dots, m-1, \tag{1}$$

where w is a known positive function and θ_k, $k = 1, \dots, m-1$, is a vector of parameters with finite dimension equal to d. The supports of the laws G_i may be known or unknown, discrete or continuous. All the m densities functions are assumed unknown but are related, however, through a tilt which determines the difference between them. This approach generalizes the classical normal-based one-way analysis of variance in the sense that it obviates the need for a completely specified parametric model (see [8]). For an application of the density ratio model to meteorological data, see [7].

The density ratio model has attracted much attention recently, because it relaxes several conventional assumptions in the context of multi-samples problems and because fitting can be easily implemented in standard software.

An example of model (1) is provided by multinomial logistic regression, one of the most popular choices for nominal data analysis, with several applications especially in econometrics and biostatistics. Model (1) is general and includes examples such as the exponential family of distribution and has been studied in detail by [8] and [6]. It is useful to say that expression (1) can be viewed as a biased sampling model with weights depending on parameter. Inference for the case $m = 2$ has been studied by [13,9].

The aim of this contribution is to estimate unknown densities in two steps. First, applying the empirical likelihood method to the model (1), then, using a modified projection density estimator.

In Section 2 we recall the estimation method of θ based on the empirical likelihood approach (see [13] and references therein). Section 3, in connection with the theory of Section 2, puts forward modified projection density estimators of the unknown probability density functions. Section 4 pertains to some asymptotic results of the discussed estimators and an idea of the proof of our main result. Some concluding remarks are given in Section 5.

2. Inference in density ratio model

Consider the m samples with corresponding densities that satisfy Eq. (1), let $n := \sum_{i=1}^{m} n_i$ be the total sample size and consider the empirical likelihood (see [10,11]) based on the pooled data $\{x_{ij},\ j = 1, \dots, n_i,\ i = 1, \dots, m\}$

$$L(\theta, G_m) = \left\{ \prod_{j=1}^{n_1} p_{1j} w(x_{1j}, \theta_1) \right\} \left\{ \prod_{j=1}^{n_2} p_{2j} w(x_{2j}, \theta_2) \right\} \dots \prod_{j=1}^{n_m} p_{mj} = \left(\prod_{i=1}^{m} \prod_{j=1}^{n_i} p_{ij} \right) \prod_{i=1}^{m-1} \prod_{j=1}^{n_i} w(x_{ij}, \theta_i)$$

where $p_{ij} = \mathrm{d}G_m(x_{ij})$ and $\theta = (\theta_1^t, \dots, \theta_{m-1}^t)^t$ is a vector of dimension $(m-1)d$. Hence, the log-likelihood writes

$$l(\theta, p) = \sum_{i=1}^{m} \sum_{j=1}^{n_i} \log(p_{ij}) + \sum_{i=1}^{m-1} \sum_{j=1}^{n_i} \log\bigl\{ w(x_{ij}, \theta_i) \bigr\}. \tag{2}$$

Maximization of Eq. (2) is carried out by following a profiling procedure (see [14]), whereby first we express each p_{ij} in terms of θ and then we substitute the p_{ij} back into the likelihood to produce a function of θ only. The profile log-likelihood (in θ) is then $l(\theta) = \sup_{p \in \mathcal{C}_\theta} l(\theta, p)$, where p is constrained to the set

$$\mathcal{C}_\theta := \left\{ p \in \mathbb{R}_+^n \text{ such that } \sum_{i=1}^{m} \sum_{j=1}^{n_i} p_{ij} = 1, \sum_{i=1}^{m} \sum_{j=1}^{n_i} p_{ij} \bigl\{ w(x_{ij}, \theta_k) - 1 \bigr\} = 0,\ k = 1, \dots, m-1 \right\}.$$

The maximization employs the method of Lagrange multipliers, and it follows that if $\mathcal{C}_\theta \neq \emptyset$,

$$p_{ij}(\lambda, \theta) = \frac{1}{n} \frac{1}{1 + \sum_{k=1}^{m-1} \lambda_k \{ w(x_{ij}, \theta_k) - 1 \}}, \tag{3}$$

where λ_k, $k = 1, \ldots, m-1$, are the Lagrange multipliers determined by the following equations:

$$\sum_{i=1}^{m}\sum_{j=1}^{n_i} \frac{w(x_{ij},\theta_k)-1}{1+\sum_{l=1}^{m-1}\lambda_l\{w(x_{ij},\theta_l)-1\}} = 0, \quad k = 1, \ldots, m-1.$$

It turns out that the vector of Lagrange multipliers is a continuously differentiable function of the parameter θ, hence Eq. (2) becomes

$$l\big(\theta,\lambda(\theta)\big) = -\sum_{i=1}^{m}\sum_{j=1}^{n_i}\log\left[1+\sum_{k=1}^{m-1}\lambda_k(\theta)\big\{w(x_{ij},\theta_k)-1\big\}\right] + \sum_{i=1}^{m-1}\sum_{j=1}^{n_i}\log\big[w(x_{ij},\theta_i)\big] - n\log n.$$

Under some regularity conditions (see [6]), as $n \to \infty$, $\hat{\theta}$ and $\hat{\lambda} = \lambda(\hat{\theta})$ exist, are consistent, satisfy the following system of estimating equations:

$$\frac{\partial l(\theta,\lambda)}{\partial\theta_k} = -\sum_{i=1}^{m}\sum_{j=1}^{n_i}\frac{\lambda_k\partial w(x_{ij},\theta_k)/\partial\theta_k}{1+\sum_{l=1}^{m-1}\lambda_l\{w(x_{ij},\theta_l)-1\}} + \sum_{j=1}^{n_k}\frac{\partial[\log\{w(x_{kj},\theta_k)\}]}{\partial\theta_k} = 0,$$

$$\frac{\partial l(\theta,\lambda)}{\partial\lambda_k} = -\sum_{i=1}^{m}\sum_{j=1}^{n_i}\frac{w(x_{ij},\theta_k)-1}{1+\sum_{l=1}^{m-1}\lambda_l\{w(x_{ij},\theta_l)-1\}} = 0, \quad k = 1, \ldots, m-1,$$

and are asymptotically normal.

We turn now to the question of modified density estimation based on the above inference output.

3. New semiparametric density estimators

Ref. [6] proposes semiparametric density estimators by modifying classical kernel density estimators (essentially by smoothing the increments of $\widehat{G}_i$, $i = 1, \ldots, m$). Moreover, he shows that in the density ratio model (1), the pooled data leads to more efficient kernel density estimators for the unknown distributions, in the sense that they have the same amount of bias but they are less variable than traditional kernel density estimators.

Refs. [5] and [15] study the problem of kernel density estimation under the two-sample density ratio model, that is model (1) with $m = 2$. The bandwidth selection criterion of [5] is based on a least square cross validation scheme whereas the resulting density estimators are employed for a goodness-of-fit test of the two-sample density ratio model, using the L_2 norm of the difference between the semiparametric and the nonparametric kernel density estimators. [15] establish the asymptotic normality of estimator.

Another density estimation method is the one obtained by projection. It consists in projecting the density to estimate on a finite dimensional space (for example the one generated by the first components of a basis of the space $\mathcal{G}$ of the possible densities) and estimate this projection by a moments method (see [4]).

In the following, we will design the pooled data $(x_{11}, \ldots, x_{1n_1}, \ldots, x_{m1}, \ldots, x_{mn_m})$ by $(t_1, \ldots, t_n)$. We suppose that for $l = 1, \ldots, m$, the l-th sample admits the density g_l with respect to μ such that $g_l \in L^2(\mu)$, where μ is a finite measure. Let $(e_j)_{j\in\mathbb{N}^*}$ be some orthonormal basis in a separable infinite-dimensional Hilbert space $L^2(\mu)$.

The classical projection density estimator (see [4]) for $g_m = \sum_{j=1}^{\infty} a_j e_j$ (the density of the m-th sample $(x_{m1}, \ldots, x_{mn_m})$) is $\tilde{g}_{m_{n_m}} = \sum_{j=1}^{k_{n_m}} \bar{a}_{j,n_m} e_j$, where (k_{n_m}) is the truncation index sequence such that $k_{n_m} \leqslant n_m$, $(k_{n_m}) \uparrow \infty$ and $\frac{k_{n_m}}{n_m} \to 0$ when $n_m \uparrow \infty$. $\bar{a}_{j,n_m} = \frac{1}{n_m}\sum_{i=1}^{n_m} e_j(x_{mi})$ is the unbiased estimate of the j-th Fourier coefficient a_j. In this Note, we consider the case in which $(e_j)_{j\in\mathbb{N}^*}$ is uniformly bounded (such that $\exists M < \infty$: $\sup_j \|e_j\|_\infty < M$).

However, one can merge the information coming from the m samples (instead of only from the last one) whose densities are linked by the model (1), in order to estimate more efficiently g_m.

Here, every t_i is associated to a p_i, $i = 1, \ldots, n$ (see (3)), which is estimated by the empirical likelihood method. We recall that the p_i verify $\sum_{i=1}^{n} p_i = 1$.

Our modified projection density estimator of g_m is then

$$\hat{g}_{m_n} = \sum_{j=1}^{k_n} \hat{a}_{j,n} e_j, \quad \text{with } \hat{a}_{j,n} := \sum_{i=1}^{n} \hat{p}_i e_j(t_i), \tag{4}$$

where (k_n) is chosen by the user and is such that $k_n \leqslant n$, $(k_n) \uparrow \infty$ and $\frac{k_n}{n} \to 0$ when $n \uparrow \infty$.

Furthermore, Eq. (4) is a semiparametric density estimator since it depends on both the unknown distribution function and the parameters of the model (1). We show that, if we choose the projection basis $(e_j)_{j\in\mathbb{N}^*}$ a good way, then the AMISE of the estimator defined in (4) can be close to $\frac{1}{n}$ (see Corollary 4.2). This rate (almost a 'parametric' one) is better than that obtained by the kernel estimation method, both in the cases of classical and semiparametric density estimation (see [5,6,15]).

We deduce projection estimators for the other densities g_l, $l = 1, \ldots, m-1$, as following:

$$\hat{g}_{l_n} = \sum_{j=1}^{k_n} \hat{a}_{j,n} e_j, \quad \text{with } \hat{a}_{j,n} := \sum_{i=1}^{n} \hat{p}_i\, w(t_i, \hat{\theta}_l) e_j(t_i),\ l = 1, \ldots, m-1, \tag{5}$$

where (k_n) is chosen in the same way as before. Obviously, these estimators enjoy the same asymptotic properties of (4).

4. Asymptotic results

In this section, we consider the asymptotic mean integrated square error (AMISE) of the semiparametric projection density estimator $\hat{g}_{m_n}$ (4) as a measure of its global accuracy.

Theorem 4.1. *Under classical regularity conditions* (*see hypotheses in* [6]*, Theorem* 1)*, we have*

$$\mathbb{E}\|\hat{g}_{m_n} - g_m\|^2 = \mathcal{O}\left(\frac{k_n}{n}\right) + \sum_{j>k_n} a_j^2.$$

Elements of proof. To demonstrate the previous result concerning the projection density estimator $\hat{g}_{m_n}$, it is useful to consider

$$\tilde{g}_{m_n} = \sum_{j=1}^{k_n} \tilde{a}_{j,n} e_j, \quad \text{with } \tilde{a}_{j,n} := \sum_{i=1}^{n} p_i e_j(t_i).$$

We show that $\tilde{a}_{j,n}$ is an unbiased estimator of a_j, that $\hat{g}_{m_n} - \tilde{g}_{m_n} = \mathcal{O}_p(\frac{1}{\sqrt{n}})$ and conclude by noting that $\operatorname{Var}(\tilde{a}_{j,n}) = \mathcal{O}(\frac{1}{n})$. □

Theorem 4.1 reveals the common trade-off problem between random and systematic error, i.e. large values of k_n reduces bias but introduce substantial variance as opposed to small values of the truncation index which lead to smaller variance but increased bias.

Corollary 4.2. *Under conditions of Theorem* 4.1*,*

(i) *If* $\forall j \geqslant 1,\ |a_j| < \gamma j^{-\rho}$ *where* $\gamma > 0$ *and* $\rho > 1/2$*, then, for* $k_n^* = n^{1/(2\rho)}$*,*

$$\mathbb{E}\|\hat{g}_{m_n} - g_m\|^2 = \mathcal{O}\big(n^{(1-2\rho)/(2\rho)}\big).$$

(ii) *If* $\forall j \geqslant 1,\ |a_j| < \alpha\beta^{-j}$ *where* $\alpha > 0$ *and* $\beta > 1$*, then, for* $k_n^* = \frac{\log n}{2\log\beta}$*,*

$$\mathbb{E}\|\hat{g}_{m_n} - g_m\|^2 = \mathcal{O}\left(\frac{\log n}{n}\right).$$

Corollary 4.2 means that a strong enough decrease of the Fourier coefficients (in case (i) with $\rho > \frac{5}{2}$ and in case (ii), included in case (i)) implies that a good choice of (k_n) gives us a density estimator which reduces the AMISE when it is compared with that of the semiparametric kernel density estimator (see [5,6] and [15]). A strong decrease of the Fourier coefficients means that the user chooses a suitable projection basis $(e_j)_{j\in\mathbb{N}^*}$ with respect to the density to estimate.

Proposition 4.3. *Under hypotheses of Theorem* 4.1*, for the same truncation index selection, if* $\exists l < m$ *such that* $n_l \neq 0$*, then*

$$MISE(\hat{g}_{m_n}) < MISE(\bar{g}_{m_{n_m}}).$$

5. Conclusions and perspectives

We studied the semiparametric inference problem that is related to the density ratio model by appealing to the methodology of empirical likelihood. We used the combined data from all the samples to calculate a new projection density estimator for the unknown distributions.

The required computation for our method can be accomplished by using the standard statistical software packages. We also established some asymptotic results on the proposed projection density estimator.

This new estimator was shown to be more efficient than the semiparametric kernel one for suitable projection bases, and it reduces the AMISE when it is compared with that of the traditional projection density estimator, in the sense that the pooled data yield estimate with the same amount of bias but which are less variable for the same truncation index selection.

Some authors have studied a data-driven version of the projection density estimator (see [1–3,12]). This estimator enjoys some local suroptimality properties for the AMISE. An extension of this work can be the use of this data driven estimator (instead of the classical one) in a semiparametric context.

Acknowledgements

We are grateful to the referee for many constructive comments and suggestions that have greatly improved our original submission and to Professor Denis Bosq for helpful discussions.

References

[1] J.B. Aubin, A. Massiani, Comportement asymptotique d'un estimateur de la densité adaptatif par méthode d'ondelettes, C. R. Acad. Sci. Paris, Ser. I 337 (4) (2003) 293–296.
[2] D. Bosq, Estimation localement suroptimale et adaptative de la densité, C. R. Acad. Sci. Paris, Ser. I 334 (7) (2002) 591–595.
[3] D. Bosq, Inférence et prévision en grandes dimensions, Economica, 2005.
[4] N.N. Cencov, Estimation of unknown distribution density from observations, Soviet Math. Dokl. 3 (1962) 1559–1562.
[5] K.F. Cheng, C.K. Chu, Semiparametric density estimation under a two-sample density ratio model, Bernoulli 10 (4) (2004) 583–604.
[6] K. Fokianos, Merging information for semiparametric density estimation, Part 4, J. R. Statist. Soc. B 66 (2004) 941–958.
[7] K. Fokianos, B. Kedem, J. Qin, J. Haferman, D.A. Short, On combining instruments, J. Appl. Meteorology 37 (1998) 220–226.
[8] K. Fokianos, B. Kedem, J. Qin, D.A. Short, A semiparametric approach to the one-way layout, Technometrics 43 (2001) 56–64.
[9] A. Keziou, S. Leoni-Aubin, Test of homogeneity in semiparametric two-sample density ratio models, C. R. Acad. Sci. Paris, Ser. I 340 (12) (2005) 905–910.
[10] A.B. Owen, Empirical likelihood ratio confidence intervals for a single functional, Biometrika 75 (2) (1988) 237–249.
[11] A.B. Owen, Empirical Likelihood, Chapman and Hall, New York, 2001.
[12] D. Picard, K. Tribouley, Adaptive confidence interval for pointwise curve estimation, Ann. Statist. 28 (1) (2000) 298–335.
[13] J. Qin, Inferences for case-control and semiparametric two-sample density ratio models, Biometrika 85 (3) (1998) 619–630.
[14] J. Qin, J. Lawless, Empirical likelihood and general estimating equations, Ann. Statist. 22 (1) (1994) 300–325.
[15] B. Qin, J. Zhang, Density estimation under a two-sample semiparametric model, Nonparametr. Statist. 17 (6) (2005) 665–683.