Metrika
DOI 10.1007/s00184-009-0246-3

Estimation of Fisher information using model selection

Jan Mielniczuk · Małgorzata Wojtyś

Received: 16 March 2008

Abstract In the paper the problem of estimation of Fisher information I_f for a univariate density supported on [0, 1] is discussed. A starting point is an observation that when the density belongs to an exponential family of a known dimension, an explicit formula for I_f there allows for its simple estimation. In a general case, for a given random sample, a dimension of an exponential family which approximates it best is sought and then estimator of I_f is constructed for the chosen family. As a measure of quality of fit a modified Bayes Information Criterion is used. The estimator, which is an instance of Post Model Selection Estimation method is proved to be consistent and asymptotically normal when the density belongs to the exponential family. Its consistency is also proved under misspecification when the number of exponential models under consideration increases in a suitable way. Moreover we provide evidence that in most of considered parametric cases the small sample performance of proposed estimator is superior to that of kernel estimators.

Keywords Fisher information · Post model selection estimation · Bayes information criterion (BIC) · Kernel estimator of a density · Sheather–Jones bandwidth

1 Introduction

Consider a univariate probability distribution having a density f and such that its Fisher information

J. Mielniczuk (✉) · M. Wojtyś
Faculty of Mathematics and Information Science, Warsaw University of Technology, Plac Politechniki 1, 00-661 Warsaw, Poland
e-mail: J.Mielniczuk@mini.pw.edu.pl

J. Mielniczuk
Institute of Computer Science, Polish Academy of Sciences, Warsaw, Poland

Published online: 28 February 2009

$$I_f = \int\limits_{\mathbb{R}} \frac{[f'(x)]^2}{f(x)}\, dx. \tag{1}$$

is finite. Obviously, $I_f = \mathbb{E}[(f'/f)^2(X)]$, where X is a random variable distributed according to density f. The Fisher information is a quantity appearing in many contexts in statistics and its nonparametric estimation is essential in construction of efficient estimators and tests, see, e.g. Bickel et al. (1998).

In the paper we deal with a problem of nonparametric estimation of this quantity. An obvious method to do this would be to estimate a density and its derivative by one of numerous nonparametric functional estimation methods and then use the definition of I_f for constructing its plug-in type estimator. This is an approach used, e.g. in Stone (1975) and van der Vaart (2000, Section 25.8.1). However, such an estimator of $f'(x)/f(x)$ could behave unstably for x corresponding to small values of the density. A remark by last cited author (p. 397), who used truncation of both density and derivative estimators, that this construction 'is not necessarily recommended for use in practice' is telling. Moreover, with such an approach one faces a delicate problem of choosing smoothing parameters for estimators of both density and its derivative. Note also that I_f equals to the integral of the squared optimal score $J(y) = f'/f(F^{-1}(y))$. Thus another possibility is to apply this representation of I_f in terms of the optimal score function and use a direct estimator of it which does not involve separate estimation of f and its derivative. An example of such direct estimator is NN estimator proposed by Csörgő and Révész (1986) and defined in Sect. 2.3.

Here we take a different route. Assume that f is supported on [0, 1] and consider k-dimensional exponential family $\mathcal{M}_k$ pertaining to orthonormal functions $b_1(x), \ldots, b_k(x)$ defined in (2)–(3). If $f \in \mathcal{M}_k$ then $I_f = \mathbb{E}\{\sum_{j=1}^{k} \theta_j b_j'(X)\}^2$ (cf. 14), where X is distributed according to density f. This equality is used to construct an estimator of I_f. In general, when it is not known whether f belongs to an exponential family, we construct a parsimonious exponential model $\mathcal{M}_k$ well approximating f in terms of a modified Bayesian Information Criterion (BIC) due to Schwarz (1978) and use an estimator constructed for this very model. The procedure known as post model selection estimation is described in detail below. An important point is that nonparametric estimation of the density f and its derivative is entirely avoided and the whole procedure relies on estimation of dimension k and parametric estimation of θ by maximum likelihood method in the exponential family. We also note that the introduced method can be applied to estimate other functionals of a density such as its entropy or an integral of its squared derivative of a given order.

In the paper we prove consistency and asymptotic normality of the introduced estimator when the underlying distribution belongs to one of parametric models under consideration and discuss some properties of the related selection rule. Its consistency is also proved under misspecification when the underlying density is sufficiently well approximated by the models on the list (cf. Theorem 5). Moreover, we compare its performance with plug-in estimators and estimator based on Csörgő and Révész estimator of a score function.

The paper is structured as follows. Section 2 contains necessary preliminary information on exponential families as well as post model selection estimators (PMSE) and introduces the proposed estimator of I_f together with its competitors. In Sect. 3 main results of the paper are stated and proved. The last section is devoted to discussion of results of simulation study, in which the case of a density with an unknown, possibly unbounded, support is also addressed.

2 Preliminaries

2.1 Exponential family

Let $\mathcal{M}_k$, for $k \geq 1$, be k-dimensional exponential family

$$\mathcal{M}_k = \{f(\cdot, \theta) \ : \ \theta \in \mathbb{R}^k\}, \tag{2}$$

where $f(\cdot, \theta) = f_\theta(\cdot)$ is a density function with respect to Lebesgue measure λ on $[0, 1]$ defined as

$$f(x, \theta) = \exp\{\theta \circ b(x) - \psi_k(\theta)\}, \tag{3}$$

where $\theta = (\theta_1, \ldots, \theta_k)^T \in \mathbb{R}^k$ and $b(x) = (b_1(x), \ldots, b_k(x))^T$ is a vector of known bounded and differentiable functions such that $\{1, b_1(x), \ldots, b_k(x)\}$ form an orthonormal system in $(L^2([0, 1]), \lambda)$. The symbol $\circ$ denotes throughout the inner product in $\mathbb{R}^k$; for $x \in \mathbb{R}^k$, $y \in \mathbb{R}^j$, when $j \leq k$, we define $x \circ y = \sum_{i=1}^{j} x_i y_i$.

Since f is a density on $[0, 1]$, a normalizing constant $\psi_k(\theta)$ is equal to

$$\psi_k(\theta) = \log \int_0^1 \exp\{\theta \circ b(x)\}\, dx. \tag{4}$$

Observe that $\psi_k(\theta)$ is finite for any $\theta \in \mathbb{R}^k$ as $b(\cdot)$ is bounded and $f(\cdot, \theta)$ has a compact support. Thus $\mathcal{M}_k$ is regular, which by definition means that the natural parameter set is open.

We assume throughout that the family $\mathcal{M}_k$ is of full rank, i.e. no linear combination $\sum_{j=1}^{k} \lambda_j b_j(X)$, for $(\lambda_1, \ldots, \lambda_k) \neq 0$, is constant with probability 1 for any $X \sim f \in \mathcal{M}_k$. This is equivalent to the condition that the covariance matrix $\mathrm{Cov}_\theta b(X)$ of vector $b(X)$ is positive definite for every $\theta \in \mathbb{R}^k$. Observe that families $\mathcal{M}_k$ are nested, i.e. $\mathcal{M}_k \subset \mathcal{M}_l$ for $k < l$ when the parameter space of $\mathcal{M}_k$ is embedded in $\mathbb{R}^l$ by appending $\theta \in \mathbb{R}^k$ with $l - k$ zeros.

Below we summarize some basic facts concerning maximum likelihood estimators (MLE) of the parameter θ in exponential family (cf. van der Vaart 2000, Sect. 4.2).

Let $X_1, \ldots, X_n$ be i.i.d. observations whose law belongs to a k-dimensional exponential family. Then for $\theta \in \mathbb{R}^k$ the log-likelihood function $\mathcal{L}_k(\theta)$ is of the form

$$\mathcal{L}_k(\theta) = \log \prod_{i=1}^{n} f(X_i, \theta) = n\{\theta \circ Y_n - \psi_k(\theta)\}, \tag{5}$$

where $Y_n = (1/n)\sum_{i=1}^n b(X_i)$. For $\theta \in \mathbb{R}^k$ define a vector-valued function

$$e_k(\theta) = (e_{k,1}(\theta), \ldots, e_{k,k}(\theta))^T = \mathbb{E}_\theta b(X),$$

where X is a random variable distributed according to density $f(\cdot, \theta) \in \mathcal{M}_k$.

In view of $\psi_k'(\theta) = e_k(\theta)$ and $\psi_k''(\theta) = \mathrm{Cov}_\theta b(X)$ (van der Vaart 2000, p. 38), we have

$$\begin{aligned} \mathcal{L}'_k(\theta) &= n[Y_n - e_k(\theta)], \\ \mathcal{L}''_k(\theta) &= -n\mathrm{Cov}_\theta b(X). \end{aligned} \tag{6}$$

For $k \in \{1, \ldots, K\}$, where $K \in \mathbb{N}$ is some fixed integer, define

$$\mathcal{L}_k = \sup_{\theta \in \mathbb{R}^k} \mathcal{L}_k(\theta), \tag{7}$$

and let $\hat{\theta} = \hat{\theta}_n^k \in \mathbb{R}^k$ (with the superscript indicating the dimension of the vector) be a MLE of the parameter $\theta = \theta^k \in \mathbb{R}^k$ in the k-dimensional exponential family $\mathcal{M}_k$. As $\hat{\theta}_n^k$ is a point at which the maximum of the log-likelihood $\mathcal{L}_k(\theta)$ is attained, it is the solution of the equation $\mathcal{L}'_k(\theta) = 0$, which in the exponential family is equivalent to the condition $Y_n = e_k(\theta)$. Moreover, we have

$$e_k'(\theta) = \mathrm{Cov}_\theta b(X), \tag{8}$$

where $e_k'(\theta) = \left(\frac{\partial e_{k,i}(\theta)}{\partial \theta_j}\right)_{1\le i,j\le k}$. Since $\mathcal{M}_k$ is assumed to be of full rank then e_k is one-to-one and if $\hat{\theta}_n^k$ exists then

$$\hat{\theta}_n^k = e_k^{-1}(Y_n). \tag{9}$$

For $j = 1, \ldots, k$ define a vector $a_j = [b_j(X_i)]_{1\le i\le n}$. We assume that $a_1, \ldots, a_k$ are linearly independent with probability 1 which is a sufficient condition for existence of MLE $\hat{\theta}_n^k$ (see Bogdan and Bogdan 2000). This condition also implies that the family $\mathcal{M}_k$ is of full rank so that $\hat{\theta}_n^k$ is an argument of the global maximum of the log-likelihood. Thus $\mathcal{L}_k(\hat{\theta}_n^k) = \mathcal{L}_k$. Moreover, if $\theta \in \mathbb{R}^k$ is the true value of the parameter, then we have $\hat{\theta}_n^k \xrightarrow{P} \theta$ and

$$\sqrt{n}(\hat{\theta}_n^k - \theta) \xrightarrow{\mathcal{D}} \mathcal{N}(0, [\mathrm{Cov}_\theta b(X)]^{-1}), \tag{10}$$

where $\xrightarrow{P}$ and $\xrightarrow{\mathcal{D}}$ denote convergence in probability and distribution, respectively.

2.2 Schwarz rule and post model selection estimators

In the preceding section we assumed that the density pertaining to the sample $X_1, X_2, \ldots, X_n$ belongs to an exponential family of known dimension k. If this is not the case, i.e. either an exponential model is misspecified or k is not known, we may first want to estimate the dimension of a model, which, in a sense specified below, describes well the sample $X_1, X_2, \ldots, X_n$, and then compute an estimator $\hat{\theta}$ of the parameter θ in a family with the dimension equal to the found value of the estimator of k. An estimator of θ constructed in this manner is known as a PMSE (see, e.g. Leeb and Pötscher 2006 for a general introduction to the subject).

One of the methods for estimating the dimension of a model is Schwarz's (1978) BIC. According to it we should choose the family $\mathcal{M}_k,\ k \in \{1, \ldots, K\}$, where K is some preassigned integer, for which the quantity

$$L_k = \mathcal{L}_k - \frac{1}{2} k \log n = n \sup_{\theta \in \mathbb{R}^k} \{\theta \circ Y_n - \psi_k(\theta)\} - \frac{1}{2} k \log n \tag{11}$$

is largest. Equivalently, the chosen dimension S of the model is given by

$$S = \min\{k : 1 \leq k \leq K,\ L_k \geq L_j,\ j = 1, \ldots, K\}.$$

The selection rule S has been applied in Ledwina (1994) in order to construct an adaptive version of Neyman's smooth goodness-of-fit test and intensively studied thereafter. In the context of estimation problems we found it advantageous to modify slightly its definition by adding the value 0 to the set of possible values. For this purpose we modify S in the following way. The symbol $\mathcal{M}_0$ will denote a model consisting of the uniform density on [0, 1]

$$\mathcal{M}_0 = \{f \in \mathcal{M}_1\ :\ \theta = 0\}.$$

Let $\mathcal{L}_0 = 0$. Observe that as $\mathcal{M}_0$ consists solely of the uniform density, $\mathcal{L}_0$ formally equals supremum of the log-likelihood for this model what is consistent with (7) for $k = 0$. Moreover, we extend the definition of L_k in (11) to $k = 0$ by letting $L_0 = 0$. We introduce estimator $\tilde{S}$ of dimension k of the model $\mathcal{M}_k,\ k \in \{0, 1, \ldots, K\}$ by

$$\tilde{S} = \min\{k : 0 \leq k \leq K,\ L_k \geq L_j,\ j = 0, \ldots, K\}. \tag{12}$$

Then the PMSE $\hat{\theta}$ of parameter θ is defined as follows

$$\hat{\theta} = \hat{\theta}^{\tilde{S}},$$

where $\hat{\theta}^{\tilde{S}}$ is the MLE of parameter θ in $\tilde{S}$-dimensional exponential family $\mathcal{M}_{\tilde{S}}$.

Note that if $\tilde{S} = 0$ then we infer that the random sample pertains to the uniform distribution. Moreover,

$$\tilde{S} = \begin{cases} S & \text{if } L_S > 0; \\ 0 & \text{otherwise.} \end{cases} \tag{13}$$

Let $KL(f, f_\theta) = \int \log(f/f_\theta) f$ be a Kullback–Leibler distance between f and f_θ. We note finally that as $\mathbb{E}_f \mathcal{L}_k(\theta)/n$ is equal $-KL(f, f_\theta) + \int \log(f) f$, maximization of L_k with respect to k can be viewed as choosing a model such that a penalized empirical KL distance between f and the model is smallest.

2.3 Estimators of Fisher information

Fisher information pertaining to the density f is defined as in (1) and assume throughout that I_f is finite. Consider the situation when $f(\cdot) = f(\cdot, \theta) \in \mathcal{M}_k,\ 1 \le k \le K$, where $\mathcal{M}_k$ is defined in (2). Then it is easy to see that for $X \sim f(\cdot, \theta)$

$$I_f = \int_0^1 \left\{ \sum_{j=1}^{k} \theta_j b_j'(x) \right\}^2 f(x, \theta)\, dx = \mathbb{E} \left\{ \sum_{j=1}^{k} \theta_j b_j'(X) \right\}^2 . \tag{14}$$

Thus in the case when k is assumed known, a natural estimator of I_f is

$$\hat{I}_f^k = \frac{1}{n} \sum_{i=1}^{n} \left\{ \sum_{j=1}^{k} \hat{\theta}_j b_j'(X_i) \right\}^2 = \frac{1}{n} \sum_{i=1}^{n} (\hat{\theta} \circ b'(X_i))^2, \tag{15}$$

where $X_1, X_2, \ldots, X_n$ is i.i.d. sample pertaining to $f(\cdot, \theta)$ and $\hat{\theta} = \hat{\theta}_n^k$ is a MLE of θ in $\mathcal{M}_k$ based on $X_1, X_2, \ldots, X_n$. Moreover, we define $\hat{I}_f^k = 0$ for $k = 0$. Note that $\hat{\theta}$ and $\hat{I}_f$ are estimated using the same sample.

In a general case, when no assumption about f belonging to one of parametric models $\mathcal{M}_k,\ 0 \le k \le K$, is made, we propose the following estimator of I_f

$$\hat{I}_f = \hat{I}_f^{\tilde{S}}, \tag{16}$$

i.e. $\hat{I}_f = \hat{I}_f^k$, when $\tilde{S} = k$. Thus we choose a parametric model $\mathcal{M}_k$ with $k = \tilde{S}$, which yields the best parsimonious fit to the data with respect to the modified Schwarz criterion (12), and then we estimate I_f in a chosen family $\mathcal{M}_{\tilde{S}}$ using (15). Thus $\hat{I}_f$ is an instance of PMSE discussed in Sect. 2.2.

Estimator (16) is the main object studied in this paper, in the next section we prove its consistency and asymptotic normality. In Sect. 4 we will compare small sample behaviour of $\hat{I}_f$ with two other estimators which are now briefly discussed.

Kernel estimator of I_f. An obvious class of estimators of I_f is defined as follows

$$\tilde{I}_f = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{\hat{f}'(X_i)}{\hat{f}(X_i)} \right)^2, \tag{17}$$

where $\hat{f}$ and $\hat{f}'$ are some estimators of the density f and its derivative f', respectively. In the following we will use as $\hat{f}$ a kernel estimator with boundary adjustment (see, e.g. Schuster 1985)

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - X_i) + K_h(x + X_i) + K_h(x + X_i - 2), \tag{18}$$

where $K_h(x) = h^{-1}K(x/h)$, K is a chosen probability density and $h = h_n$ is a sequence of bandwidths such that $h_n \to 0$ and $nh_n \to \infty$ as $n \to \infty$. Denote the bandwidth used for density estimation as $h_1 = h_{1,n}$. Moreover, f' is estimated by the derivative of $\hat{f}(x)$ with bandwidth $h_2 = h_{2,n}$. The considered choices of bandwidths are discussed in Sect. 4.1.

NN-estimator of I_f. The second estimator is based on an estimator of the optimal score function $J(y)$ defined as the derivative of $f(F^{-1}(y))$. It is easy to check that $I_f = \int_0^1 J^2(y)\,dy$. NN-estimator of $J(y)$ was introduced by Csörgő and Révész (1986) and is motivated as follows. Let k_n be a sequence of positive even integers and $X_{i:n}$ denotes ith order statistics in the X-sequence. Observe that $f(F^{-1}(u))$ may be estimated by $(k_n/n)(X_{i+k_n/2:n} - X_{i-k_n/2:n})^{-1}$ where $i = [nu]$. Moreover, $J(y)$ is approximated by

$$\int K_h(y-u)\,df(F^{-1}(u)) = \int K_h'(y-u)f(F^{-1}(u))\,du.$$

Plugging in the estimator of $f(F^{-1}(u))$ an letting $h_n = k_n/n$ we arrive at

$$\hat{J}_n(y) = \frac{1}{k_n}\sum_{i=k_n/2}^{n-k_n/2} K'\left(\frac{y - i/n}{h_n}\right)\frac{1}{(X_{i+k_n/2:n} - X_{i-k_n/2:n})}. \tag{19}$$

Finally, the estimator of I_f is defined as $n^{-1}\sum_{i=1}^{n}\hat{J}_n^2(i/n)$.

3 Main results

We first prove an auxiliary result on consistency of selection rule S. The result has been stated in Ledwina (1994) and the proof used Haughton's (1988) result for general k-dimensional exponential families. We give a direct proof here.

Lemma 1 *If $X_1, \ldots, X_n$ are i.i.d. with density $f(\cdot, \theta) \in \mathcal{M}_k$, where $1 \le k \le K$, and $\theta_k \neq 0$, then $\lim_{n\to\infty} P(S = k) = 1$.*

Proof First we show that $\lim_{n\to\infty} P(S < k) = 0$.

We have

$$P(S < k) = \sum_{l=1}^{k-1} P(S = l) \le \sum_{l=1}^{k-1} P(L_l \ge L_k).$$

Note that

$$\sup_{t\in\mathbb{R}^k} \{t\circ Y_n - \psi_k(t)\} \xrightarrow{P} \sup_{t\in\mathbb{R}^k} \{t\circ (e_{k,1}(\theta),\ldots,e_{k,k}(\theta)) - \psi_k(t)\} \tag{20}$$

and for $l \in \{1,\ldots,k-1\}$

$$\sup_{t\in\mathbb{R}^l} \{t\circ Y_n - \psi_l(t)\} \xrightarrow{P} \sup_{t\in\mathbb{R}^l} \{t\circ (e_{k,1}(\theta),\ldots,e_{k,l}(\theta)) - \psi_i(t)\}, \tag{21}$$

which follows from the fact that function $Y \mapsto \sup_{t\in\mathbb{R}^i} \{t\circ Y - \psi_l(t)\}$ is convex and hence continuous, $i = l, k$.

Moreover, we have for $(t,0) \in R^k$ denoting t appended with $k-l$ zero coordinates

$$\begin{aligned}\sup_{t\in\mathbb{R}^l}\{t\circ(e_{k,1}(\theta),\ldots,e_{k,l}(\theta))-\psi_l(t)\} &= \sup_{(t,0)\in\mathbb{R}^k}\{(t,0)\circ(e_{k,1}(\theta),\ldots,e_{k,k}(\theta))-\psi_k((t,0))\}\\ &< \sup_{t\in\mathbb{R}^k}\{t\circ(e_{k,1}(\theta),\ldots,e_{k,k}(\theta))-\psi_k(t)\}\end{aligned}$$

as the last supremum is uniquely attained at $t = \theta$ and $\theta_k \neq 0$.

From this inequality together with (20) and (21) we have

$$\sup_{t\in\mathbb{R}^k} \{t\circ Y_n - \psi_k(t)\} - \sup_{t\in\mathbb{R}^l} \{t\circ Y_n - \psi_l(t)\} - (k-l)\frac{\log n}{2n} \xrightarrow{P} a > 0,$$

where $a = \sup_{t\in\mathbb{R}^k} \{t\circ(e_{k,1}(\theta),\ldots,e_{k,k}(\theta)) - \psi_k(t)\} - \sup_{t\in\mathbb{R}^l}\{t\circ(e_{k,1}(\theta),\ldots,e_{k,l}(\theta)) - \psi_l(t)\}$.

Hence

$$P(L_l \geq L_k) = P\left(\sup_{t\in\mathbb{R}^k} \{t\circ Y_n - \psi_k(t)\} - \sup_{t\in\mathbb{R}^l} \{t\circ Y_n - \psi_l(t)\} - (k-l)\frac{\log n}{2n} \leq 0\right) \to 0$$

and $\lim_{n\to\infty} P(S < k) = 0$.

Now, for $l \in \{k+1,\ldots,K\}$ we have

$$P(S = l) \leq P(L_l > L_k) = P\left(\mathcal{L}_l - \mathcal{L}_k > \frac{1}{2}(l-k)\log n\right).$$

Note that $\mathcal{L}_l - \mathcal{L}_k \leq \mathcal{L}_l - \mathcal{L}_l(\theta^l)$, where $\theta^l \in \mathbb{R}^l$ equals θ appended with $l-k$ zero coordinates. Now the result follows from two-term Taylor expansion of $\mathcal{L}_l - \mathcal{L}_l(\theta^l)$, the fact that $\sqrt{n}(\hat{\theta}_n^l - \theta^l) = O_P(1)$ and continuity of $\mathcal{L}_l''(\cdot)$ which together imply that $\mathcal{L}_l - \mathcal{L}_l(\theta^l) = O_P(1)$. □

We prove now that selection rule $\tilde{S}$ is consistent.

Theorem 1 *Let $X_1, \ldots, X_n$ be i.i.d. with density $f(\cdot, \theta) \in \mathcal{M}_k$, $0 \le k \le K$. If $k \ge 1$ and $\theta_k \neq 0$ then* $\lim_{n\to\infty} P(\tilde{S} = k) = 1$. *If $\theta = 0$ then* $\lim_{n\to\infty} P(\tilde{S} = 0) = 1$.

Proof Assume first that $\theta_k \neq 0$. Then $\sup_{t\in\mathbb{R}^k}\{t \circ (e_{k,1}(\theta), \ldots, e_{k,k}(\theta)) - \psi_k(t)\} = \sup_{t\in\mathbb{R}^k} \mathbb{E}_\theta \log f(X_1, t) = \mathbb{E}_\theta \log f(X_1, \theta) > \mathbb{E}_\theta \log f(X_1, 0) = 0$ and arguing as in the proof of Lemma 1, we obtain that

$$\sup_{t\in\mathbb{R}^k}\{t \circ Y_n - \psi_k(t)\} - k \log n/(2n) \xrightarrow{P} \sup_{t\in\mathbb{R}^k}\{t \circ ((e_{k,1}(\theta), \ldots, e_{k,k}(\theta)) - \psi_k(t)\} > 0.$$

Hence $L_k \xrightarrow{P} \infty$ which in view of (13) and Lemma 1 implies that $\lim_{n\to\infty} P(\tilde{S} = k) = \lim_{n\to\infty} P(S = k) = 1$.

Now assume that $\theta = 0$. Then $\mathbb{E}_\theta b(X_1) = 0$ and similarly to the proof of Theorem 2 we can show that $L_l \xrightarrow{P} -\infty$ for $l \in \{1, \ldots, K\}$. Hence $\lim_{n\to\infty} P(L_l > 0) = 0$ for $l \in \{1, \ldots, K\}$ and $\lim_{n\to\infty} P(\tilde{S} = 0) = 1$. □

Theorem 2 *If $X_1, \ldots, X_n$ are i.i.d. with an arbitrary distribution P on* $[0, 1]$ *and* $\mathbb{E}_P b_1(X_1) = \cdots = \mathbb{E}_P b_{k-1}(X_1) = 0$, $\mathbb{E}_P b_k(X_1) \neq 0$ *for some $1 \le k \le K$, then* $\lim_{n\to\infty} P(\tilde{S} \ge k) = 1$.

Proof Kallenberg and Ledwina (1995, p. 1600) proved that the above assumptions imply $L_k \xrightarrow{P} \infty$. Thus to complete the proof it suffices to show that $L_l \xrightarrow{P} -\infty$ for $l \in \{1, \ldots, k-1\}$. However, as $b_1(X), \ldots, b_k(X)$ may be correlated, Prohorov's inequality, as suggested in the last reference, cannot be used, and we use Yurinskii's inequality instead.

Let $M > 0$ and $0 < \varepsilon < 2/3$. We have

$$P(L_l \ge -M) = P\left(\sup_{t\in\mathbb{R}^l}\{t \circ Y_n - \psi_l(t)\} \ge (l \log n - 2M)/(2n)\right).$$

Using Theorem 7.4 in Inglot and Ledwina (1996) we get for sufficiently large n

$$\begin{aligned} &P\left(\sup_{t\in\mathbb{R}^l}\{t \circ Y_n - \psi_l(t)\} \ge (l \log n - 2M)/(2n)\right) \\ &\quad \le P\left(\|Y_n\|^2 \ge (2-\varepsilon)(l \log n - 2M)/(2n)\right) \\ &\quad = P\left(\left\|\sum_{i=1}^n b(X_i)\right\| \ge [n(2-\varepsilon)(l \log n - 2M)/2]^{1/2}\right). \end{aligned}$$

Random vector $b(X_1) = (b_1(X_1), \ldots, b_l(X_1))^T$ satisfies Cramér's condition:

$$\mathbb{E}_P b(X_1) = 0, \quad \mathbb{E}_P \|b(X_1)\|^s \le \frac{s!}{2} b^2 \mathcal{K}^{s-2}, \quad s = 2, 3, \ldots$$

with $\mathcal{K} = \sqrt{l} V_l$ and $b^2 = l V_l^2$, where $V_l = \max_{1\le j\le l} \sup_{x\in[0,1]} |b_j(x)|$.

Thus we can apply Yurinskii's (1976) inequality (see also Inglot and Ledwina 1996, p. 2011):

$$P\left(\left\|\sum_{i=1}^{n} b(X_i)\right\| \geq [n(2-\varepsilon)(l\log n - 2M)/2]^{1/2}\right)$$
$$\leq 2\exp\left\{-\frac{x_n^2}{2}\left(1 + 1.62\frac{x_n\mathcal{K}}{B_n}\right)^{-1}\right\},$$

where $x_n = [(2-\varepsilon)(l\log n - 2M)/(2lV_l^2)]^{1/2}$ and $B_n = (nb^2)^{1/2} = V_l(nl)^{1/2}$.

Since $x_n \to \infty$ and $x_n/B_n \to 0$ as $n \to \infty$, we have $\lim_{n\to\infty} P(L_l \geq -M) = 0$. Thus $L_l \overset{P}{\to} -\infty$ as $n \to \infty$. □

Remark 1 Another possible method to choose the dimension of a model that is suggested in literature is the so-called simplified Schwarz selection rule which uses Neyman's smooth test statistic T_k defined below instead of the supremum $\mathcal{L}_k$ of log-likelihood function. Namely, let

$$T_k = \sum_{j=1}^{k}\left\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n} b_j(X_i)\right\}^2.$$

Then the selection rule introduced in Kallenberg and Ledwina (1997) is defined by

$$S2 = \min\{k : 1 \leq k \leq K,\ T_k - k\log n \geq T_j - j\log n,\ j = 1, \ldots, K\}.$$

An obvious advantage of $S2$ over S is its computational simplicity as calculation of MLE is avoided. In Ledwina (2000) it is shown that under assumptions of Theorem 2 $\lim_{n\to\infty} P(S2 \geq k) = 1$. Thus a similar criterion which is based on the augmented family of models including $\mathcal{M}_0$ can be defined as:

$$\tilde{S}2 = \min\{k : 0 \leq k \leq K,\ T_k - k\log n \geq T_j - j\log n,\ j = 0, \ldots, K\},$$

where $T_0 = 0$. If assumptions of Theorem 2 are satisfied then $\lim_{n\to\infty} P(\tilde{S}2 \geq k) = \lim_{n\to\infty} P(S2 \geq k) = 1$. When $X_1, \ldots, X_n$ are independent and uniformly distributed on $[0, 1]$ then $\lim_{n\to\infty} P(S2 = 1) = 1$ (see Inglot and Ledwina 2001, p. 814). This together with the fact that $T_1 \overset{\mathcal{D}}{\to} \chi_1^2$ yields $T_j - j\log n \overset{P}{\to} -\infty$ for $j = 1, 2, \ldots, K$. Hence $\lim_{n\to\infty} P(\tilde{S}2 = 0) = 1$. Note, however, that in the case of Fisher information we use MLE estimators for computing $\hat{I}_f$ and advantage of $\tilde{S}2$ over $\tilde{S}$ is lost.

Theorem 3 *If $X_1, \ldots, X_n$ are i.i.d. with density $f(\cdot, \theta) \in \mathcal{M}_k$, $0 \leq k \leq K$, and $\mathbb{E}(b'(X_1))^2 < \infty$, then $\hat{I}_f \overset{P}{\to} I_f$.*

Proof If $f \in \mathcal{M}_0$ then $P(\hat{I}_f^k = I_f) \geq P(\tilde{S} = 0) \to 1$ as $n \to \infty$ in view of Theorem 1. Let $\theta \in \mathbb{R}^k$, $\theta_k \neq 0$ and $\hat{\theta}_n^k = (\hat{\theta}_{1,n}^k, \ldots, \hat{\theta}_{k,n}^k)$. Note that $P(\hat{I}_f \to I_f) \geq P(\{\hat{I}_f^k \to I_f\} \cap \{\hat{S} = k\})$ and since $P(\tilde{S} = k) \to 1$ as $n \to \infty$ it suffices to show that $\hat{I}_f^k \xrightarrow{P} I_f$. We have

$$\hat{I}_f^k = \frac{1}{n}\sum_{i=1}^n (\hat{\theta}^k \circ b'(X_i))^2 = \frac{1}{n}\sum_{i=1}^n \left(\sum_{j=1}^k (\hat{\theta}_{j,n}^k - \theta_j) b_j'(X_i) + \sum_{j=1}^k \theta_j b_j'(X_i)\right)^2$$
$$= I_{n,1} + I_{n,2} + I_{n,3},$$

where

$$I_{n,1} = \frac{1}{n}\sum_{i=1}^n \left(\sum_{j=1}^k (\hat{\theta}_{j,n}^k - \theta_j) b_j'(X_i)\right)^2, \quad I_{n,2} = \frac{1}{n}\sum_{i=1}^n \left(\sum_{j=1}^k \theta_j b_j'(X_i)\right)^2,$$
$$I_{n,3} = \frac{2}{n}\sum_{i=1}^n \left[\left(\sum_{j=1}^k (\hat{\theta}_{j,n}^k - \theta_j) b_j'(X_i)\right)\left(\sum_{j=1}^k \theta_j b_j'(X_i)\right)\right].$$

By the weak law of large numbers we have $I_{n,2} \xrightarrow{P} I_f$. Using the Schwarz inequality we obtain

$$I_{n,1} \leq \frac{k}{n}\sum_{i=1}^n \sum_{j=1}^k (\hat{\theta}_{j,n}^k - \theta_j)^2 (b_j'(X_i))^2 = k\sum_{j=1}^k \left((\hat{\theta}_{j,n}^k - \theta_j)^2 \frac{1}{n}\sum_{i=1}^n (b_j'(X_i))^2\right).$$

We have $\hat{\theta}_{j,n}^k - \theta_j \xrightarrow{P} 0$ and by the weak law of large numbers $n^{-1}\sum_{i=1}^n (b_j'(X_i))^2 \xrightarrow{P} \mathbb{E}_\theta(b_j'(X))^2$. Hence $I_{n,1} \xrightarrow{P} 0$. By Schwarz inequality $I_{n,3} \leq 2(I_{n,1} \cdot I_{n,2})^{1/2}$ so $I_{n,3} \xrightarrow{P} 0$. Hence $\hat{I}_f^k \xrightarrow{P} I_f$. □

Theorem 4 *If $X, X_1, \ldots, X_n$ are i.i.d. with density $f(\cdot, \theta) \in \mathcal{M}_k$, $0 \leq k \leq K$, and $\mathbb{E}(b'(X_1))^4 < \infty$, then*

$$\sqrt{n}(\hat{I}_f - I_f) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma^2),$$

where

$$\sigma^2 = \sigma_1^2 + \sigma_2^2 + 4Cov_\theta(\theta^T B'\, \theta, b)(Cov_\theta b)^{-1} \Delta\, \theta, \tag{22}$$

with $B' = b'(X)(b'(X))^T$, $\Delta = \mathbb{E}_\theta(B')$, $\sigma_1^2 = 4\theta^T \Delta (Cov_\theta b(X))^{-1} \Delta\, \theta$, $\sigma_2^2 = Var_\theta(\theta^T B' \theta)$.

Proof Using the reasoning from the proof of Theorem 3, it is enough to prove the result for $\hat{I}_f^k$. Define B' as a $k \times k$ matrix consisting of all terms $b_i'(X)b_j'(X)$ where $i, j \in \{1, \dots, k\}$:

$$B' = B'(X) = [b_i'(X)b_j'(X)]_{1\le i,j\le k} = b'(X)(b'(X))^T$$

and let

$$\bar{B}' = \frac{1}{n}\sum_{i=1}^{n} B'(X_i).$$

$\hat{I}_f^k$ can be written as

$$\begin{aligned}\hat{I}_f^k &= \frac{1}{n}\sum_{i=1}^{n}(\hat{\theta}\circ b'(X_i))^2 = \hat{\theta}^T\,\frac{1}{n}\sum_{i=1}^{n} b'(X_i)(b'(X_i))^T\,\hat{\theta}\\ &= \hat{\theta}^T\bar{B}'\,\hat{\theta} = (\hat{\theta}^T - \theta^T)\bar{B}'(\hat{\theta}+\theta) + \theta^T\bar{B}'\,\theta\end{aligned}$$

with $\hat{\theta} = \hat{\theta}^k$. Thus

$$\sqrt{n}(\hat{I}_f - I_f) = \sqrt{n}(\hat{\theta}^T - \theta^T)\bar{B}'(\hat{\theta}+\theta) + \sqrt{n}(\check{I}_f - I_f), \tag{23}$$

where $\check{I}_f = \theta^T\bar{B}'\theta$. Note that since $\theta^T\bar{B}'\theta = n^{-1}\sum_{i=1}^{n}(\theta\circ b'(X_i))^2$, $\check{I}_f$ can be regarded as an estimator of I_f when θ is known. In view of (23) we have

$$\sqrt{n}(\hat{I}_f - I_f) = g(\sqrt{n}(\hat{\theta}-\theta), \sqrt{n}(\check{I}_f - I_f), \bar{B}'(\hat{\theta}+\theta)), \tag{24}$$

where $g(x, y, z) = x\circ z + y$ for $x, z \in \mathbb{R}^k$, $y \in \mathbb{R}$, is a continuous mapping. We will apply delta method (cf. van der Vaart 2000, p. 25) and (24) to prove the result.

By (10) and the weak law of large numbers, which implies $\bar{B}'(\hat{\theta}+\theta) \xrightarrow{P} 2\Delta\theta$, where $\Delta = \mathbb{E}_\theta(B')$, we have

$$\sqrt{n}(\hat{\theta}^T - \theta^T)\bar{B}'(\hat{\theta}+\theta) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma_1^2),$$

where $\sigma_1^2 = 4\theta^T\Delta(\mathrm{Cov}_\theta b(X))^{-1}\Delta\theta$. Since $\check{I}_f = n^{-1}\sum_{i=1}^{n}\theta^T B'(X_i)\,\theta$ then, by the Central Limit Theorem,

$$\sqrt{n}(\check{I}_f - I_f) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma_2^2),$$

where $\sigma_2^2 = Var(\theta^T B'\theta)$.

Define auxiliary variables

$$S = S(X) = \begin{bmatrix} b(X) \\ \theta^T B'\theta \end{bmatrix} \in \mathbb{R}^{k+1} \quad \text{and} \quad \bar{S} = \frac{1}{n}\sum_{i=1}^{n} S(X_i).$$

Then by the Central Limit Theorem

$$\sqrt{n}(\bar{S} - \mathbb{E}_\theta S) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \mathrm{Cov}_\theta S),$$

where $\mathrm{Cov}_\theta S$ is the covariance matrix of vector S:

$$\mathrm{Cov}_\theta S = \begin{bmatrix} \mathrm{Cov}_\theta b & \mathrm{Cov}_\theta(b, \theta^T B'\,\theta) \\ \mathrm{Cov}_\theta(\theta^T B'\,\theta, b) & \sigma_2^2 \end{bmatrix}.$$

Recall that $\hat{\theta}_n^k = e_k^{-1}(Y_n)$ (see 9) and define function $\nu : \mathbb{R}^{k+1} \to \mathbb{R}^{k+1}$ as follows:

$$\nu(u_1, \ldots, u_k, u_{k+1}) = [e^{-1}(u_1, \ldots, u_k)^T, u_{k+1}]^T.$$

In view of (8) the Jacobian matrix of function ν is equal to

$$\nabla\nu = \begin{bmatrix} (\mathrm{Cov}_\theta b)^{-1} & 0 \\ 0 & 1 \end{bmatrix}$$

By the delta method

$$\sqrt{n}(\nu(\bar{S}) - \nu(\mathbb{E}_\theta S)) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \Sigma_1),$$

where $\nu(\bar{S}) = (\hat{\theta}^T, \check{I}_f)^T,\ \nu(\mathbb{E}_\theta S) = (\theta^T, I_f)^T$ and

$$\begin{aligned} \Sigma_1 &= \nabla\nu\ \mathrm{Cov}_\theta S\ (\nabla\nu)^T \\ &= \begin{bmatrix} (\mathrm{Cov}_\theta b)^{-1} & (\mathrm{Cov}_\theta b)^{-1}\mathrm{Cov}_\theta(b, \theta^T B'\,\theta) \\ \mathrm{Cov}_\theta(\theta^T B'\,\theta, b)(\mathrm{Cov}_\theta b)^{-1} & \sigma_2^2 \end{bmatrix}. \end{aligned}$$

Thus a random vector $(\sqrt{n}(\hat{\theta} - \theta), \sqrt{n}(\check{I}_f - I_f), \bar{B}'(\hat{\theta} + \theta))^T$ converges in law to the distribution of $(Z, 2\Delta\,\theta)$, where $Z \sim \mathcal{N}(0, \Sigma_1)$. Hence in view of (24) $\sqrt{n}(\hat{I}_f - I_f) \xrightarrow{\mathcal{D}} g((Z, 2\Delta\,\theta))$, i.e.

$$\sqrt{n}(\hat{I}_f - I_f) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma^2),$$

where

$$\begin{aligned} \sigma^2 &= [2\theta^T\Delta, 1]\ \Sigma_1 \begin{bmatrix} 2\Delta\,\theta \\ 1 \end{bmatrix} \\ &= 4\theta^T\Delta(\mathrm{Cov}_\theta b(X))^{-1}\Delta\,\theta + 4\mathrm{Cov}_\theta(\theta^T B'\,\theta, b)(\mathrm{Cov}_\theta b)^{-1}\Delta\,\theta + Var(\theta^T B'\theta) \\ &= \sigma_1^2 + \sigma_2^2 + 4\mathrm{Cov}_\theta(\theta^T B'\,\theta, b)(\mathrm{Cov}_\theta b)^{-1}\Delta\,\theta. \end{aligned}$$

□

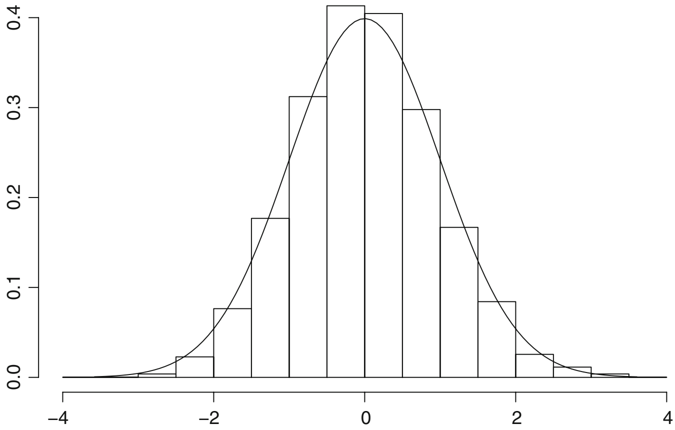


Fig. 1 Histogram of $\sqrt{n}(\hat{I}_f - I_f)/\sigma$ for $n = 3{,}000$ with $\mathcal{N}(0, 1)$ density overlaid. Density f belongs to $\mathcal{M}_1$ with $\theta_1 = 0.4$

Remark 2 Note that the asymptotic variance in (22) corresponds to decomposition in (23) with the term $4\mathrm{Cov}_\theta(\theta^T B'\, \theta, b)(\mathrm{Cov}_\theta b)^{-1}\Delta\theta$ being the asymptotic covariance of $\sqrt{n}(\hat{\theta}^T - \theta^T)\bar{B}'(\hat{\theta} + \theta)$ and $\sqrt{n}(\check{I}_f - I_f)$.

The asymptotic behaviour of estimator $\hat{I}_f = \hat{I}_f^{\tilde{S}}$ is exemplified in Fig. 1. We computed the value of $\hat{I}_f$ for 10^4 random samples $X_1, \ldots, X_n \sim f(\cdot, \theta) \in \mathcal{M}_k$ with $k = 1,\ \theta = 0.4$ and $n = 3{,}000$. In Fig. 1 the standard normal density function is compared to the histogram of $\sqrt{n}(\hat{I}_f - I_f)/\sigma$ where I_f and σ are theoretical values defined in (14) and (22), respectively. Visible conformity of the histogram and theoretical curve illustrates the result of Theorem 4.

Remark 3 Note that Theorems 1–4 remain valid when a term $\log n$ in the penalty is replaced with arbitrary k_n such that $k_n \to \infty$ and $k_n = o(n)$. In the following section we investigate by means of simulations influence of larger penalty on performance of $\hat{I}_f$.

We consider now a general case when underlying density f does not necessarily belong to an exponential family. Specifically, we assume that f has the following form

$$f(x) = \exp\left\{\sum_{j=1}^{\infty} \theta_j b_j(x) - \psi(\theta)\right\}, \tag{25}$$

where $\theta = (\theta_1, \theta_2, \ldots) \in \mathbb{R}^\infty$, $||\theta|| = (\sum_{j=1}^{\infty} |\theta_j|^2)^{1/2} < \infty$, and $\{1, b_1(x), b_2(x), \ldots\}$ form an orthonormal system in $(L^2([0, 1]), \lambda)$. Define $||a||_\infty = \sup_{x\in[0,1]} |a(x)|$ and let

$$V_m = \max_{j=1,\ldots,m} ||b_j||_\infty,$$
$$V'_m = \max_{j=1,\ldots,m} ||b'_j||_\infty.$$

Let $f_m(x) = \exp\{\sum_{j=1}^m \theta_j b_j(x) - \psi_m(\theta^m)\} \in \mathcal{M}_m$, where $\theta^m = (\theta_1, \ldots, \theta_m)^T$, be the approximation of f. We assume that for $m \to \infty$

$$||f - f_m||_{L^2} = O(m^{-r}) \quad \text{for some } \ r > 0. \tag{26}$$

Moreover, we impose the condition that

$$\sum_{j=1}^{\infty} |\theta_j| \, ||b_j||_\infty < \infty. \tag{27}$$

In the case when the density of a random sample is of the form (25) with the series $\sum_{j=1}^\infty \theta_j b'_j(x)$ converging uniformly in [0, 1] the Fisher information equals

$$I_f = \mathbb{E}_f(\rho(X))^2, \tag{28}$$

where $\rho(x) = f'(x)/f(x) = \sum_{j=1}^\infty \theta_j b'_j(x)$. In particular (28) is satisfied when

$$\sum_{j=1}^{\infty} |\theta_j| \, ||b'_j||_\infty < \infty. \tag{29}$$

Let $m = m(n)$ be a deterministic sequence such that $m(n) \to \infty$ when $n \to \infty$. The following result on consistency of $\hat{I}_n^m$ holds.

Theorem 5 *Assume that $V_m = O(m^{\omega_1})$, $V'_m = O(m^{\omega_2})$ for some $\omega_1, \omega_2 \geq 0$ and $r > \frac{1}{2} + \max(\omega_1, \omega_2)$. If $m = O(n^{1/k})$, where $k > 2 + 4\max(\omega_1, \omega_2)$ and conditions* (26)–(27) *and* (29) *hold, then $\hat{I}_n^m \xrightarrow{P} I_f$.*

In order to prove Theorem 5, we state two lemmas. Their proofs are given at the end of the section. The first lemma is a version of Lemma 5 in Barron and Sheu (1991).

Lemma 2 *Let $\theta_0 \in \mathbb{R}^m$, $\alpha_0 = \int b f_{\theta_0}$ and $\alpha \in \mathbb{R}^m$ be given. Let $c = e^{||\log f_{\theta_0}||_\infty}$. If*

$$||\alpha_0 - \alpha|| \leq \frac{1}{4ce\sqrt{m}V_m} \tag{30}$$

then the solution $\theta(\alpha)$ to $\int b f_\theta = \alpha$ exists and satisfies

$$||\theta_0 - \theta(\alpha)|| \leq 2ce||\alpha - \alpha_0||.$$

Moreover,

$$||\log f_{\theta_0}/f_{\theta(\alpha)}||_\infty \leq 4ce\sqrt{m}V_m||\alpha - \alpha_0|| \leq 1.$$

Let $\theta_m^* \in \mathbb{R}^m$ be the vector that satisfies the equation $\int bf_{\theta_m^*} = \int bf$. Then $f_m^* := f_{\theta_m^*}$ is called the information projection of f onto the exponential family $\mathcal{M}_m$.

In the next lemma properties of θ_m^* are used to establish consistency of MLE $\hat{\theta}_n^m$ when $m \to \infty$.

Lemma 3 *If $V_m = O(m^{\omega_1})$ for some $\omega_1 \geq 0$, $r > \frac{1}{2} + \omega_1$ and (26)–(27) are satisfied then for sufficiently large m the vector θ_m^* exists and*

$$||\theta_m^* - \theta^m|| = O(m^{-r}). \tag{31}$$

If additionally $m(n) = O(n^{1/k})$ and $k > 2 + 4\omega_1$ then the MLE $\hat{\theta}_n^m$ exists with probability tending to 1 as $n \to \infty$ and

$$||\hat{\theta}_n^m - \theta_m^*|| = O_P\left(\left(m^{1+2\omega_1}/n\right)^{1/2}\right). \tag{32}$$

Proof of Theorem 5 $|\hat{I}_n^m - I_f|$ is bounded by

$$\left|\frac{1}{n}\sum_{i=1}^{n}(\hat{\rho}_{n,m}(X_i))^2 - \int (\hat{\rho}_{n,m}(x))^2 f(x)\,dx\right|$$
$$+ \left|\int (\hat{\rho}_{n,m}(x))^2 f(x)\,dx - \int (\rho(x))^2 f(x)\,dx\right|,$$

where $\hat{\rho}_{n,m}(x) = \sum_{j=1}^{m}(\hat{\theta}_n^m)_j\ b_j'(x) = \hat{\theta}_n^m \circ b'(x)$. The first term of rhs can be written as $|(\hat{\theta}_n^m)^T(n^{-1}\sum_{i=1}^n B'(X_i) - \mathbb{E}B'(X))\hat{\theta}_n^m|$, where $B'(x) = b'(x)(b'(x))^T$, and whence is bounded by $||\hat{\theta}_n^m||^2||n^{-1}\sum_{i=1}^n Y^i||$, where $Y^i = (Y_{j,l}^i)_{1\leq j,l\leq m} \in \mathbb{R}^{m^2}$, $Y_{j,l}^i = b_j'(X_i)b_l'(X_i) - \int b_j'b_l'f$.

Yurinskii (1976) inequality implies that

$$P\left(\frac{1}{n}\left|\left|\sum_{i=1}^n Y^i\right|\right| > \varepsilon\right) \leq 2\exp\left\{-\frac{x^2}{2}\left(1 + 1.62\frac{x\kappa}{B_n}\right)\right\},$$

where $\kappa = 2m(V_m')^2$, $B_n = \sqrt{n}\kappa$, $x = n\varepsilon/B_n$. Whence for some positive constant C

$$x = \frac{\varepsilon}{2}\left(\frac{n}{m^2(V_m')^4}\right)^{1/2} \geq C\left(\frac{n}{m^{2+4\omega_2}}\right)^{1/2} \to \infty \quad \text{as } n \to \infty$$

and as $|x\kappa/B_n|$ is bounded it implies that $||n^{-1}\sum_{i=1}^n Y^i||$ converges to 0 in probability.

Moreover,

$$|\hat{\rho}_{n,m}(x) - \rho(x))| \leq ||\hat{\theta}_n^m - \theta^m||\sqrt{m}V_m' + \sum_{j=m+1}^{\infty}|\theta_j|||b'||_\infty$$

and using Lemma 3 we obtain

$$\begin{aligned}||\hat{\theta}_n^m - \theta^m||\sqrt{m}V'_m &\leq (||\theta^m - \theta_m^*|| + ||\hat{\theta}_n^m - \theta_m^*||)\sqrt{m}V'_m \\ &= O(m^{1/2+\omega_2-r}) + O_P\left(\left(m^{2+2\omega_1+2\omega_2}/n\right)^{1/2}\right),\end{aligned}$$

which implies that $\int(\hat{\rho}_{n,m}(x))^2 f(x)\,dx - \int(\rho(x))^2 f(x)\,dx$ converges to 0 in probability. □

Proof of Lemma 2 The proof proceeds similarly to the proof of Lemma 5 in Barron and Sheu (1991) with $q(x) = 1$ for $x \in [0, 1]$ and $\tau = 1$, the only change being that instead of using the result of Lemma 4 therein we apply the inequality

$$D(f_{\theta_1}||f_{\theta_2}) \geq \frac{1}{2}e^{-||\log q/f_{\theta_1}||_\infty}e^{-2||\theta_1-\theta_2||\sqrt{m}V_m}||\theta_1 - \theta_2||^2,$$

which follows from the fact that

$$||\log f_{\theta_1}/f_{\theta_2}||_\infty \leq 2||(\theta_1 - \theta_2)\circ b||_\infty \leq 2||\theta_1 - \theta_2||\sqrt{m}V_m$$

(cf. the proof of Lemma 4 in Barron and Sheu 1991). □

Proof of Lemma 3 To establish equality (31) we use Lemma 2 with $\alpha_0 = \int bf_m$, $\alpha = \int bf$ and $c = e^{||\log f_m||_\infty}$. Note that

$$||\log f_m||_\infty \leq ||\log f||_\infty + 2\sum_{j=1}^{\infty}|\theta_j|||b_j||_\infty \tag{33}$$

so $||\log f_m||_\infty$ is bounded in m. Since $\{b_j\}_{j=1}^\infty$ form an orthonormal system the Bessel inequality yields

$$||\int bf_m - \int bf|| = ||\int b(f_m - f)|| \leq ||f_m - f||_{L^2}.$$

Thus the fact that $m^{\frac{1}{2}+\omega_1}||f_m - f||_{L^2} \to 0$ as $m \to \infty$ implies (30) for sufficiently large m and hence the existence of θ_m^*. Moreover,

$$||\theta_m^* - \theta^m|| \leq 2ce||f_m - f||_{L^2},$$

which proves (31). It also follows from Lemma 2 that

$$||\log f_m^*/f_m||_\infty \leq 1. \tag{34}$$

For the equality (32) we apply Lemma 2 once again with $\alpha_0 = \int bf_{\theta_m^*}$, $\alpha = n^{-1}\sum_{i=1}^n b(X_i)$ and $c = e^{||\log f_m^*||_\infty}$. Using (34) and (33) we obtain that $||\log f_m^*||_\infty$ is bounded since $||\log f_m^*||_\infty \le ||\log f_m^*/f_m||_\infty + ||\log f_m||_\infty \le 1 + ||\log f_m||_\infty$. Yurinskii (1976) inequality yields

$$\left\|\frac{1}{n}\sum_{i=1}^{n} b(X_i) - \mathbb{E}b(X)\right\| = O_P\left(\left(m^{1+2\omega_1}/n\right)^{1/2}\right). \tag{35}$$

Thus $\sqrt{m}V_m||n^{-1}\sum_{i=1}^n b(X_i) - \mathbb{E}b(X)|| = O_P\left(\left(m^{2+4\omega_1}/n\right)^{1/2}\right)$ and the assumption (30) of Lemma 2 is satisfied with probability tending to 1 as $n \to \infty$. Whence

$$||\hat{\theta}_n^m - \theta_m^*|| \le 2ce\left\|\frac{1}{n}\sum_{i=1}^{n} b(X_i) - \mathbb{E}b(X)\right\|$$

and (32) follows. □

Remark 4 It is shown in Barron and Sheu (1991, pp. 1351–1352), that when $f \in W_2^r$, Sobolev space of functions on [0,1] such that $f^{(r-1)}$ is absolutely continuous with squared integrable derivative then f satisfies (26) when $\{b_j\}_{j=0}^\infty$ is the Legendre system. Moreover, in this case $V_m = (2m+1)^{1/2}$ and $V_m' = m(m+1)(2m+1)^{1/2}$, i.e. $\omega_1 = 1/2$, $\omega_2 = 5/2$, $k > 12$ and $f \in W_2^3$ in Theorem 5.

Consider a family of exponential models $\{\mathcal{M}_k\}$, $k \in \{1,\ldots,m(n)\}$ and corresponding $\tilde{S}$ defined as in (12) with K replaced by $m(n)$. The following corollary states consistency of $\hat{I}_n^{\tilde{S}}$. Note that its last assumption is in particular satisfied under conditions of Lemma 3 in view of (31).

Corollary 1 *Assume that $m(n)$ satisfies conditions of Theorem 5, there is an infinite number of nonzero θ_i in the representation (25), for sufficiently large m θ_m^* exists and $||\theta_m^* - \theta^m|| \to 0$ when $m \to \infty$. Then $\hat{I}_n^{\tilde{S}} \xrightarrow{P} I_f$.*

Proof The proof follows by examination of the proof of the Theorem 5, which indicates that its conclusion remains valid for a random sequence $m(n)$ provided $m(n) \xrightarrow{P} \infty$ and $m(n) = O_P(n^{1/k})$. In the case of $\tilde{S}$ the second condition is trivially satisfied as $\tilde{S} \le m$. To prove the first assertion observe that it is impossible that, starting from a certain m_0, all coordinates $(\theta_m^*)_l$ of θ_m^* for $m_0 \le l \le m$ and $m \ge m_0$ would be 0. Namely, considering $k \ge m_0$ such that $\theta_k \ne 0$ this would imply $||\theta_m^* - \theta_m|| \ge |\theta_k| \ne 0$ for all $m \ge m_0$ which contradicts the assumptions. Thus for every $M > 0$ there exists $m_1 \ge M$ such that $(\theta_m^*)_{m_2} \ne 0$ for some $m_2 \ge M$. Let $\tilde{m}_1$ be the smallest integer having this property. For $0 \le l < \tilde{m}_1$ we have

$$\begin{aligned}\sup_{t\in\mathbb{R}^l}(t\circ\mathbb{E}b - \psi_l(t)) &= \theta_l^*\circ\mathbb{E}b - \psi_l(\theta_l^*) < \theta_{\tilde{m}_1}^*\circ\mathbb{E}b - \psi_{\tilde{m}_1}(\theta_{\tilde{m}_1}^*)\\ &= \sup_{t\in\mathbb{R}^{\tilde{m}_1}}(t\circ\mathbb{E}b - \psi_{\tilde{m}_1}(t)).\end{aligned}$$

Reasoning as in the proof of Lemma 1 leads to

$$
\begin{aligned}
&P(\tilde{S} < M) \leq P(\tilde{S} < \tilde{m}_1) \leq \sum_{l=0}^{m_1-1} P(L_l \geq L_{\tilde{m}_1})\\
&\leq \sum_{l=0}^{m_1-1} P\left(\sup_{t\in\mathbb{R}^l}\{t\circ Y_n - \psi_l(t)\} - \sup_{t\in\mathbb{R}^{\tilde{m}_1}}\{t\circ Y_n - \psi_{\tilde{m}_1}(t)\} + (\tilde{m}_1 - l)\frac{\log n}{2n} \geq 0\right) \to 0
\end{aligned}
$$

and thus $\tilde{S} \xrightarrow{P} \infty$. □

4 Simulations

In this section we discuss results of a simulation study in which the performance of $\hat{I}_f = \hat{I}_f^{\tilde{S}}$ has been compared with performance of several competing estimators described in Sect. 2.3. In all numerical experiments first k Legendre polynomials on [0, 1] have been considered as the functions $b_1(x), \ldots, b_k(x)$. The chosen sample sizes have been equal to $n = 100$ and $n = 500$. The number K of dimensions considered in (11) depended on a sample size, namely $K(100) = 4$ and $K(500) = 10$. The number m of repetitions of each experiment in the study equaled 10^4.

4.1 Estimators under consideration

We consider two kernel-type estimators $\tilde{I}_f$ with a kernel function K equal to the density of the standard normal law $\mathcal{N}(0, 1)$ for various choices of bandwidths h_1 and h_2 used for estimating the density and its derivative, respectively. A kernel estimator of the derivative equals derivative of the kernel estimator of f defined in (18).

Let h_{SJ} be a bandwidth of a kernel density estimator computed using the method of Sheather and Jones (1991) which is a commonly used method of data-dependent bandwidth choice. We consider two estimators of I_f that use h_{SJ}: for the first one, called $\tilde{I}_{\mathrm{SJ}}$, we have taken $h_1 = h_2 = h_{\mathrm{SJ}}$. For the second one, we calculated a bandwidth $h_{2,\mathrm{opt}}$, which is the minimizer of the asymptotic Mean Integrated Squared Error of $\hat{f}'$ for normal $\mathcal{N}(\mu, \sigma^2)$ distribution and expressed it in terms of $h_{1,\mathrm{opt}}$, which is analogously defined asymptotically optimal bandwidth for $\hat{f}$. This yields

$$h_{2,\mathrm{opt}} = (3n/4)^{2/35}(3/5)^{1/7}h_{1,\mathrm{opt}}.$$

The estimator $\tilde{I}'_{\mathrm{SJ}}$ uses $h_1 = h_{\mathrm{SJ}}$ and h_2 calculated from the above formula when $h_{1,\mathrm{opt}}$ is replaced by h_{SJ}. Additionally, we define a ‘prophet’ kernel estimator $\tilde{I}_{\mathrm{best}}$ for which we choose $h_1 = h_2$ such that it minimizes an integrated squared error (ISE) of the corresponding estimator of f'/f. ISE is defined as $m^{-1}\sum_{i=1}^{m}[(\hat{f}'/\hat{f}(x_i)) - (f'/f)(x_i)]^2$, where (x_i) is a grid of m equidistant points in [0, 1]. Note that we use here the theoretical density of observations which is not available in practice.

We also consider 'prophet' NN estimator $\tilde{I}_{\text{NN}}^{\text{best}}$ with density of Beta(3,3) distribution on $[-1, 1]$ as a kernel function K and with the choice of parameter h_n that minimizes ISE of the estimated function $\hat{J}_n$ defined in (19). To the best of our knowledge no proposal of data-dependent bandwidth choice for NN estimator is available.

4.2 Employed distributions

The following distributions supported on [0, 1] has been considered in numerical experiments:

- the uniform distribution;
- truncated normal distribution with mean 0.5 and $\sigma = 0.25,\ 0.5,\ 0.75,\ 1$;
- Beta(a, b) distribution with $a = b = 3,\ 4,\ 5,\ 6,\ 7,\ 8,\ 9$;
- distribution $h4(a)$ with a density: $f_4^a(x) = a2^{a-1}(\min(x, 1-x))^{a-1}$ for $a = 3,\ 4,\ 5,\ 6,\ 7$;
- distribution $\text{l1}(\theta)$ belonging to $\mathcal{M}_1$ with a density: $f(x) = \exp(\theta b_1(x) - \psi_1(\theta))$

where $\theta = 0.4$ and $b_1(x) = \sqrt{3}(2x - 1)$ is the first Legendre polynomial on [0, 1].

We considered also two distributions with unbounded support:

- normal distribution with mean 0 and $\sigma = 0.25,\ 0.5,\ 0.75,\ 1$;
- exponential distribution with $\lambda = 0.5,\ 1,\ 1.5$.

In the last two cases, the following modification of the original definition of $\hat{I}_f$ has been made. The range $[X_{1:n}, X_{n:n}]$ of data is mapped onto [0, 1] by a linear transformation and $\hat{I}_f$ is calculated based on transformed data. Then it is multiplied by $(X_{n:n} - X_{1:n})^{-2}$ what corresponds to the change of Fisher information when the density is transformed back from [0, 1] to $[X_{1:n}, X_{n:n}]$.

4.3 Comparison of performance

We compared estimator $\hat{I}_f$ with kernel estimators $\tilde{I}_{\text{SJ}}$, $\tilde{I}'_{\text{SJ}}$ and their 'prophet' counterparts $\tilde{I}_{\text{best}}$, $\tilde{I}_{\text{NN}}^{\text{best}}$ for the families described above. The results are summarized in Tables 1, 2, 3, 4 and 5, where empirical means, standard deviations and mean squared errors based on $m = 10^4$ repetitions are given. Also means of empirical distributions of $\hat{I}_f$, $\tilde{I}_{\text{SJ}}$ and $\tilde{I}'_{\text{SJ}}$ are plotted in Fig. 2 as a function of pertaining parameter (for Beta distribution the range of parameter a is larger than that given in the corresponding table). The plots clearly indicate that in most displayed cases $\hat{I}_f$ is the least biased estimator among considered ones. It follows from the tables that when the distribution belongs to an exponential family (i.e. the case of the uniform, l1 and truncated normal distribution), $\hat{I}_f$ performed much better in terms of MSE than $\tilde{I}_{\text{SJ}}$ and $\tilde{I}'_{\text{SJ}}$. In particular, for the uniform distribution the ratios of MSEs are larger than 3 for $n = 100$. Surprisingly, $\hat{I}_f$ fares better than 'prophet' estimators $\tilde{I}_{\text{best}}$ and $\tilde{I}_{\text{NN}}^{\text{best}}$ in some cases, e.g. for l1 distribution with $n = 500$. The same conclusion is true in most cases for Beta and h4 distributions (the results not shown in the last case). The superiority of $\hat{I}_f$

Table 1 Performance of estimators for the uniform (left) and l1(0.4) (right) distributions

I_f	0			1.92		
	Mean	SD	MSE	Mean	SD	MSE
$n = 100,\ K = 4$						
$\hat{I}_f$	0.09	1.18	1.4	2.33	2.79	7.96
$\tilde{I}'_{\mathrm{SJ}}$	1.53	1.69	5.18	3.78	2.67	10.57
$\tilde{I}_{\mathrm{SJ}}$	2.49	2.63	13.1	5.27	3.9	26.46
$\tilde{I}_{\mathrm{best}}$	0.008	0.006	8.7e-05	0.72	0.53	1.7
$\tilde{I}^{\mathrm{best}}_{\mathrm{NN}}$	0.14	0.01	0.02	0.47	0.06	2.12
$n = 500,\ K = 10$						
$\hat{I}_f$	0.005	0.1	0.009	1.96	0.53	0.28
$\tilde{I}'_{\mathrm{SJ}}$	0.86	0.61	1.12	3.14	1.2	2.95
$\tilde{I}_{\mathrm{SJ}}$	1.87	1.23	5.03	4.97	2.16	13.96
$\tilde{I}_{\mathrm{best}}$	0.003	0.003	1.8e-05	0.66	0.24	1.64
$\tilde{I}^{\mathrm{best}}_{\mathrm{NN}}$	0.26	0.01	0.07	0.31	0.02	2.6

Table 2 Performance of estimators for Beta(a, a) distribution

a	3			4			5			6		
I_f	40			42			48			55		
	Mean	SD	MSE	Mean	SD	MSE	Mean	SD	MSE	Mean	SD	MSE
$n = 100,\ K = 4$												
$\hat{I}_f$	26.2	8.3	259	36.3	8.23	101	45.3	8.73	83	54.1	9.88	99
$\tilde{I}'_{\mathrm{SJ}}$	15.3	4.58	631	23.9	5.57	359	28	6.61	445	34.7	7.67	470
$\tilde{I}_{\mathrm{SJ}}$	17.7	5.6	529	26.6	6.71	281	30.8	8.25	363	38	9.59	380
$\tilde{I}_{\mathrm{best}}$	44.5	18	343	48	16.2	300	51.4	14.51	222	55.6	13.2	176
$\tilde{I}^{\mathrm{best}}_{\mathrm{NN}}$	6	5.25	605	22.1	6.48	439	28.1	7.72	458	33.8	9.17	534
$n = 500,\ K = 10$												
$\hat{I}_f$	28.1	5.74	175	36.6	4.58	50	44.7	4	27	52.7	4.09	22
$\tilde{I}'_{\mathrm{SJ}}$	18.5	2.07	468	26.1	2.38	258	33.7	2.74	212	39.7	3.27	246
$\tilde{I}_{\mathrm{SJ}}$	20.9	2.19	368	28.1	2.53	199	35.5	2.88	164	41.7	3.65	190
$\tilde{I}_{\mathrm{best}}$	45	1.93	168	47.3	10.86	146	50.2	9.7	99	54.2	8.94	81
$\tilde{I}^{\mathrm{best}}_{\mathrm{NN}}$	21.2	2.36	360	28.1	3.01	203	34.9	3.67	184	41.8	4.3	194

is more apparent for $n = 100$ and larger values of the parameter a in cases of Beta and h4 and parameter σ for truncated normal distributions. There is only one case, namely that of truncated normal distribution with $\sigma = 0.5$, when MSE of $\tilde{I}'_{\mathrm{SJ}}$ is smaller than MSE of $\hat{I}_f$.

'Prophet' methods $\tilde{I}_{\mathrm{best}}$ and $\tilde{I}^{\mathrm{best}}_{\mathrm{NN}}$ which availed themselves of the knowledge of true density performed best in some cases (e.g. for the truncated normal and the uniform

Table 3 Performance of estimators for truncated normal distribution on [0, 1] with mean 0.5 and standard deviation σ

σ	0.25			0.5			0.75			1		
I_f	12.38			1.164			0.248			0.081		
	Mean	SD	MSE	Mean	SD	MSE	Mean	SD	MSE	Mean	SD	MSE
$n = 100,\ K = 4$												
$\hat{I}_f$	13.3	4.81	24	0.58	2.31	5.66	0.13	1.23	1.53	0.12	1.29	1.68
$\tilde{I}'_{\mathrm{SJ}}$	7.75	3.46	33.38	1.76	1.67	3.15	1.46	1.54	3.85	1.48	1.6	4.51
$\tilde{I}_{\mathrm{SJ}}$	9.43	4.37	27.8	2.66	2.54	8.67	2.32	2.42	10.29	2.34	2.53	11.76
$\tilde{I}_{\mathrm{best}}$	7.93	2.22	24.77	0.47	0.35	0.61	0.05	0.07	0.04	0.017	0.002	0.004
$\tilde{I}^{\mathrm{best}}_{\mathrm{NN}}$	10.64	3.06	12.39	1.03	0.14	0.04	0.15	0.04	0.01	0.14	0.012	0.004
$n = 500,\ K = 10$												
$\hat{I}_f$	12.5	1.83	3.36	0.86	1.07	1.24	0.05	0.3	0.13	0.014	0.164	0.031
$\tilde{I}'_{\mathrm{SJ}}$	8.1	1.46	20.47	1.19	0.57	0.32	0.83	0.52	0.61	0.81	0.56	0.85
$\tilde{I}_{\mathrm{SJ}}$	9.65	1.74	10.46	1.97	1.01	1.67	1.68	1.04	3.13	1.73	1.13	3.98
$\tilde{I}_{\mathrm{best}}$	9.64	1.58	9.98	0.61	0.27	0.38	0.08	0.07	0.03	0.017	0.0008	0.004
$\tilde{I}^{\mathrm{best}}_{\mathrm{NN}}$	14.2	1.67	6.1	1.01	0.06	0.03	0.28	0.01	0.001	0.27	0.01	0.04

Table 4 Performance of estimators for normal distribution $\mathcal{N}(0, \sigma)$

σ	0.25			0.5			0.75			1		
I_f	16			4			1.778			1		
	Mean	SD	MSE	Mean	SD	MSE	Mean	SD	MSE	Mean	SD	MSE
$n = 100,\ K = 4$												
$\hat{I}_f$	13	3.16	19.18	3.23	0.86	1.33	1.42	0.377	0.27	0.8	0.219	0.089
$\tilde{I}'_{\mathrm{SJ}}$	11.4	2.37	26.47	2.19	0.69	3.75	0.75	0.234	1.11	0.43	0.108	0.337
$\tilde{I}_{\mathrm{SJ}}$	12.1	2.77	22.86	2.4	0.79	3.2	0.82	0.282	1	0.46	0.132	0.308
$\tilde{I}_{\mathrm{best}}$	10.1	1.32	37.04	2.2	0.46	3.46	0.67	0.141	1.25	0.38	0.065	0.385
$\tilde{I}^{\mathrm{best}}_{\mathrm{NN}}$	11.8	2.78	25.74	2.95	0.7	1.59	1.33	0.315	0.3	0.75	0.174	0.094
$n = 500,\ K = 10$												
$\hat{I}_f$	15.2	1.12	1.91	3.79	0.29	0.13	1.68	0.131	0.03	0.95	0.072	0.008
$\tilde{I}'_{\mathrm{SJ}}$	13.1	1.02	9.39	2.45	0.31	2.5	0.82	0.102	0.92	0.47	0.045	0.283
$\tilde{I}_{\mathrm{SJ}}$	13.5	1.12	7.46	2.62	0.33	2.02	0.87	0.115	0.84	0.49	0.051	0.266
$\tilde{I}_{\mathrm{best}}$	12.1	1.04	16.5	1.65	0.27	5.57	0.68	0.071	1.22	0.44	0.043	0.315
$\tilde{I}^{\mathrm{best}}_{\mathrm{NN}}$	13.9	1.42	6.43	3.47	0.35	0.41	1.54	0.155	0.08	0.87	0.088	0.026

distributions) but in general they performed on par or even worse than $\hat{I}_f$. One of the reasons is possibly that the criterion which has been used to find a bandwidth for $\tilde{I}_{\mathrm{best}}$ and $\tilde{I}^{\mathrm{best}}_{\mathrm{NN}}$ involved f'/f and not directly I_f, the other is possible superior-

Table 5 Performance of estimators for exponential distribution $\mathcal{E}(\lambda)$

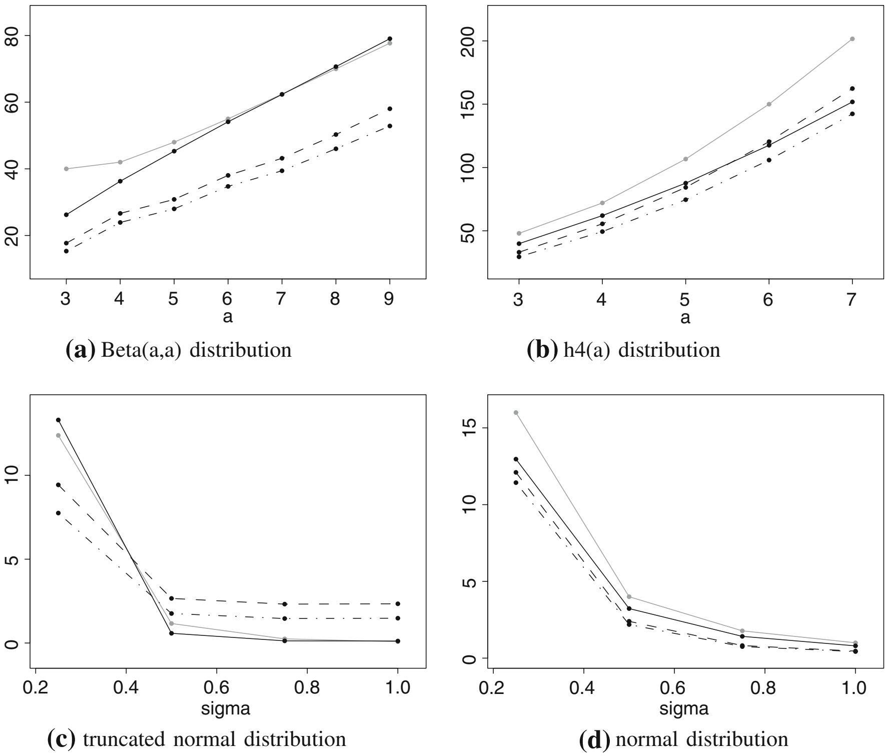
λ	0.5			1			1.5		
I_f	0.25			1			2.25		
	Mean	SD	MSE	Mean	SD	MSE	Mean	SD	MSE
$n = 100,\ K = 4$									
$\hat{I}_f$	0.256	0.114	0.013	1.021	0.474	0.225	2.28	0.9	0.81
$\tilde{I}'_{\mathrm{SJ}}$	0.474	0.160	0.076	1.473	0.674	0.678	2.782	1.603	2.851
$\tilde{I}_{\mathrm{SJ}}$	0.549	0.202	0.13	1.651	0.79	1.048	3.084	1.79	3.901
$\tilde{I}_{\mathrm{best}}$	0.166	0.018	0.007	0.312	0.174	0.503	1.227	0.563	1.362
$\tilde{I}^{\mathrm{best}}_{\mathrm{NN}}$	0.287	0.117	0.015	1.19	0.358	0.165	2.494	0.767	0.648
$n = 500,\ K = 10$									
$\hat{I}_f$	0.244	0.025	0.0007	0.971	0.098	0.01	2.187	0.222	0.054
$\tilde{I}'_{\mathrm{SJ}}$	0.657	0.129	0.182	2.425	0.583	2.372	5.121	1.516	10.543
$\tilde{I}_{\mathrm{SJ}}$	0.831	0.182	0.371	2.773	0.694	3.625	5.021	1.524	10
$\tilde{I}_{\mathrm{best}}$	0.17	0.008	0.0063	0.315	0.072	0.474	1.68	0.496	0.571
$\tilde{I}^{\mathrm{best}}_{\mathrm{NN}}$	0.278	0.045	0.0027	1.076	0.166	0.033	2.396	0.36	0.151

ity of log-linear PMS estimator which avoids a problem of choosing two different bandwidths.

In the case of two densities supported on unbounded sets estimator $\hat{I}_f$ was always less biased and had much smaller MSE than kernel estimators. For normal distribution this was true even in comparison with ‘prophet’ estimators. Also, for exponential distribution estimator $\hat{I}_f$ had in all cases smaller standard deviation than kernel estimators. Thus the proposed method seems promising also for the cases when the support of f is unknown and possibly unbounded.

4.4 A choice of penalty and comparison of $\hat{I}_f^{\tilde{S}}$ and $\hat{I}_f^{S}$.

We also studied the effect of adding family $\mathcal{M}_0$ to the list and changing a penalty in BIC criterion. Namely, we compared MSEs for $\hat{I}_f = \hat{I}_f^{\tilde{S}}$ and $\hat{I}_f^{S}$ when the penalty $k \log n/2$ in (11) is multiplied by C. Obviously, $C = 1$ corresponds to the original criterion. Due to space constraints only selected results for the uniform and the truncated normal distribution are shown for $C = 1$, 2 and $n = 100$ in Tables 6 and 7, respectively. In general, for both estimators taking $C > 2$ does not lead to decrease of MSE and in some cases (e.g. for truncated normal with $\sigma = 0.25$ or Beta(3, 3) for $n = 100$ in case of $\hat{I}_f$) causes its substantial increase. This also happens in the case of 11 distribution when adoption of a larger penalty results in a larger probability of $k = 0$, i.e. of wrong decision, in the case of $\hat{I}_f^{\tilde{S}}$, whereas in the case of $\hat{I}_f^{S}$ probability of $k = 1$ (correct decision) becomes larger simply because model $\mathcal{M}_0$ is excluded from the list of models. It turned out that the choice of a large penalty ($C \geq 4$) in the case of the uniform and the truncated normal distribution resulted in choice of $k = 0$ for *all*



(a) Beta(a,a) distribution

(b) h4(a) distribution

(c) truncated normal distribution

(d) normal distribution

Fig. 2 Empirical means of estimators for selected parametric distributions and for $n = 100$. *Solid line* $\hat{I}_f$, *dashed line* $\tilde{I}_{\mathrm{SJ}}$, *dotted-dashed line* $\tilde{I}'_{\mathrm{SJ}}$. *Grey line* I_f

Table 6 An influence of penalty on $\hat{I}_f^S$ (left) and $\hat{I}_f^{\tilde{S}}$ (right) for uniform distribution on [0, 1] for $n = 100$, $K = 4$

I_f	0			0		
	Mean	SD	MSE	Mean	SD	MSE
$C = 1$	0.391	2.400	5.911	0.091	1.178	1.395
$C = 2$	0.146	0.572	0.349	0.002	0.056	0.003

10^4 repetitions. $\hat{I}_f^{\tilde{S}}$ for $C = 2$ performs better than $\hat{I}_f^S$ for the uniform distribution and the truncated normal in the case $\sigma \geq 0.5$, i.e. in cases when the distribution is close to the uniform. For beta, h4 and exponential distributions both estimators performed exactly the same. For l1 and normal distributions $\hat{I}_f^{\tilde{S}}$ performed worse. $\hat{I}_f^{\tilde{S}}$ with $C = 2$ performed in general better than for $C = 1$. Thus it should be kept in mind that changing C from 1 to 2 for comparison with other estimators discussed above will underline superiority of $\hat{I}_f^{\tilde{S}}$ when compared with kernel estimators even more.

Table 7 An influence of penalty on $\hat{I}^S$ and $\hat{I}^{\tilde{S}}$ for truncated normal distribution on [0, 1] with mean 0.5 and standard deviation σ for $n = 100,\ K = 4$

σ	0.25			0.5			0.75			1		
I_f	12.38			1.164			0.248			0.081		
	Mean	SD	MSE	Mean	SD	MSE	Mean	SD	MSE	Mean	SD	MSE
$\hat{I}^S$												
$C = 1$	13.33	4.74	23.4	1.36	2.92	8.55	0.55	2.41	5.89	0.39	1.88	3.63
$C = 2$	12.96	4.25	18.4	0.48	1.52	2.79	0.17	0.58	0.35	0.14	0.52	0.27
$\hat{I}^{\tilde{S}}$												
$C = 1$	13.3	4.81	24	0.58	2.31	5.66	0.13	1.23	1.53	0.12	1.29	1.68
$C = 2$	11.8	6.16	38.3	0.03	0.51	1.56	0.006	0.18	0.09	0.004	0.15	0.03

Acknowledgments We thank T. Ledwina for suggesting the subject of this paper, insightful comments and supplying us with a computer program for calculating ML estimators in exponential families created under her Ministry of National Education Grant 341 046.

References

Barron AR, Sheu CH (1991) Approximation of density functions by sequences of exponential families. Ann Stat 19:1347–1369

Bickel P, Klassen C, Ritov Y, Wellner J (1998) Semiparametrics. In: Kotz S, Read C, Banks D (eds) Encyclopedia of statistical sciences, vol 2. Wiley, New York, pp 602–614

Bogdan K, Bogdan M (2000) On existence of maximum likelihood estimators in exponential families. Statistics 34:137–149

Csörgő M, Révész P (1986) A nearest neighbour-estimator for the score function. Prob Theory Relat Fields 71:293–305

Haughton DMA (1988) On the choice of a model to fit data from an exponential family. Ann Stat 16:342–355

Inglot T, Ledwina T (1996) Asymptotic optimality of data-driven Neyman's tests for uniformity. Ann Stat 24:1982–2019

Inglot T, Ledwina T (2001) Intermediate approach to comparison of some goodness-of-fit tests. Ann Inst Stat Math 53:810–834

Kallenberg WCM, Ledwina T (1995) Consistency and Monte Carlo simulation of a data driven version of smooth goodness-of-fit tests. Ann Stat 23:1594–1608

Kallenberg WCM, Ledwina T (1997) Data driven smooth test when the hypothesis is composite. J Am Stat Assoc 92:1094–1104

Ledwina T (1994) Data-driven version of Neyman's smooth test of fit. J Am Stat Assoc 89:1000–1005

Ledwina T (2000) Limiting behaviour of data driven Neyman's statistic under fixed alternatives. Manuscript

Leeb H, Pötscher BM (2006) Model Selection, manuscript. http://www.stat.yale.edu/~hl284/handbook.pdf

Schuster EF (1985) Incorportating support constraints into nonparametric estimation of densities. Commun Stat Theory Methods 14:1123–1136

Schwarz G (1978) Estimating the dimension of a model. Ann Stat 6:461–464

Sheather SJ, Jones MC (1991) A reliable data-based bandwidth selection method for kernel density estimation. J R Stat Soc Ser B 53:683–690

Stone CJ (1975) Adaptive maximum likelihood estimator of a location parameter. Ann Stat 3:267–284

van der Vaart AW (2000) Asymptotic statistics. Cambridge University Press, Cambridge

Yurinskii VV (1976) Exponential inequalities for sums of random vectors. J Mult Anal 6:473–499