Minimum Entropy Estimators in Semiparametric Regression Problems

Eric Wolsztynski, Eric Thierry, and Luc Pronzato

Laboratoire I3S, UNSA-CNRS
2000, route des lucioles, Les Algorithmes - bât. Euclide B, BP.121
06903 Sophia Antipolis Cedex, France
(e-mail: {pronzato,et,wolsztyn}@i3s.unice.fr)

Abstract. We consider semiparametric regression problems for which the response function is known up to some vector of parameters θ and the errors have an unknown density f, treated as an infinite-dimensional nuisance parameter for the estimation of θ. The maximum likelihood (ML) estimator is clearly unapplicable in this context, and classical approaches like least squares or M-estimation may perform poorly. Since the results of Stein in 1956, a large amount of work was dedicated to the construction of adaptive estimators that have the same asymptotic behavior as the ML estimator (*asymptotic efficiency*). The focus has been mainly set on the asymptotic theory and the practical results seem to be restricted to the case of scalar observations.

We presented in [Pronzato *et al.*, 2004] an estimator that minimizes the entropy of the symmetrized sample of the residuals. In [Wolsztynski *et al.*, 2005] we show the link between this Minimum Entropy (ME) estimator, the ML estimator, and the two-stage adaptive estimator of [Bickel, 1982]. Also, we show that the shift-invariance property of entropy confers some robustness to ME estimation.

Adaptive estimation has important applications in Signal and Image Processing. The present paper summarizes the theoretical aspects of the ME approach and focuses on such applications. Although asymptotic properties are commonly the main concern, we illustrate the performances of estimators for finite samples through simulations, including multidimensional situations. The examples we consider also illustrate the robustness of ME estimation.
Keywords: Adaptivity, efficiency, entropy estimation, multivariate regression, semiparametric estimation.

1 Introduction

We consider nonlinear regression models that we assume to be known up to some vector of parameters $\theta \in \Theta \subset \mathbb{R}^p$. We denote $\eta(\bar{\theta}, x)$ the model response, where $\bar{\theta} \in \mathrm{int}(\Theta)$ is the true unknown value of θ and $x \in \mathcal{X} \subset \mathbb{R}^q$ are the design variables. In what follows the design can be randomized or fixed. The observations $Y_i \in \mathbb{R}^d$, $d \geq 1$, are given by

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

with (ε_i) a sequence of independent and identically distributed (i.i.d.) random variables with probability density function (p.d.f.) f with respect to the

Lebesgue measure. For a given measure μ on the design variable x we suppose that the identifiability condition $\left[\int_{\mathcal{X}} [\eta(\theta, x) - \eta(\bar{\theta}, x)]^2 \mu(dx) = 0 \;\Rightarrow\; \theta = \bar{\theta}\right]$ is satisfied. The only assumptions that we make on f, along with some usual regularity conditions, are that it is (centrally) symmetric about 0 and has unbounded support. The density of the noise then corresponds to an infinite-dimensional nuisance parameter for the estimation of θ, and an estimator that remains *asymptotically efficient* in this context is termed *adaptive* (whenever it exists). [Bickel, 1982] and then [Manski, 1984] established that adaptivity was possible for nonlinear regression models.

Consider the residuals $e_i(\theta)$ obtained from the observations (1),

$$e_i(\theta) = Y_i - \eta(\theta, X_i) = \varepsilon_i + \eta(\bar{\theta}, X_i) - \eta(\theta, X_i),\ i = 1, \ldots, n. \tag{2}$$

We suggest in [Pronzato *et al.*, 2004] an estimator of θ that minimizes an estimate of the entropy of the residuals in the univariate case. Since entropy is shift-invariant, we use the $2n$ symmetrized residuals $\pm e_i(\theta)$ with density given X_i

$$f^s_{e,\,X_i}(u) = \frac{1}{2}\left[f(u - \eta(\bar{\theta}, X_i) + \eta(\theta, X_i)) + f(u + \eta(\bar{\theta}, X_i) - \eta(\theta, X_i))\right]\ . \tag{3}$$

Using classical results of Information Theory, we show in [Wolsztynski *et al.*, 2005] that the (Shannon) entropy $H(f^s_e) = -\int f^s_e(e) \log f^s_e(e) \mu(de)$ of the marginal distribution of the symmetrized residuals, $f^s_e(u) = \int_{\mathcal{X}} f^s_{e,\,x}(u) \mu(dx)$, is minimum for $\theta = \bar{\theta}$ when the identifiability condition given above is satisfied. When f is unknown, an estimator of $H(f^s_e)$ thus provides a criterion for the estimation of θ. Moreover, we shall see that the shift-invariance property of the entropy makes minimum entropy (ME) estimation robust with respect to the presence of outlying data.

In [Wolsztynski *et al.*, 2005] we show the link between a two-step ME estimation procedure and the adaptive Stone-Bickel approach for univariate observations. The construction involves data splitting, which allows for the estimate of the density to be independent of that of the entropy, and the application of a single Newton step onto a preliminary locally sufficient estimator then provides an asymptotically efficient estimator of θ.

In the next section we consider two direct ME estimation procedures (without data splitting) for multidimensional data samples. Two examples illustrate the performance of our technique in Section 3.

2 Direct Minimum Entropy estimation procedures

The direct ME estimator that we proposed for univariate data is constructed by plugging a kernel density estimate $\hat{f}^{\theta}_n$ of f^s_e based on the $2n$ symmetrized

residuals $\pm e_i(\theta)$ in an empirical expression of the entropy. The density estimate we use is given by

$$\hat{f}_n^\theta(u) = \frac{1}{2nh_n} \sum_{i=1}^{n} \left[K\left(\frac{u - e_i(\theta)}{h_n} \right) + K\left(\frac{u + e_i(\theta)}{h_n} \right) \right] ,$$

with h_n a smoothing parameter, and is used to construct the Ahmad-type plug-in entropy estimator

$$\hat{H}_n(\theta) = -\frac{1}{n} \sum_{i=1}^{n} \log \hat{f}_n^\theta(e_i(\theta)) \,. \tag{4}$$

This provides a fully non parametric estimate that can be used for the estimation of θ without data splitting. An alternative entropy estimator can also be constructed by using a truncated integral of $\hat{f} \log \hat{f}$ instead of the sum in (4), but in practice the two estimators turn out to be quite close in performance (although the results might vary in function of the selected bandwidth and of the nature of the problem). For the simple case of the location model we give in [Pronzato *et al.*, 2004] a justification for this method and in [Wolsztynski *et al.*, 2004] we show that $\hat{H}_n(\theta) \xrightarrow{p} H(f_e^s) \geq H(f)$ uniformly in θ, $n \to \infty$, with $H(f_e^s) = H(f)$ for $\theta = \bar{\theta}$, provided that the kernel bandwidth h_n decreases slowly enough and f and K satisfy some regularity conditions. Under slightly stronger conditions, we also prove that $\nabla^2 \hat{H}_n(\theta) \xrightarrow{p} \nabla^2 H(f_e^s)$ uniformly in θ, $n \to \infty$, with $\nabla^2 H(f_e^s) = \nabla^2 H(f) = \mathcal{I}(f)$ for $\theta = \bar{\theta}$. However, proving adaptivity of this direct approach remains an open challenge.

Consider now multidimensional observations. In the case of independent components, the entropy of the residuals is the sum of the entropies of each component (i.e. $H(f_e^s) = \sum_{j=1}^{d} H(f_{e^j}^s)$). The construction used in (4) is therefore suitable to obtain the entropy of the residuals as the sum of the entropies of each marginal distribution.

In the general situation where independence of components does not necessarily hold, we can extend the procedure above by simply using techniques of multivariate density estimation. For small dimensions (2 or 3), techniques based on products of univariate kernels are computationally efficient, see for instance [Scott, 1992]. Given $K(.)$ a univariate density that is symmetric about zero, we thus propose to use

$$\hat{f}_n^\theta(u) = \frac{1}{2\,n} \left[\sum_{i=1}^{n} \prod_{j=1}^{d} \frac{1}{h_j} K\left(\frac{u^j - e_i^j(\theta)}{h_j} \right) + \sum_{i=1}^{n} \prod_{j=1}^{d} \frac{1}{h_j} K\left(\frac{u^j + e_i^j(\theta)}{h_j} \right) \right] , \tag{5}$$

where $h_j = h_j(\pm e_1^j(\theta), \ldots, \pm e_n^j(\theta), K)$ is the bandwidth of the univariate kernel K based on the j-th component of the symmetrized sample of

residuals, with K satisfying common regularity conditions. One can choose, e.g., K as the standard normal, in which case the optimal bandwidth (in the sense of the asymptotic mean integrated squared error) is given by $h^\star = (4/(d+2))^{1/(d+4)}\,\sigma_i n^{-1/(d+4)}$, see [Scott, 1992]. Plugging (5) into an Ahmad-type estimate of the entropy similar to (4), we obtain a criterion for estimating θ from multidimensional data. In practice, one can define a data-driven selection of h by substituting the estimated standard deviation of the residuals on each component for its exact value into the expression of $h^\star$. Notice that $\hat{H}_n(\theta)$ is two times continuously differentiable w.r.t. $\theta \in \mathrm{int}(\Theta)$ when $\eta(\theta, x)$ is smooth enough.

In higher dimensions, however, kernel estimation techniques rapidly become inefficient. The major limitation comes from the choice of the bandwidth $h(\pm e_1(\theta), \ldots, \pm e_n(\theta), K)$: due to the curse of dimensionality, the bandwidth for each kernel must be large enough to take a sufficient number of data points into account, which causes oversmoothing. The main alternatives involve kernels that are not positive everywhere [Härdle and Linton, 1994], which is not suitable for computing entropy, non-differentiable density estimates, see for instance [Türlach, 1994], and kernel methods with variable bandwidth [Scott, 1992, Devroye and Lugosi, 2000]. We consider now a special case of the latter.

We suggest here a simple alternative that uses the k-nearest neighbor (kNN) entropy estimator as introduced by [Kozachenko and Leonenko, 1987] for $k = 1$ and extended to $k > 1$ in [Goria *et al.*, 2005], where its consistency is proved for general dimension d under very weak conditions on f.

Consider the open ball $v(x, r)$ centered on $x \in \mathbb{R}^d$ with radius $r > 0$; its volume is given by $|v(x, r)| = r^d c_1(d)$, where $c_1(d) = 2\pi^{d/2}/(d\Gamma(d/2))$. Denote the Euclidean distance from $e_i(\theta)$ to its k-th nearest neighbor by $\rho_{i,\,k}(\theta)$. For the symmetrized residuals $\pm e_j(\theta)$, the kNN-ME estimator of θ then minimizes

$$H_{k,\,n}(\theta) = d \log \bar{\rho}_k(\theta) + T(n, k)\,, \tag{6}$$

where $\bar{\rho}_k(\theta) = \left(\Pi_{i=1}^{2n} \rho_{i,\,k}(\theta)\right)^{1/2n}$ is the geometric mean of the kNN distances and $T(n, k) = \log(n-1) - \psi(k) + \log c_1(d)$ does not depend on θ, with $\psi(k) = \Gamma'(k)/\Gamma(k)$, the digamma function.

The parameter k can be chosen so that $k/n \to 0$, $k \to \infty$ when $n \to \infty$; a typical choice is $k = \sqrt{n}$. We shall take $k > p$ where $p = \dim(\theta)$ to avoid singularities. Notice that (6) is not differentiable in θ.

Although asymptotic results are not yet available for this procedure, we present it here as a simple computational alternative. In the next section we present two examples in image processing for 3-dimensional data.

Note that one could consider the estimate (6) of the entropy as another plug-in entropy estimate, that is, as a generalization of the method of kernels (but avoiding the tricky problem of bandwidth selection). Indeed, consider the ball $v(x, \rho_k)$ mentioned above, where $\rho_k = \rho_k(x)$ is the distance from $x \in \mathbb{R}^d$ to its k-th closest point; [Devroye and Wagner, 1977] proved the strong consistency of the kNN p.d.f. estimate $\hat{f}_n(x) = k \left[n \left(\rho_k(x) \right)^d c_1(d) \right]^{-1}$. The ME estimator based on (6) can thus be written as a multivariate plug-in estimator (with a different bias correction term).

3 Examples

We present some simulation results obtained on images, where the estimator of θ is obtained through an exhaustive search on a finite grid. In this context, entropy is a very natural criterion given its key role in coding theory for the definition of maximum compression rates (or equivalently of minimum description lengthes). Minimizing the entropy of the errors between two signals or two images is equivalent to choosing the parameters for which we achieve the maximum compression rate.

We take a 176×144 png picture for the first example (scalar residuals), and a 352×288 jpg one for the second example, which gives 3-dimensional residuals. Here the observations correspond to a bloc A of an image that is contaminated with additive noise. The problem is to locate the corresponding bloc in a copy X of the original image, also contaminated with noise. We suppose that this copy has not suffered from any nonlinear transformation. The coordinates $\bar{\theta}$ of A are measured from the top-left corner of the original image, and θ is therefore a two-dimensional vector. The dimension of the observations corresponds to the number of channels that make each pixel: 1 channel describes the gray level in the black and white png file, whereas 3 channels (RGB) contain the levels of coloring in the color jpg file. Figure 1 shows, clockwise from top left, (a) the 15×15 bloc A, within the small square, to be identified in (b), the working image, that contains 2×6 outliers (white patch); (c) the 30×30 bloc A to be located in the color image (d). (a) and (b) are black and white pictures contaminated by gaussian noise of variance 6; (c) and (d) are in color and are contaminated by salt and pepper noise, where 6% of pixel values are replaced by the maximum or minimum possible values and contaminated pixels are randomly distributed on the image.

In the first example, images (a) and (b), we compare the LS estimator, the Least Absolute Values (LAV) estimator (which minimizes the sum of the absolute values of the residuals), the plug-in ME (piME) estimator given by (4) and the kNN-ME estimator given by (6). The bandwidth h for piME is set to $.2345\ \sigma\,(2n)^{-1/5}$ (which is optimal in the sense of minimal mean integrated squared error for gaussian kernels, see e.g. [Berlinet and

Fig. 1. *Images a, b (black and white, top), c, d (color, bottom).*

Devroye, 1994]) and the value k for the kNN estimator is set to 5. The true value of the parameters of interest is $(80, 70)$. Table 1 contains the means of the estimates obtained for 100 runs of the experiment described above. The two ME estimators estimate $\bar{\theta}$ without error in 100% of the runs and thus appear insensitive to the presence of outliers (the white patch).

In the second example, images (c) and (d), we compare the kNN-ME estimator (6) with a piME-o estimator using the optimal bandwidth $h^\star$, a second piME-e estimator using the estimated bandwidth $\hat{h}$ defined by $\hat{h}_j = \hat{\sigma_j}(2n)^{-1/(d+4)}$ for each component of the observations (with $\hat{\sigma_j} = \hat{\sigma_j}(\theta)$ the estimated value of the standard deviation of the $e_i^j(\theta)$, $i = 1, \ldots, n$), see [Scott, 1992], and the standard LS estimator. Figure 2 shows (clockwise from top left) a typical plot of the respective criteria as functions of θ; here $\bar{\theta} = (140, 170)$. Note the good behavior of the kNN and piME-o estimators, and the loss of accuracy due to the estimation of the smoothing parameter h for piME-e. The LS criterion gives $\hat{\theta}_{\texttt{LS}} = (136, 173)$ and its shape suggests that it is not suitable for such problems. The value of the entropy of the symmetrized residuals estimated by (6) is 9.99 for $\hat{\theta}_{\texttt{LS}}$, as opposed to -0.23 for $\hat{\theta}_{\texttt{kNN}} = \bar{\theta}$.

Table 1. Mean values of the estimates for 100 runs on a black and white picture; $\bar{\theta} = (80, 70)$.

LS	LAV	kNN	piME
$(94.25, 67.51)$	$(94.17, 68.26)$	$(80.00, 70.00)$	$(80.00, 70.00)$

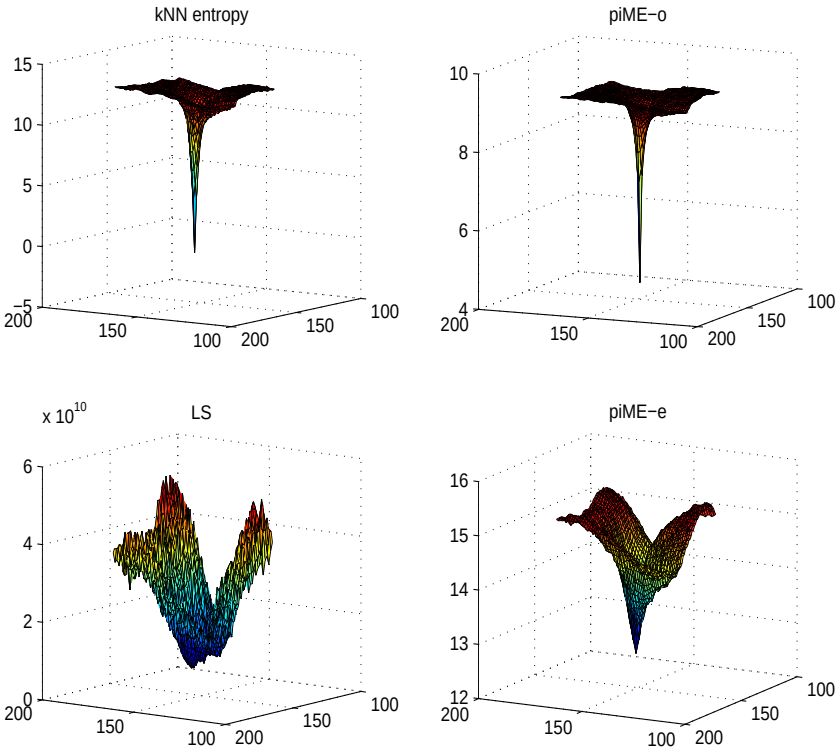

Fig. 2. *criteria vs θ in Example 2, clockwise from top-left : kNN, piME-o, piME-e, LS.* $\bar{\theta} = (140, 170)$.

4 Acknowledgements

We are thankful to Professor Michel Barlaud and Thomas André from Laboratoire I3S for their suggestion of applications in image processing, and for having provided us with the images used in the examples.

References

[Berlinet and Devroye, 1994]A. Berlinet and L. Devroye. A comparison of kernel density estimates. *Publications de l'Institut de Statistique de l'Université de Paris*, 38(3):3–59, 1994.

[Bickel, 1982]P.J. Bickel. On adaptive estimation. *Ann. Stat.*, 10:647–671, 1982.

[Devroye and Lugosi, 2000]L.P. Devroye and G. Lugosi. Variable kernel estimates: on the impossibility of tuning the parameters. In D. Mason E. Giné and J.A. Wellner, editors, *High-Dimensional Probability II*, pages 405–424. Springer-Verlag, New York, 2000.

[Devroye and Wagner, 1977]L.P. Devroye and T.J. Wagner. The strong uniform consistency of nearest neighbor density estimates. *Ann. Stat.*, 5(3):536–540, 1977.

[Goria *et al.*, 2005]M.N. Goria, N.N. Leonenko, V.V. Mergel, and P.L. Novi Inverardi. A new class of random vector entropy estimators and its applications in testing statistical hypotheses. *Jour. Nonparam. Stat.*, 2005. Submitted.

[Härdle and Linton, 1994]W. Härdle and O. Linton. *Applied Nonparametric methods. In "Hanbook of Econometrics"*, volume IV. Elsevier Science B.V., 1994.

[Kozachenko and Leonenko, 1987]L. Kozachenko and N. Leonenko. On statistical estimation of entropy of random vector. *Problems Infor. Transmiss.*, 23(2):95–101, 1987.

[Manski, 1984]C. Manski. Adaptive estimation of nonlinear regression models. *Econometric Reviews*, 3(2):145–194, 1984.

[Pronzato *et al.*, 2004]L. Pronzato, E. Thierry, and E. Wolsztynski. Minimum entropy estimation in semi-parametric models: a candidate for adaptive estimation? In Di Bucchianico, Läuter, and Wynn, editors, *mODa'7 – Advances in Model–Oriented Design and Analysis, Heeze (Netehrlands)*, pages 125–132, Heidelberg, 2004. Physica Springer-Verlag.

[Scott, 1992]D.W. Scott. *Multivariate Density Estimation: Theory, Practice, and Visualization*. Wiley, 1992.

[Stein, 1956]C. Stein. Efficient nonparametric testing and estimation. In *Proc. 3rd Berkeley Symp. Math. Stat. Prob.*, volume 1, pages 187–196. University of California Press, Berkeley, 1956.

[Türlach, 1994]B.A. Türlach. Fast implementation of density-weighted average derivative estimation. *Computationally Intensive Statistical Methods*, 26:28–33, 1994.

[Wolsztynski *et al.*, 2004]E. Wolsztynski, E. Thierry, and L. Pronzato. Consistency of a minimum-entropy estimator of location. Internal Report No I3S/RR-2004-38-FR, 30 pages, http://www.i3s.unice.fr/~mh/RR/rapports.html, 2004.

[Wolsztynski *et al.*, 2005]E. Wolsztynski, E. Thierry, and L. Pronzato. Minimum-entropy estimation in semi-parametric models. *Signal Processing, Special Issue on Information Theoretic Signal Processing*, 2005.