***P*-value model selection criteria for exponential families of increasing dimension**

Jan Mielniczuk · Małgorzata Wojtyś

Received: 16 September 2011 / Published online: 21 March 2013
© The Author(s) 2013. This article is published with open access at Springerlink.com

Abstract Let $\mathcal{M}_{\underline{l}}$ be an exponential family of densities on $[0, 1]$ pertaining to a vector of orthonormal functions $b_{\underline{l}} = (b_{l_1}(x), \dots, b_{l_p}(x))^T$ and consider a problem of estimating a density f belonging to such family for unknown set $\underline{l} \subset \{1, 2, \dots, m\}$, based on a random sample $X_1, \dots, X_n$. Pokarowski and Mielniczuk (2011) introduced model selection criteria in a general setting based on p -values of likelihood ratio statistic for $H_0 : f \in \mathcal{M}_0$ versus $H_1 : f \in \mathcal{M}_{\underline{l}} \setminus \mathcal{M}_0$, where $\mathcal{M}_0$ is the minimal model. In the paper we study consistency of these model selection criteria when the number of the models is allowed to increase with a sample size and f ultimately belongs to one of them. The results are then generalized to the case when the logarithm of f has infinite expansion with respect to $(b_i(\cdot))_1^\infty$. Moreover, it is shown how the results can be applied to study convergence rates of ensuing post-model-selection estimators of the density with respect to Kullback–Leibler distance. We also present results of simulation study comparing small sample performance of the discussed selection criteria and the post-model-selection estimators with analogous entities based on Schwarz’s rule as well as their greedy counterparts.

Keywords Density estimation · Exponential family · Information projection · Likelihood ratio test · Model selection · p -value criterion

The research was supported by Polish Science and Higher Education Ministry grant no. N201 526138 and the European Union in the framework of European Social Fund through the Warsaw University of Technology Development Programme.

J. Mielniczuk · M. Wojtyś (✉)
Faculty of Mathematics and Information Science, Warsaw University of Technology,
ul. Koszykowa 75, Warsaw 00-662, Poland
e-mail: M.Wojtys@mini.pw.edu.pl

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

1 Introduction

Consider the following general scenario of a model selection. Let $\{\mathcal{M}_{\underline{i}}\}_{\underline{i} \in I}$ be a list of models, where $I \subset 2^{\text{full}}$ and $\text{full} = \{1, 2, \dots, m\}$, consisting of densities described by finite-dimensional parameter $\theta^{\underline{i}}$, such that for every $\underline{i} \in I$ we have $\mathcal{M}_0 \subseteq \mathcal{M}_{\underline{i}} \subseteq \mathcal{M}_{\text{full}}$ for some minimal model $\mathcal{M}_0$. In the case of a correct specification of the list of models the density f from which iid sample of size n is available belongs to a model $\mathcal{M}_{\underline{i}}$ for some $\underline{i} \in I$ of cardinality $|\underline{i}|$ such that $f \in \mathcal{M}_{\underline{i}}$ implies $\mathcal{M}_{\underline{i}} \subseteq \mathcal{M}_{\underline{j}}$. Throughout, $|\underline{i}|$ will denote cardinality of set $\underline{i}$. Pokarowski and Mielniczuk (2011) considered general parametric models $\mathcal{M}_{\underline{i}}$ and introduced (unscaled) minimal p -value criterion M_m^n based on asymptotic p -values of likelihood ratio test statistics $\Lambda_{n,0,\underline{i}}$ for testing $H_0 : f \in \mathcal{M}_0$ versus $H_1 : f \in \mathcal{M}_{\underline{i}} \setminus \mathcal{M}_0$. It is proved there that under mild conditions on the regularity of the models and for a general lists of parametric models of fixed size m which does not depend on the sample size the introduced minimal p -value criterion (mPVC) is consistent in the sense considered for selection rules, i.e. that $P(M_m^n = \underline{i}) \rightarrow 1$. Moreover, it is shown that Bayesian Information Criterion BIC (cf. Schwarz 1978) is an approximation of mPVC.

In the present paper we focus on a special family of models, namely exponential models and generalize and strengthen these consistency results to the case of possibly growing families and allow for misspecification of the true density. We also consider maximal p -value criterion (MPVC) as well as greedy versions of both mPVC and MPVC.

We note that permitting that the size of the list of models grows together with the sample size is of interest as it makes possible to handle nonparametric situations, e.g. the case when logarithm of the underlying density has infinite orthonormal expansion. Therefore, beside results on consistency of introduced rules and on Kullback–Leibler distance from the ensuing post-model-selection estimator to the true density when it belongs to one of the underlying models, we were able to prove similar results in the case when the list of models is misspecified. In particular, Theorem 3 states that the selector based on the minimal p -value criterion is then conservative in a sense defined in Sect. 3.1.

BIC-based selection of an appropriate exponential model was applied by Ledwina (1994) to construct data-adaptive Neyman smooth tests. It turned out to be a powerful tool in many testing problems (cf. e.g. Inglot and Ledwina 1996; Kallenberg and Ledwina 1997). In the present paper we investigate usefulness of such approach in a parallel problem of estimation. In particular, in Sect. 4 we compare numerically performance of post-model-selection estimators based on p -values with those based on BIC. It is important to stress that unlike in testing problems, which usually consider nested family of models, we deal here with an arbitrary family of exponential models.

The presented approach can lead also to effective methods of estimating functionals of density which compare favorably with kernel plug-in method. In particular, Mielniczuk and Wojtyś (2010) investigated this problem in the case of Fisher information with BIC applied as a selection criterion, and a similar method may be used also to approximate entropy of a density, which is usually estimated with the use of a kernel estimate truncated from below.

Moreover, let us note that the construction of selection criteria using p -values is quite general and can be applied in other scenarios. In particular, an analogous approach may be used to choose the most parsimonious linear model from a growing list of linear models provided that a design matrix satisfies mild regularity conditions.

Organization of the paper is as follows. In Sect. 2 we introduce exponential families of densities, define p -value criteria and state auxiliary lemmas. Section 3 contains main results on consistency and conservativeness of mPVC and MPVC criteria as well as their greedy counterparts. The results concern both the situation when an underlying density belongs to one of the exponential models considered (correct specification of family of models) and the opposite misspecification case. We also show how these results can be applied to bound convergence rates of ensuing post-model-selection estimators of density with respect to Kullback–Leibler distance. In Sect. 4 behavior of post-model-selection estimators is studied by means of numerical experiments.

2 Model selection procedures based on p -values in exponential families of densities and auxiliary results

2.1 Minimal and maximal p -value selection criteria

We specify $\{\mathcal{M}_{\underline{i}}\}_{\underline{i} \in I}$ to be a list of exponential models. Let $B = (b_j(x))_{j=0}^\infty$ be some orthonormal system in $L^2([0, 1], \lambda)$, where λ is the Lebesgue measure on $[0, 1]$ and $b_0 \equiv 1$. Observe that equality $\sum_{j=0}^k a_j b_j(x) = 0$ λ -a.e. for some constants $(a_j)_{j=0}^k$ and $k \in \mathbb{N}$ implies $a_j = 0$ for $j = 0, 1, \dots, k$. For fixed $m \in \mathbb{N}$ and every nonempty subset $\underline{i}$ of $\{1, \dots, m\}$ define regular exponential family of densities on $[0, 1]$ of the form

$$\mathcal{M}_{\underline{i}} = \left\{ f_{\theta}(x) = \exp \left\{ \sum_{j \in \underline{i}} \theta_j b_j(x) - \psi_{\underline{i}}(\theta) \right\} : \theta_j \in \mathbb{R} \text{ for } j \in \underline{i} \right\}, \quad (1)$$

where $\theta = (\theta_j)_{j \in \underline{i}}$, $\psi_{\underline{i}}(\theta) = \log \int_0^1 \exp\{\sum_{j \in \underline{i}} \theta_j b_j(x)\} dx$ is normalizing constant corresponding to model $\mathcal{M}_{\underline{i}}$ under consideration. Let $\mathbb{1}(\cdot)$ be an indicator function and $\mathcal{M}_0 = \{f_0(x) = \mathbb{1}(x \in [0, 1])\}$ the smallest (null) model consisting of the uniform density only. We refer to [van der Vaart \(2000\)](#), Sect. 4.2, and [Wainwright and Jordan \(2008\)](#) for an introduction to exponential families.

Let $f \in L^2([0, 1], \lambda)$ be some unknown density on $[0, 1]$ and consider the problem of choosing one of the above 2^m models in order to estimate f with the use of a random sample $X_1, \dots, X_n \sim f$. We consider the following model selection criteria which are proposed in [Pokarowski and Mielniczuk \(2011\)](#) and are based on the idea of likelihood ratio goodness of fit tests. Namely, for a fixed $m \in \mathbb{N}$ and for every nonempty $\underline{i} \subset \{1, \dots, m\}$ consider testing null hypothesis:

$$H_0 : f \in \mathcal{M}_0$$

versus

$$H_1 : f \in \mathcal{M}_{\underline{i}} \setminus \mathcal{M}_0$$

with test statistic

$$\Lambda_{n,0,\underline{i}} = 2 \log \prod_{j=1}^n f_{\hat{\theta}_{\text{ML}}^{\underline{i}}}(X_j), \quad (2)$$

calculated for the random sample $X_1, \dots, X_n \sim f$, where $\hat{\theta}_{\text{ML}}^{\underline{i}}$ is the maximum likelihood estimator of parameter θ in family $\mathcal{M}_{\underline{i}}$ based on $X_1, \dots, X_n$. Note that since for $\underline{i} = \emptyset$ estimator $f_{\hat{\theta}_{\text{ML}}^{\underline{i}}}$ is equal to $f_0(x) = \mathbb{1}(x \in [0, 1])$, the statistic $\Lambda_{n,0,\underline{i}}$ is the likelihood ratio test statistic (LRT) for the above testing problem.

In order to compute p -value of (2) knowledge of the distribution of likelihood ratio statistic under the null hypothesis is needed. Since its exact distribution is not known in the case of the exponential family (1), we use its asymptotically valid approximation by chi-squared distribution. Namely Wilks' theorem (cf. e.g. Theorem 5.6.3 in [Sen and Singer 1993](#)) implies that likelihood ratio statistic defined as in (2) has an asymptotic $\chi_{|\underline{i}|}^2$ distribution provided $H_0 : f \in \mathcal{M}_0$ holds.

For $x \in \mathbb{R}$ and $k \in \mathbb{N}$ denote

$$p(x|k) = 1 - F_k(x), \quad (3)$$

where F_k is the cumulative distribution function of χ_k^2 distribution. We consider $p(\Lambda_{n,0,\underline{i}} | |\underline{i}|)$ as an approximate p -value of the test statistic (2), i.e. the conditional probability that, given $\Lambda_{n,0,\underline{i}}$, a random variable having $\chi_{|\underline{i}|}^2$ distribution exceeds $\Lambda_{n,0,\underline{i}}$. Observe that if G_n is c.d.f. of $\Lambda_{n,0,\underline{i}}$ then c.d.f. of random variable $p(\Lambda_{n,0,\underline{i}} | |\underline{i}|)$ equals $1 - G_n(F_{|\underline{i}|}^{-1}(1-x))$ and if H_0 holds is approximately c.d.f. of the uniform distribution. As the best fitted model for f we choose the one having the smallest scaled p -value:

$$\hat{\underline{i}}_{\text{mPVC}} = \arg \min_{\underline{i} \subset \{1, \dots, m\}} p(\Lambda_{n,0,\underline{i}} | |\underline{i}|) e^{a_n |\underline{i}|},$$

where $a_n \geq 0$, $a_n/n \rightarrow 0$ as $n \rightarrow \infty$ and we put $p(\Lambda_{n,0,0} | 0) := n^{-1/2}$. In the case of ties the model $\mathcal{M}_{\underline{i}}$ having the smallest number of parameters $|\underline{i}|$ is chosen. Model selection criterion based on choosing the smallest p -value is introduced in [Pokarowski and Mielniczuk \(2011\)](#) where it is called **minimal p -value criterion** (mPVC). Its original definition considered only the case $a_n = 0$. In this case from among pairs $\{H_0, H_{\underline{i}}\}$ we choose the pair for which we are most inclined to reject H_0 , i.e. we select a model corresponding to the most convincing alternative hypothesis. If $a_n > 0$ the scaling factor is interpreted as an additional penalization for the complexity of the model.

As an alternative method **maximal p -value criterion** (MPVC) is also proposed in Pokarowski and Mielniczuk (2011), which involves test statistics for a family of 2^m null hypotheses of the form

$$H_0 : f \in \mathcal{M}_{\underline{l}}$$

versus

$$H_1 : f \in \mathcal{M}_{\text{full}} \setminus \mathcal{M}_{\underline{l}},$$

where $\text{full} = \{1, \dots, m\}$ corresponds to the largest considered model. In this case we choose the model for which likelihood ratio statistic

$$\Lambda_{n,\underline{l},\text{full}} = 2 \log \prod_{j=1}^n \frac{f_{\hat{\theta}_{\text{ML}}^{\text{full}}}(X_j)}{f_{\hat{\theta}_{\text{ML}}^{\underline{l}}}(X_j)} \quad (4)$$

attains the largest scaled p -value:

$$\hat{\underline{l}}_{\text{MPVC}} = \arg \max_{\underline{l} \in \{1, \dots, m\}} p(\Lambda_{n,\underline{l},\text{full}} | m - |\underline{l}|) e^{-a_n |\underline{l}|}, \quad (5)$$

where $a_n \geq 0$, $a_n/n \rightarrow 0$ as $n \rightarrow \infty$ and, as before, $p(\Lambda_{n,\underline{l},\text{full}} | m - |\underline{l}|)$ is an approximate p -value defined as in (3). We also put $p(\Lambda_{n,\text{full},\text{full}} | 0) := 1$. Here we use the fact that likelihood ratio statistic (4) has under $H_0 : f \in \mathcal{M}_{\underline{l}}$ and for fixed m an asymptotic $\chi_{m-|\underline{l}|}^2$ distribution. The motivation is similar to the motivation of mPVC, namely we choose a model which we are the least inclined to reject when compared to the full model.

The aim of the paper is to study properties of the above criteria when fixed m is replaced by m_n possibly depending on n . In other words, the number of models is allowed to change with the sample size. Such assumption allows us to consider the case when unknown density f belongs to one of the models only ultimately as well as the case when the logarithm of f has infinite expansion with respect to $(b_i)_1^\infty$. We assume throughout that the size m_n of the list is nondecreasing function of n . In the paper we establish conditions under which consistency properties of mPVC and MPVC hold in such settings. In particular it turns out that under the introduced scaling conditions MPVC rule in general is conservative. The assumption $a_n \rightarrow \infty$ is a sufficient condition for consistency of MPVC (cf. Theorems 5 and 6). This is a difference between mPVC and MPVC criteria as mPVC may be consistent for $\limsup a_n < \infty$. In the following we use the notation m_n wherever possible, however in some formulas in order to avoid subscripts of multiple levels we abbreviate the symbol m_n to m .

Observe that both p -value criteria are strictly monotone functions of the maximized likelihood of a model provided the number of degrees of freedom is fixed. The same property holds for BIC and AIC criteria. It follows that if two criteria enjoying this property choose models having the same number of parameters, e.g. if $|\hat{\underline{l}}_{\text{mPVC}}| = |\hat{\underline{l}}_{\text{MPVC}}|$, and these models are uniquely determined then the chosen models necessarily coincide.

Before presenting the main results of the paper, we provide several auxiliary results concerning the properties of likelihood ratio statistic Λ_n , which are crucial to prove the consistency of mPVC and MPVC in the case of exponential families of growing dimensions. Some of them, in particular Lemmas 3 and 4, are of independent interest. In their statements the notions of relative entropy and information projection will be used. Thus, let $D(f||g)$ denote relative entropy (or Kullback–Leibler distance) between densities f and g , which is equal to $\int_{-\infty}^{\infty} f \log(f/g)$ if f is absolutely continuous with respect to g , and ∞ otherwise.

For the sake of simplicity of notation in the following we use the symbol $\hat{f}_{\underline{i}}$ to denote the maximum likelihood estimator of density f in family $\mathcal{M}_{\underline{i}}$, i.e. $\hat{f}_{\underline{i}} = f_{\hat{\theta}_{\text{ML}}^{\underline{i}}}$. Maximum likelihood estimator $\hat{\theta}_{\text{ML}}^{\underline{i}}$, if it exists, satisfies $E_{\hat{f}_{\underline{i}}}(b_j(X)) = \frac{1}{n} \sum_{i=1}^n b_j(X_i)$ for $j \in \underline{i}$. It easily follows that $\Lambda_{n,0,\underline{i}}$ defined in (2) satisfies

$$\frac{1}{2n} \Lambda_{n,0,\underline{i}} = \hat{\theta}_{\text{ML}}^{\underline{i}} \circ \bar{b}_n^{\underline{i}} - \psi_{\underline{i}}(\hat{\theta}_{\text{ML}}^{\underline{i}}) = D(\hat{f}_{\underline{i}}||f_0), \quad (6)$$

where $\bar{b}_n^{\underline{i}} = (n^{-1} \sum_{i=1}^n b_j(X_i))_{j \in \underline{i}}$, while $\Lambda_{n,\underline{i},\text{full}}$ defined in (4) equals $2n(D(\hat{f}_{\text{full}}||f_0) - D(\hat{f}_{\underline{i}}||f_0))$. Thus as p -value $p(\Lambda_{n,0,\underline{i}} | |\underline{i}|)$ is a strictly monotone function of $\Lambda_{n,0,\underline{i}}$ for a fixed $|\underline{i}|$ it follows that on the stratum $|\underline{i}| = j$ mPVC criterion chooses an alternative $\underline{i}$ such that $\hat{f}_{\underline{i}}$ has the largest KL distance from the uniform density.

2.2 Information projection and basic assumptions

Suppose that $\theta_{\underline{i}}^* \in \mathbb{R}^{|\underline{i}|}$ is the unique vector which satisfies the equation $\int b^{\underline{i}} f_{\theta_{\underline{i}}^*} = \int b^{\underline{i}} f$ where $b^{\underline{i}}(x) = (b_j(x))_{j \in \underline{i}}$. Then $f_{\theta_{\underline{i}}^*}$ minimizes $D(f||\cdot)$ on $\mathcal{M}_{\underline{i}}$. The density $f_{\underline{i}}^* := f_{\theta_{\underline{i}}^*}$ is called an information projection of f onto the exponential family $\mathcal{M}_{\underline{i}}$. It is also characterized by the Pythagorean-like equality $D(f||g) = D(f||f_{\underline{i}}^*) + D(f_{\underline{i}}^*||g)$ valid for any $g \in \mathcal{M}_{\underline{i}}$. This implies uniqueness of the projection. For more of its properties see e.g. Barron and Sheu (1991).

Moreover, define the Kullback–Leibler distance between density f and the set $\mathcal{M}$ of densities as

$$D(f, \mathcal{M}) = \inf\{D(f||g) : g \in \mathcal{M}\}$$

and let $\underline{t}_m^*$ be defined as

$$\underline{t}_m^* = \{j : \theta_{\text{full},j}^* \neq 0\},$$

i.e. $\underline{t}_m^*$ is the set of indices of nonzero coefficients of the parameter $\theta_{\text{full}}^* \in \mathbb{R}^m$ of information projection of f onto $\mathcal{M}_{\text{full}}$. As $f_{\text{full}}^* = f_{\underline{t}_m^*}^*$ it follows that $\underline{t}_m^*$ is the minimal with respect to inclusion set of indices such that

$$\underline{t}_m^* = \arg \min_{\underline{i} \subset \{1, \dots, m\}} D(f, \mathcal{M}_{\underline{i}}).$$

Indeed, the fact that every other model which attains the minimum of $D(f, \cdot)$ includes $\mathcal{M}_{\underline{t}_m^*}$ as its subset is implied by Pythagorean-like equality mentioned above and assumed linear independence of the system $(b_i)_{i=0}^\infty$ with $b_0 \equiv 1$. Our aim in the paper is to identify $\underline{t}_m^*$ for a given family $\{\mathcal{M}_{\underline{i}}\}_{\underline{i} \subset \{1, \dots, m_n\}}$ using the introduced model selection criteria.

In the following f will denote an arbitrary density on $[0, 1]$. Assumptions (A1), (A2) and (A5) below are general assumptions about f whereas (A6) stipulates that $\log f$ has an expansion with respect to functions $(b_i)_{i=0}^\infty$ which converges in $L^2[0, 1]$. Existence of information projection is discussed in Theorem 3.3 in [Wainwright and Jordan \(2008\)](#). Assumption (A3) concerns growth of $\|b_j\|_\infty$, where $\|b\|_\infty = \sup_{x \in [0, 1]} |b(x)|$ denotes supremum norm of b , and (A4) constrains growth of m_n . For all main results (A6) will be assumed.

- (A1) for every $m_n \in \mathbb{N}$ and $\underline{i} \subset \{1, \dots, m_n\}$ the information projection $f_{\underline{i}}^*$ exists,
- (A2) the sequence $(\max_{\underline{i} \subset \{1, \dots, m_n\}} \|\log f_{\underline{i}}^*\|_\infty)_n$ is bounded in $n \in \mathbb{N}$,
- (A3) $V_k = \max_{j=1, \dots, k} \|b_j\|_\infty = O(k^\omega)$ for some $\omega \geq 0$ as $k \rightarrow \infty$,
- (A4) $n/m_n^{2+4\omega} \rightarrow \infty$ as $n \rightarrow \infty$,
- (A5) $\|f\|_\infty < \infty$ and $\inf_{x \in [0, 1]} f(x) > 0$,
- (A6) $f(x) = \exp\{\sum_{j=1}^\infty \theta_j b_j(x) - \psi(\theta)\}$ for $x \in [0, 1]$ and some $\theta \in l^2$ such that $\psi(\theta) < \infty$, where $\psi(\theta) = \log \int_0^1 \exp\{\sum_{j=1}^\infty \theta_j b_j(x)\} dx$.

The lemmas below are stated under minimal subsets of assumptions from (A1) to (A6). Their proofs rely on methods developed by [Barron and Sheu \(1991\)](#). In particular Lemma 2 is an extension of a result proved there to the case when quantities considered in (8)–(9) are maximized over all subsets of $\{1, \dots, m_n\}$. Such properties are useful when selection rule $\hat{\underline{i}}$ based on an exhaustive search of all subsets is considered. Lemmas 3–5 are to the best of our knowledge new.

Let $(\theta_j)_{j=1}^\infty$ be the vector of coefficients given in the representation (A6) of f . We use throughout the following notation:

$$\underline{t} := \{j : \theta_j \neq 0\}$$

for the set of indices of nonzero coefficients of $(\theta_j)_{j=1}^\infty$. Note that if f belongs to an exponential family set $\underline{t}$ corresponds to the minimal model containing it discussed in the Introduction. For every vector $v \in \mathbb{R}^\infty$ let

$$v^{\underline{i}} := (v_j)_{j \in \underline{i}}$$

be the vector formed from $(v_j)_{j=1}^\infty$ by choosing only those coordinates whose indices are elements of the set $\underline{i}$, $\underline{i} \subset \mathbb{N}$. Moreover for $m \in \mathbb{N}$ we set $v^m = v^{\{1, \dots, m\}}$.

If f belongs to some exponential family of distributions $\mathcal{M}_{\underline{t}}$, i.e. $|\underline{t}| < \infty$, and the maximal element of $\underline{t}$ denoted as $\max \underline{t} \leq m_n$ then $\underline{t} = \underline{t}_m^*$ since in this case $\theta_{\text{full}}^* = \theta_{\text{full}}$. Observe also that if $\underline{t}$ is infinite then f does not belong to any exponential model (1). We recall that $m = m_n$ may depend on a sample size n .

2.3 Auxiliary lemmas

In this subsection we give several auxiliary lemmas which are crucial to prove the consistency of mPVC and MPVC in the case of exponential families of growing dimensions. The proofs of them are deferred to the Appendix.

Lemma 1 *If conditions (A2), (A5) and (A6) hold then sequence $(\max_{\underline{i} \in \{1, \dots, m_n\}} \|\theta_{\underline{i}}^*\|)_n$ is bounded in $n \in \mathbb{N}$.*

Lemma 2 *If conditions (A1)–(A4) hold then $P(\hat{\theta}_{\text{ML}}^{\underline{i}} \text{ exists for every } \underline{i} \in \{1, \dots, m_n\}) \rightarrow 1$ as $n \rightarrow \infty$ and*

$$\max_{\underline{i} \in \{1, \dots, m_n\}} D(f_{\underline{i}}^* \| \hat{f}_{\underline{i}}) = O_P \left(m_n^{1+2\omega} / n \right), \quad (7)$$

$$\max_{\underline{i} \in \{1, \dots, m_n\}} D(\hat{f}_{\underline{i}} \| f_{\underline{i}}^*) = O_P \left(m_n^{1+2\omega} / n \right). \quad (8)$$

Lemmas 3 and 4 concern the asymptotic properties of likelihood ratio statistic $\Lambda_{n,0,\underline{i}}$ in the case of exponential families of growing dimensions and as such are of independent interest. For $\underline{i} \in \{1, \dots, m_n\}$ define

$$\lambda_{\underline{i}} = \theta_{\underline{i}}^* \circ E_f(b^{\underline{i}}(X)) - \psi_{\underline{i}}(\theta_{\underline{i}}^*).$$

Observe that in view of the definition of $f_{\underline{i}}^*$ we have $\lambda_{\underline{i}} = \int f \log f_{\underline{i}}^* = \int f_{\underline{i}}^* \log f_{\underline{i}}^* = D(f_{\underline{i}}^* \| f_0)$. The last equality implies $\lambda_{\underline{i}} \geq 0$.

Lemma 3 *If conditions (A1)–(A6) hold then:*

(i)

$$\max_{\underline{i} \in \{1, \dots, m_n\}} \left| \frac{1}{2n} \Lambda_{n,0,\underline{i}} - \lambda_{\underline{i}} \right| = O_P \left(\left(\frac{m_n^{1+2\omega}}{n} \right)^{1/2} \right) \text{ as } n \rightarrow \infty. \quad (9)$$

Moreover,

- (ii) if $\underline{t}_m^* \not\subset \underline{i}$ then $\lambda_{\underline{i}} < \lambda_{\underline{t}_m^*}$,
- (iii) if $\underline{t}_m^* \subset \underline{i}$ then $\lambda_{\underline{i}} = \lambda_{\underline{t}_m^*}$.

Remark 1 Note that $\hat{f}_{\underline{i}}$ and $f_{\underline{i}}^*$ do not depend on the specific system of functions $(b_j(x))_{j \in \underline{i}}$ provided the systems span the same linear space. In particular we will investigate in Lemma 4 the behavior of the quantities $D(f_{\underline{i}}^* \| \hat{f}_{\underline{i}})$ and $\Lambda_{n,0,\underline{i}}$ considering an additional system of functions $(\tilde{b}_j(x))_{j=0}^\infty$ orthonormal in $L^2([0, 1], f)$, such that $\tilde{b}_0(x) \equiv 1$ and that $sp(b_i, i \in \underline{i}_k) = sp(\tilde{b}_i, i \in \underline{i}_k)$, $k = 1, 2$, where $sp(B)$ stands for a linear space spanned by functions from system B and $\underline{i}_1 \subset \underline{i}_2 = \{1, \dots, m_n\}$. Such system can be obtained by Gram-Schmidt orthonormalisation procedure. Obviously $\tilde{b}_j$ depend on f but it follows from Lemma 7 in the Appendix that in order to bound $D(f_{\underline{i}}^* \| \hat{f}_{\underline{i}})$ using the constructed system it is enough to evaluate the constant $A_m(f)$ appearing there. Such useful approach was used by Barron and Sheu (1991) (cf. proof

of (6.7) in their paper) and it yields better rates of convergence of $D(f_{\underline{l}}^* || \hat{f}_{\underline{l}})$ than a direct method based on [Yurinski \(1976\)](#) inequality.

Lemma 4 *If conditions (A1)–(A5) hold then*

$$\max_{\underline{l}: \underline{l}_m^* \subset \underline{l} \subset \{1, \dots, m_n\}} |\Lambda_{n,0,\underline{l}} - \Lambda_{n,0,\underline{l}_m^*}| = \Lambda_{n,0,\text{full}} - \Lambda_{n,0,\underline{l}_m^*} = O_P(m_n) \quad \text{as } n \rightarrow \infty.$$

Lemma 5 *If assumptions (A1), (A2), (A5) and (A6) hold and $|\underline{t}| < \infty$ then there exists a constant $a > 0$ such that for every $n \in \mathbb{N}$*

$$\min_{\underline{l}: \underline{l} \subset \{1, \dots, m_n\}; \underline{t} \not\subset \underline{l}} D(f || f_{\underline{l}}^*) > a.$$

3 Main results

Sections 3.1 and 3.2 deal with consistency of minimal and maximal p -value criterions, respectively. In Sect. 3.3 greedy counterparts of the criteria are introduced and their consistency is proved. We state first a lemma of a different character than those presented in Sect. 2. It concerns bounds for a tail of the χ_k^2 distribution. Recall that $p(x|k)$ is the p -value defined in (3).

For $x > 0$ and $k \in \mathbb{N}$ let

$$C(x, k) = e^{-x/2} \left(\frac{x}{2}\right)^{k/2-1} \left(\Gamma\left(\frac{k}{2}\right)\right)^{-1}$$

and

$$B(x, k) = (x/(x - k + 2))C(x, k).$$

Lemma 6 *We have*

- (i) *for $k = 1$ and $x > 0$, $B(x, 1) \leq p(k|1) \leq C(x, 1)$;*
- (ii) *for $k > 1$ and $x > 0$, $p(x|k) \geq C(x, k)$ and for $k > 1$ and $x > k - 2$, $p(x|k) \leq B(x, k)$;*

Part (i) of the above lemma was proved by [Gordon \(1941\)](#) whereas part (ii) follows from [Ingolot and Ledwina \(2006\)](#) after noticing that the product $c(k)\mathcal{E}_k(x)$ of the functions defined there equals $C(x, k)$.

For all the results below we assume that assumption (A6) holds, i.e. the logarithm of the underlying density f has L^2 expansion w.r.t. system $(b_i)_{i=0}^\infty$.

3.1 Minimal p -value criterion mPVC

The first main result states that if m is held constant then $\hat{\underline{l}}_{\text{mPVC}}$ identifies with probability tending to 1 the indices corresponding to the nonzero coefficients of the information projection of f on $\mathcal{M}_{\text{full}}$.

Theorem 1 (mPVC consistency) *If conditions (A1), (A5) and (A6) hold and $m_n = m$ is constant then $\lim_{n \rightarrow \infty} P(\hat{\underline{l}}_{\text{mPVC}} = \underline{l}_m^*) = 1$.*

Proof Assume that $|t_m^*| > 1$ and let $\tilde{\Lambda}_{n,0,\underline{i}} = \max(\Lambda_{n,0,\underline{i}}, 2)$. We have

$$\begin{aligned} & P \left(\min_{\substack{\underline{i}: t_m^* \notin \underline{i} \subset \{1, \dots, m_n\}}} p(\Lambda_{n,0,\underline{i}} \mid |\underline{i}|) e^{a_n |\underline{i}|} < p(\Lambda_{n,0,t_m^*} \mid |t_m^*|) e^{a_n |t_m^*|} \right) \\ & \leq P \left(\min_{\substack{\underline{i}: t_m^* \notin \underline{i} \subset \{1, \dots, m_n\}}} p(\tilde{\Lambda}_{n,0,\underline{i}} \mid |\underline{i}|) e^{a_n |\underline{i}|} \leq p(\Lambda_{n,0,t_m^*} \mid |t_m^*|) e^{a_n |t_m^*|} \right). \end{aligned}$$

Lemma 6(ii) implies for $\Lambda_{n,0,t_m^*} > |t_m^*| - 2$,

$$p(\Lambda_{n,0,t_m^*} \mid |t_m^*|) \leq C(\Lambda_{n,0,t_m^*}, |t_m^*|) \frac{\Lambda_{n,0,t_m^*}}{\Lambda_{n,0,t_m^*} - |t_m^*| + 2}.$$

Parts (i) and (ii) of Lemma 6 yield

$$p(\tilde{\Lambda}_{n,0,\underline{i}} \mid |\underline{i}|) \geq C(\tilde{\Lambda}_{n,0,\underline{i}}, |\underline{i}|) \frac{\tilde{\Lambda}_{n,0,\underline{i}}}{\tilde{\Lambda}_{n,0,\underline{i}} + 1}.$$

Moreover, since $\left| \frac{1}{2n} \tilde{\Lambda}_{n,0,\underline{i}} - \lambda_{\underline{i}} \right| \leq \left| \frac{1}{2n} \Lambda_{n,0,\underline{i}} - \lambda_{\underline{i}} \right| + \frac{1}{2n}$, Lemma 3(i), (ii) and the fact that m is constant yield that for some $\varepsilon > 0$

$$P \left(\max_{\substack{\underline{i}: t_m^* \notin \underline{i} \subset \{1, \dots, m\}}} \frac{1}{2n} \tilde{\Lambda}_{n,0,\underline{i}} < \frac{1}{2n} \Lambda_{n,0,t_m^*} - \varepsilon \right) \rightarrow 1.$$

Note that condition $|t_m^*| > 0$ yields $\Lambda_{n,0,t_m^*} \xrightarrow{\mathcal{P}} \infty$ as $\frac{1}{2n} \Lambda_{n,0,t_m^*} \xrightarrow{\mathcal{P}} D(f_{t_m^*}^* \mid f_0) > 0$. Thus the definition of $C(x, p)$ together with $a_n/n \rightarrow 0$, the fact that $\tilde{\Lambda}_{n,0,\underline{i}}^{(|\underline{i}|-2)/2} \geq 1$ for $|\underline{i}| \geq 2$ and the inequalities above imply

$$P \left(\min_{\substack{\underline{i}: |\underline{i}| > 1 \text{ \& } t_m^* \notin \underline{i} \subset \{1, \dots, m\}}} p(\Lambda_{n,0,\underline{i}} \mid |\underline{i}|) e^{a_n |\underline{i}|} \leq p(\Lambda_{n,0,t_m^*} \mid |t_m^*|) e^{a_n |t_m^*|} \right) \rightarrow 0.$$

In order to obtain analogous convergence for $\underline{i}$ such that $|\underline{i}| = 1$ we reason similarly using inequality

$$e^{-x/2} \left(\frac{x}{2} \right)^{-1/2} \frac{x}{x+1} \geq \frac{2}{3} e^{-x/2 - \ln(x/2)/2} \geq \frac{2}{3} e^{-x}$$

for $x = \tilde{\Lambda}_{n,0,\underline{i}} \geq 2$.

On the other hand for $\underline{i} \subset \{1, \dots, m\}$ such that $t_m^* \subsetneq \underline{i}$ we have

$$\begin{aligned} & \log \left(e^{-\frac{\Lambda_{n,0,\underline{i}}}{2}} \left(\frac{\Lambda_{n,0,\underline{i}}}{2} \right)^{\frac{|\underline{i}|-2}{2}} \Gamma \left(\frac{|\underline{i}|}{2} \right)^{-1} e^{a_n |\underline{i}|} \right) \\ & - \log \left(e^{-\frac{\Lambda_{n,0,t_m^*}}{2}} \left(\frac{\Lambda_{n,0,t_m^*}}{2} \right)^{\frac{|t_m^*|-2}{2}} \Gamma \left(\frac{|t_m^*|}{2} \right)^{-1} \left(1 + \frac{|t_m^*|-2}{\Lambda_{n,0,t_m^*} - |t_m^*| + 2} \right) e^{a_n |t_m^*|} \right) \end{aligned}$$

$$\geq \frac{1}{2} \log n + a_n - \max_{\substack{i: t_m^* \subset \underline{i}}} \frac{1}{2} |\Lambda_{n,0,\underline{i}} - \Lambda_{n,0,t_m^*}| + \frac{|\underline{i}| - 2}{2} \log \frac{\Lambda_{n,0,\underline{i}}}{2n} \\ - \frac{|t_m^*| - 2}{2} \log \frac{\Lambda_{n,0,t_m^*}}{2n} - \log \frac{\Gamma(m/2)}{\Gamma(|t_m^*|/2)} - \log \left(1 + \frac{t_m^* - 2}{\Lambda_{n,0,t_m^*} - |t_m^*| + 2} \right).$$

which implies, in view of Lemma 4 and Lemma 6(i) by noting that $|\underline{i}| > 1$ and bounding $p(\Lambda_{n,0,\underline{i}} | |\underline{i}|)$ from below and $p(\Lambda_{n,0,t_m^*} | |t_m^*|)$ from above, that

$$P \left(\min_{\substack{i: t_m^* \subset \underline{i} \subset \{1, \dots, m\}}} p(\Lambda_{n,0,\underline{i}} | |\underline{i}|) e^{a_n |\underline{i}|} \leq p(\Lambda_{n,0,t_m^*} | |t_m^*|) e^{a_n |t_m^*|} \right) \rightarrow 0.$$

The proof in the case $|t_m^*| = 1$ is similar. If $|t_m^*| = 0$, i.e. projected f on the full model $f_{\text{full}}^* \equiv 1$, then $\Lambda_{n,0,t_m^*} = 0$ and Lemma 4 implies that $\max_{\underline{i} \subset \{1, \dots, m\}} \Lambda_{n,0,\underline{i}} = O_P(1)$. Thus

$$P \left(\min_{\underline{i} \subset \{1, \dots, m\}} p(\Lambda_{n,0,\underline{i}} | |\underline{i}|) e^{a_n |\underline{i}|} < n^{-1/2} \right) \rightarrow 0.$$

□

In the case of a correct specification of the list of models we call a selection rule $\hat{\underline{i}} = \hat{\underline{i}}(X_1, \dots, X_n)$ conservative if $P(\underline{t} \subset \hat{\underline{i}}) \rightarrow 1$ when $n \rightarrow \infty$. Theorem below states the conditions for conservativeness and consistency of mPVC criterion when the true density f belongs to one of the models on the list. Observe that some growth conditions on m_n have to be imposed in this context as the method of calculating approximate p -values relies on Wilks' theorem and the quality of approximation $\Lambda_{n,0,\underline{i}}$ by $\chi_{|\underline{i}|}^2$ deteriorates when $|\underline{i}|$ increases together with n .

Theorem 2 (mPVC consistency) *Assume (A1)–(A6), $|t| < \infty$ and $\lim_{n \rightarrow \infty} m_n \geq \max \underline{t}$. Then*

- (i) $\lim_{n \rightarrow \infty} P(\underline{t} \subset \hat{\underline{i}}_{\text{mPVC}}) = 1$,
- (ii) if $m_n \log m_n = o(\log n + a_n)$ as $n \rightarrow \infty$ then $\lim_{n \rightarrow \infty} P(\hat{\underline{i}}_{\text{mPVC}} = \underline{t}) = 1$.

Proof Consider the case when $|t| > 0$. Assume first that $\underline{i}$ is such that $\underline{t} \not\subset \underline{i} \subset \{1, \dots, m_n\}$. It is sufficient to show that with $\tilde{\Lambda}_{n,0,\underline{i}} = \max(\Lambda_{n,0,\underline{i}}, 2)$

$$P \left(\min_{\substack{i: \underline{t} \not\subset \underline{i} \subset \{1, \dots, m_n\}}} p(\tilde{\Lambda}_{n,0,\underline{i}} | |\underline{i}|) e^{a_n |\underline{i}|} \leq p(\Lambda_{n,0,\underline{t}} | |t|) e^{a_n |t|} \right) \rightarrow 0. \quad (10)$$

Lemma 6 yields for $\tilde{\Lambda}_{n,0,\underline{i}} > 0$ and $|\underline{i}| > 0$

$$p(\tilde{\Lambda}_{n,0,\underline{i}} | |\underline{i}|) \geq e^{-\tilde{\Lambda}_{n,0,\underline{i}}/2} (\tilde{\Lambda}_{n,0,\underline{i}}/2)^{|\underline{i}|/2-1} \Gamma(|\underline{i}|/2)^{-1} \tilde{\Lambda}_{n,0,\underline{i}} / (\tilde{\Lambda}_{n,0,\underline{i}} + 1) \quad (11)$$

and for $\Lambda_{n,0,\underline{t}} > |t| - 2$ and $|t| > 1$

$$p(\Lambda_{n,0,\underline{t}} | |t|) \leq e^{-\Lambda_{n,0,\underline{t}}/2} (\Lambda_{n,0,\underline{t}}/2)^{|t|/2-1} \Gamma(|t|/2)^{-1} \Lambda_{n,0,\underline{t}} / (\Lambda_{n,0,\underline{t}} - |t| + 2). \quad (12)$$

Assume that $|\underline{t}| > 1$. The case $|\underline{t}| = 1$ is analogous. As $\min_{i: \underline{t} \not\subset \underline{i}} \tilde{A}_{n,0,\underline{i}} > 0$, $A_{n,0,\underline{t}} \xrightarrow{\mathcal{P}} \infty$ (since $f \not\equiv 1$) and $|\underline{t}| < \infty$, the assumptions of Lemma 6 are satisfied with probability tending to 1 as $n \rightarrow \infty$. Observe that $D(f||f_{\underline{i}}^*) = D(f_{\underline{t}}||f_{\underline{i}}^*) = \lambda_{\underline{t}} - \lambda_{\underline{i}}$. We have

$$\begin{aligned} \frac{1}{2n} A_{n,0,\underline{t}} - \frac{1}{2n} \tilde{A}_{n,0,\underline{i}} &= \left(\frac{1}{2n} A_{n,0,\underline{t}} - \lambda_{\underline{t}} \right) - \left(\frac{1}{2n} \tilde{A}_{n,0,\underline{i}} - \lambda_{\underline{i}} \right) + (\lambda_{\underline{t}} - \lambda_{\underline{i}}) \\ &\geq \left(\frac{1}{2n} A_{n,0,\underline{t}} - \lambda_{\underline{t}} \right) - \max_{i: \underline{t} \not\subset \underline{i} \subset \{1, \dots, m_n\}} \left| \frac{1}{2n} \tilde{A}_{n,0,\underline{i}} - \lambda_{\underline{i}} \right| \\ &\quad + \min_{i \subset \{1, \dots, m_n\}: \underline{t} \not\subset \underline{i}} D(f||f_{\underline{i}}^*). \end{aligned}$$

As in the proof of Theorem 1 Lemmas 3(i), (ii) and 5 imply that there exists $\varepsilon > 0$ such that

$$P \left(\min_{i: \underline{t} \not\subset \underline{i} \subset \{1, \dots, m_n\}} \frac{1}{2n} A_{n,0,\underline{t}} - \frac{1}{2n} A_{n,0,\underline{i}} > \varepsilon \right) \rightarrow 1. \quad (13)$$

It is easily seen that the inequality $\Gamma(p/2) \leq (\lceil p/2 \rceil - 1)!$ for $p \in \mathbb{N} \setminus \{1\}$ implies

$$\max_{i: \underline{i} \subset \{1, \dots, m_n\}} \log \frac{\Gamma(|\underline{i}|/2)}{\Gamma(|\underline{t}|/2)} = O(m_n \log m_n) = o(n).$$

Moreover,

$$\begin{aligned} \frac{|\underline{t}| - 2}{2} \log \frac{A_{n,0,\underline{t}}}{2} &= O_p(\log n); \\ \min_{i: \underline{t} \not\subset \underline{i} \subset \{1, \dots, m_n\}} \log \left(\frac{\tilde{A}_{n,0,\underline{i}}}{\tilde{A}_{n,0,\underline{i}} + 1} \frac{A_{n,0,\underline{t}} - |\underline{t}| + 2}{A_{n,0,\underline{t}}} \right) &= O_p(1); \end{aligned}$$

for $L(n, \underline{i}) := \frac{|\underline{i}| - 2}{2} \log \left(\frac{\tilde{A}_{n,0,\underline{i}}}{2} \right)$ we have $L(n, \underline{i}) \geq 0$ for $|\underline{i}| \geq 2$ and Lemma 3 easily implies $\min_{i: |\underline{i}|=1} L(n, \underline{i}) \geq -cn$ with probability tending to 1 for any positive c . Thus using the assumption $a_n/n \rightarrow 0$, bounds (11) and (12) and the relations above, we obtain equation (10) and it follows that $\lim_{n \rightarrow \infty} P(\underline{t} \subset \hat{I}_{\text{mPVC}}) = 1$.

Assume now that $\underline{t} \subsetneq \underline{i} \subset \{1, \dots, m_n\}$, where $f \not\equiv 1$. As $\lambda_{\underline{i}} = \lambda_{\underline{t}} > 0$, Lemma 6 may be applied to both p -values. As in proof of Theorem 1 we have

$$\begin{aligned} &\log \left(e^{-\frac{A_{n,0,\underline{i}}}{2}} \left(\frac{A_{n,0,\underline{i}}}{2} \right)^{\frac{|\underline{i}| - 2}{2}} \Gamma \left(\frac{|\underline{i}|}{2} \right)^{-1} e^{a_n |\underline{i}|} \right) \\ &\quad - \log \left(e^{-\frac{A_{n,0,\underline{t}}}{2}} \left(\frac{A_{n,0,\underline{t}}}{2} \right)^{\frac{|\underline{t}| - 2}{2}} \Gamma \left(\frac{|\underline{t}|}{2} \right)^{-1} \frac{A_{n,0,\underline{t}}}{A_{n,0,\underline{t}} - |\underline{t}| + 2} e^{a_n |\underline{t}|} \right) \end{aligned}$$

$$\begin{aligned} &\geq \frac{1}{2} \log n + a_n - \max_{\underline{i}: \underline{t} \subset \underline{i}} \frac{1}{2} |A_{n,0,\underline{i}} - A_{n,0,\underline{t}}| + \frac{|\underline{i}| - 2}{2} \log \frac{A_{n,0,\underline{i}}}{2n} \\ &\quad - \frac{|\underline{t}| - 2}{2} \log \frac{A_{n,0,\underline{t}}}{2n} - \log \frac{\Gamma(m_n/2)}{\Gamma(|\underline{t}|/2)} + \log \left(\frac{A_{n,0,\underline{t}} - |\underline{t}| + 2}{A_{n,0,\underline{t}}} \right). \end{aligned}$$

Lemma 4 implies $\max_{\underline{i}: \underline{t} \subset \underline{i} \subset \{1, \dots, m_n\}} |A_{n,0,\underline{i}} - A_{n,0,\underline{t}}| = O_P(m_n)$.

Moreover observe that

$$\begin{aligned} &\frac{|\underline{i}| - 2}{2} \log \frac{A_{n,0,\underline{i}}}{2n} - \frac{|\underline{t}| - 2}{2} \log \frac{A_{n,0,\underline{t}}}{2n} \\ &= \frac{|\underline{i}| - 2}{2} \left(\log \frac{A_{n,0,\underline{i}}}{2n} - \log \frac{A_{n,0,\underline{t}}}{2n} \right) + \frac{|\underline{i}| - |\underline{t}|}{2} \log \frac{A_{n,0,\underline{t}}}{2n}. \end{aligned}$$

The first term in the above decomposition is positive as $|\underline{i}| \geq 2$ and $\underline{t} \subset \underline{i}$ and the second is easily seen to be larger than $-c \log n$ with probability tending to 1 for any positive c . Finally, using the fact that $\log \Gamma(\frac{m_n}{2}) / \Gamma(\frac{|\underline{t}|}{2}) = O(m_n \log m_n)$ together with the assumption $m_n \log m_n = o(\log n + a_n)$ we obtain

$$\lim_{n \rightarrow \infty} P \left(\min_{\underline{i}: \underline{t} \subset \underline{i} \subset \{1, \dots, m_n\}} p(A_{n,0,\underline{i}} | |\underline{i}|) e^{a_n |\underline{i}|} > p(A_{n,0,\underline{t}} | |\underline{t}|) e^{a_n |\underline{t}|} \right) = 1.$$

The proof in the case $|\underline{t}| = 1$ is similar. Consider now the case when $f \equiv 1$, i.e. the minimal model is true. We will show that

$$P \left(\min_{\underline{i}: \underline{i} \subset \{1, \dots, m_n\}} p(A_{n,0,\underline{i}} | |\underline{i}|) e^{a_n |\underline{i}|} < n^{-1/2} \right) \rightarrow 0.$$

As Lemma 4 yields that when the minimal model is true $\max_{\underline{i} \subset \{1, \dots, m_n\}} A_{n,0,\underline{i}} = O_P(m_n)$ it is enough to show that for any $C > 0$

$$n^{1/2} \min_{k=1, \dots, m_n} e^{a_n k} P(Z_k > C m_n) \rightarrow \infty,$$

where Z_k pertains to χ_k^2 distribution. This is easily shown using Lemma 6, assumption $m_n \log m_n = o(a_n + \log n)$ and the fact that $\max_{k \leq m_n} \log \Gamma(k/2) = O(m_n \log m_n)$. $\square$

Remark 2 Careful examination of the previous proof yields that under conditions of Theorem 2

$$P(\underline{t} \notin \hat{\mathcal{I}}_{\text{mPVC}}) \leq C(m_n^{1+2\omega}/n)^{1/2}$$

for some absolute constant $C > 0$.

Theorem 3 below states that any index i_0 corresponding to a nonzero coefficient in the expansion of $\log f$ will be eventually included in $\hat{t}_{\text{mPVC}}$. We call a selection rule $\hat{t}$ conservative for f satisfying assumption (A6) if for any $M \in \mathbb{N}P(\underline{t} \cap \{1, \dots, M\} \subset \hat{t}) \rightarrow 1$. Note that this notion coincides with the usual definition of conservativeness under correct specification given above Theorem 2 when $|\underline{t}| < \infty$. In Theorem 3 $\underline{t}$ can be either infinite or finite. However, an interesting part of the result corresponds to the former case, i.e. when $\log f$ has infinite expansion w.r.t. system $(b_i)_{i=0}^\infty$. As it was noticed before this corresponds to misspecification case when density f does not belong to any model $\mathcal{M}_{m_n}$, $n = 1, 2, \dots$. For finite $\underline{t}$ Theorem 3 coincides with Theorem 2(i).

Theorem 3 Assume (A1)–(A6), moreover that (i) $\sum_{i=1}^\infty \theta_i b_i(x)$ is uniformly convergent on $[0, 1]$ and (ii) $a_n |t_{m_n}(f)| = o(n)$, where $\underline{t}_{m_n} = \underline{t} \cap \{1, \dots, m_n\}$. If $\lim_{n \rightarrow \infty} m_n \geq \max \underline{t}$ then $\hat{t}_{\text{mPVC}}$ is conservative.

Proof It is enough to prove that for any $i_0 \in \underline{t}$ $P(i_0 \in \hat{t}_{\text{mPVC}}) \rightarrow 1$ as $n \rightarrow \infty$. The proof proceeds analogously to the proof of Theorem 2 (a). In the following we prove that there exists $\varepsilon > 0$ such that

$$\max_{i: i_0 \notin \underline{t} \subset \{1, \dots, m_n\}} \lambda_{\underline{t}} < \lambda_{\underline{t}_{m_n}} - \varepsilon \quad \text{as } n \rightarrow \infty. \quad (14)$$

First note that reasoning analogously as in the proof of Lemma 5 we have

$$\liminf_{n \rightarrow \infty} \min_{i: i_0 \notin \underline{t} \subset \{1, \dots, m_n\}} D(f \| f_{\underline{t}}^*) > 0. \quad (15)$$

Indeed, assumptions (A2) and (A5) imply as in Lemma 1 that there exists a constant $C > 0$ such that $D(f \| f_{\underline{t}}^*) \geq C \int_0^1 (\log(f/f_{\underline{t}}^*))^2$ and orthogonality of $(b_j)_{j=0}^\infty$ yields

$$\int_0^1 (\log(f/f_{\underline{t}}^*))^2 = \sum_{j \in \underline{t}} (\theta_j - \theta_j^*)^2 + \sum_{j \notin \underline{t}} \theta_j^2 + (\psi(\theta) - \psi_{\underline{t}}(\theta_{\underline{t}}^*))^2 \geq \theta_{i_0}^2.$$

(throughout, the integration is performed w.r.t. Lebesgue measure λ on $[0, 1]$). Thus $D(f \| f_{\underline{t}}^*) > C \theta_{i_0}^2 > 0$ which proves (15).

On the other hand we have

$$\lim_{n \rightarrow \infty} D(f \| f_{\underline{t}_{m_n}}^*) = 0. \quad (16)$$

To see this note that since $f_{\underline{t}_{m_n}}^*$ minimizes $D(f \| \cdot)$ on $\mathcal{M}_{\underline{t}_{m_n}}$ then $D(f \| f_{\underline{t}_{m_n}}^*) \leq D(f \| f_{\theta^{\underline{t}_{m_n}}})$, where $\theta^{\underline{t}_{m_n}}$ is defined in Sect. 2.2. Lemma 1 in Barron and Sheu (1991) implies

$$D(f||f_{\theta^{L_m}}) \leq \frac{1}{2} e^{\|\log(f/f_{\theta^{L_m}})\|_\infty} \int_0^1 f(\log(f/f_{\theta^{L_m}}))^2.$$

It is easy to see that assumption (i) implies $\inf_x f_{\theta^{L_m}}(x) > \varepsilon$ for some $\varepsilon > 0$ and sufficiently large n . Thus assumption (A5) yields $D(f||f_{\theta^{L_m}}) \leq C \int_0^1 (\log(f/f_{\theta^{L_m}}))^2 = C (\sum_{i>m_n} \theta_i^2 + (\psi(\theta) - \psi_{t_m}(\theta^{L_m}))^2)$ for some constant $C > 0$. Now (16) follows from the fact that $\sum_{i>m_n} \theta_i^2 \rightarrow 0$ and $\psi(\theta) - \psi_{t_m}(\theta^{L_m}) \rightarrow 0$ as $n \rightarrow \infty$. The last convergence follows from the proof of Lemma 4 in Barron and Sheu (1991) which implies that $|\psi(\theta) - \psi_{t_m}(\theta^{L_m})| \leq \|\sum_{i=m+1}^\infty \theta_i b_i\|_\infty \rightarrow 0$.

Properties (15) and (16) imply (14). The rest of the proof is analogous to that of Theorem 2(a). $\square$

Remark 3 Observe that the above proof also yields the following more general fact. Assume that subsets $M_n \subset 2^{\{1, \dots, m_n\}}$ are such that the following separability condition generalizing (14) holds

$$\max_{\underline{i} \in M_n} \lambda_{\underline{i}} < \lambda_{t_{m_n}} - \varepsilon \quad (17)$$

for all $n \in \mathbb{N}$ and some $\varepsilon > 0$, or equivalently $\min_{\underline{i} \in M_n} D(f||f_{\underline{i}}^*) > D(f||f_{t_{m_n}}^*) + \varepsilon$. Then under conditions of Theorem 2 $P(\hat{t}_{\text{mpvc}} \in M_n^c) \rightarrow 1$. If $M_n := \{\underline{i} \in 2^{\text{full}} : i_0 \notin \underline{i}\}$ the conclusion of Theorem 3 is obtained.

Recall that $\hat{f}_{\underline{i}}$ denotes the maximum likelihood estimator of density f in family $\mathcal{M}_{\underline{i}}$ equal to $f_{\hat{\theta}_{\text{ML}}^{\underline{i}}}$. Let $\hat{f}_{\text{mpvc}}$ be the post-model-selection estimator of f based on mpvc method, i.e. $\hat{f}_{\text{mpvc}} = \hat{f}_{\hat{t}_{\text{mpvc}}}$.

Theorem 4 Assume (A1)–(A6), $|t| < \infty$, $\lim_{n \rightarrow \infty} m_n \geq \max t$ and $m_n \log m_n = o(\log n + a_n)$. Then $D(f||\hat{f}_{\text{mpvc}}) = O_P(m_n/n)$.

Proof Observe that we have for $\hat{\underline{i}} = \hat{t}_{\text{mpvc}}$

$$D(f||\hat{f}_{\text{mpvc}}) = D(f||f_{\hat{\underline{i}}}^*) + D(f_{\hat{\underline{i}}}^*||\hat{f}_{\text{mpvc}}). \quad (18)$$

The first term is 0 with probability tending to 1 in view of Theorem 2 (i). Using Lemma 7 analogously as in the proof of Lemma 4 yields $D(f_{\hat{\underline{i}}}^*||\hat{f}_{\underline{i}}) = O_P(m_n/n)$. This and convergence $P(\hat{\underline{i}} = t) \rightarrow 1$ implied by Theorem 2 (i) completes the proof. $\square$

3.2 Maximal p-value criterion MPVC

Theorems 5 and 6 below are analogous to Theorems 1 and 2, respectively, for MPVC criterion.

Theorem 5 (MPVC consistency) *If conditions (A1), (A5) and (A6) hold, $a_n \rightarrow \infty$ and $m_n = m$ is constant then $\lim_{n \rightarrow \infty} P(\hat{t}_{\text{MPVC}} = t_m^*) = 1$.*

Proof Let $\underline{i} \subset \{1, \dots, m\}$ and assume that $\underline{t}_m^* \subsetneq \underline{i}$. Then $|\underline{i}| > |\underline{t}_m^*|$ and we have

$$\begin{aligned} P \left(p(\Lambda_{n, \underline{t}_m^*, \text{full}} | m - |\underline{t}_m^*|) e^{-a_n |\underline{t}_m^*|} > \max_{\underline{i}: \underline{t}_m^* \subsetneq \underline{i} \subset \{1, \dots, m\}} p(\Lambda_{n, \underline{i}, \text{full}} | m - |\underline{i}|) e^{-a_n |\underline{i}|} \right) \\ \geq P \left(p(\Lambda_{n, \underline{t}_m^*, \text{full}} | m - |\underline{t}_m^*|) e^{-a_n |\underline{t}_m^*|} > \max_{\underline{i}: \underline{t}_m^* \subsetneq \underline{i} \subset \{1, \dots, m\}} e^{-a_n |\underline{i}|} \right) \\ \geq P \left(p(\Lambda_{n, \underline{t}_m^*, \text{full}} | m - |\underline{t}_m^*|) > e^{-a_n} \right) \rightarrow 1. \end{aligned}$$

The last convergence follows from the assumption $a_n \rightarrow \infty$ and the fact that $\Lambda_{n, \underline{t}_m^*, \text{full}} = O_P(1)$ implied by Lemma 4.

Assume now that $\underline{i} \subset \{1, \dots, m\}$ is such that $\underline{t}_m^* \not\subset \underline{i}$. As $\frac{1}{2n} \Lambda_{n, \underline{i}, \text{full}} \xrightarrow{\mathcal{P}} D(f || f_i^*) - D(f || f_{\text{full}}^*) > 0$ then analogously as in the proof of Theorem 1 we obtain using Lemma 3

$$P \left(\min_{\underline{i}: \underline{t}_m^* \not\subset \underline{i} \subset \{1, \dots, m\}} \frac{1}{2n} \Lambda_{n, \underline{i}, \text{full}} > \frac{1}{2n} \Lambda_{n, \underline{t}_m^*, \text{full}} + \varepsilon \right) \rightarrow 1 \quad (19)$$

for some $\varepsilon > 0$. Moreover, Lemma 6 implies

$$p(\Lambda_{n, \underline{t}_m^*, \text{full}} | m - |\underline{t}_m^*|) \geq C(\Lambda_{n, \underline{t}_m^*, \text{full}}, m - |\underline{t}_m^*|)$$

for $\Lambda_{n, \underline{t}_m^*, \text{full}} > 0$ and $m - |\underline{t}_m^*| > 1$ and

$$\begin{aligned} p(\Lambda_{n, \underline{i}, \text{full}} | m - |\underline{i}|) &\leq C(\Lambda_{n, \underline{i}, \text{full}}, m - |\underline{i}|) \left(1 + \frac{m - |\underline{i}| - 2}{\Lambda_{n, \underline{i}, \text{full}} - m + |\underline{i}| + 2} \right) \\ &= C(\Lambda_{n, \underline{i}, \text{full}}, m - |\underline{i}|) (1 + o_P(1)) \end{aligned}$$

for $\Lambda_{n, \underline{i}, \text{full}} \rightarrow \infty$ and $m - |\underline{i}| > 1$. Analogous inequality holds for $\underline{i}$ such that $m - |\underline{i}| = 1$. This together with (19) and assumption $a_n/n \rightarrow 0$ imply that when $m - |\underline{t}_m^*| > 1$

$$P \left(\max_{\underline{i}: \underline{t}_m^* \not\subset \underline{i} \subset \{1, \dots, m\}} p(\Lambda_{n, \underline{i}, \text{full}} | m - |\underline{i}|) e^{-a_n |\underline{i}|} < p(\Lambda_{n, \underline{t}_m^*, \text{full}} | m - |\underline{t}_m^*|) e^{-a_n |\underline{t}_m^*|} \right) \rightarrow 1.$$

The proof in the cases $1 \geq m - |\underline{t}_m^*| \geq 0$ is analogous. $\square$

Theorem 6 (MPVC consistency) *Assume (A1)–(A6), $|\underline{t}| < \infty$ and $\lim_{n \rightarrow \infty} m_n \geq \max \underline{t}$. Then*

- (a) $\lim_{n \rightarrow \infty} P(\underline{t} \subset \hat{\underline{t}}_{\text{MPVC}}) = 1$,
- (b) if $m_n \log m_n = o(a_n)$ then $\lim_{n \rightarrow \infty} P(\hat{\underline{t}}_{\text{MPVC}} = \underline{t}) = 1$.

Proof We first consider the case when model $\underline{i}$ is misspecified: $\underline{t} \not\subset \underline{i}$ and $\underline{t} \neq \text{full}$. Observe that Lemma 3(i) implies that $\max_{\underline{i} \subset \{1, \dots, m_n\}} |\Lambda_{n, \underline{i}, \text{full}} / (2n) - (\lambda_{\underline{t}} - \lambda_{\underline{i}})| \xrightarrow{\mathcal{P}} 0$ and thus it follows in view of Lemma 5 that $\min_{\underline{i} \subset \{1, \dots, m_n\}: \underline{t} \not\subset \underline{i}} \Lambda_{n, \underline{i}, \text{full}} \rightarrow \infty$ in

probability. Assume first that $m_n - |\underline{i}| > 1$. Whence using Lemma 6(ii) it is enough to show that with probability tending to 1 we have

$$\begin{aligned} & e^{-a_n|\underline{t}|} C(\Lambda_{n,\underline{t},\text{full}}, m_n - |\underline{t}|) \\ & > e^{-a_n|\underline{i}|} C(\Lambda_{n,\underline{i},\text{full}}, m_n - |\underline{i}|) \left(1 + \frac{m_n - |\underline{i}| - 2}{\Lambda_{n,\underline{i},\text{full}} - m_n + |\underline{i}| + 2} \right) \end{aligned} \quad (20)$$

for all $\underline{i} \subset \{1, \dots, m_n\} : \underline{t} \not\subset \underline{i}$ and such that $m_n - |\underline{i}| > 1$. This easily follows as in the case of mPVC consistency from

$$P \left(\min_{\underline{i}: \underline{t} \not\subset \underline{i} \subset \{1, \dots, m_n\}} \frac{1}{2n} \Lambda_{n,\underline{i},\text{full}} - \frac{1}{2n} \Lambda_{n,\underline{t},\text{full}} > \varepsilon \right) \rightarrow 1 \quad (21)$$

for some $\varepsilon > 0$. The proof for $m_n - |\underline{i}| = 1$ is analogous. The case $\underline{t} = \text{full}$ follows easily from $a_n/n \rightarrow 0$ after noting that in this case the right hand side of (20) is $O(e^{-\varepsilon n})$ for some $\varepsilon > 0$, whereas the left hand side equals $e^{-a_n|\underline{t}|}$. This yields the proof of (a).

We now consider the case when model $\underline{i}$ contains the minimal true model: $\underline{t} \subsetneq \underline{i}$. As in the proof of Theorem 5 we obtain

$$\begin{aligned} & P \left(\max_{\underline{i}: \underline{t} \subset \underline{i} \subset \{1, \dots, m_n\}} p(\Lambda_{n,\underline{i},\text{full}} | m_n - |\underline{i}|) e^{-a_n|\underline{i}|} < p(\Lambda_{n,\underline{t},\text{full}} | m_n - |\underline{t}|) e^{-a_n|\underline{t}|} \right) \\ & \geq P \left(p(\Lambda_{n,\underline{t},\text{full}} | m_n - |\underline{t}|) > e^{-a_n} \right). \end{aligned}$$

Observe that in view of Lemma 4 $\Lambda_{n,\underline{t},\text{full}} = O_P(m_n)$. Thus it is enough to show that

$$e^{a_n} P(Z_{\chi_{m_n-|\underline{t}|}^2}^2 > C m_n) \rightarrow \infty \quad (22)$$

for any fixed $C > 0$, where $Z_{\chi_{m_n-|\underline{t}|}^2}^2 \sim \chi_{m_n-|\underline{t}|}^2$. This follows easily from Lemma 6 and the assumed condition on m_n as the above expression is bounded from below by

$$\exp \left\{ a_n - \frac{C m_n}{2} + \left(\frac{\bar{m}_n}{2} - 1 \right) \log \left(\frac{C m_n}{2} \right) - \log \left(\frac{\bar{m}_n}{2} \right) \frac{\bar{m}_n}{2} \right\} \rightarrow \infty,$$

where $\bar{m}_n = m_n - |\underline{t}|$. This yields the proof of (b). $\square$

Remark 4 Note that the assumption of Theorem 6 (b) implies that $a_n \rightarrow \infty$ whereas Theorem 2 asserts that consistency of mPVC criterion holds also for constant a_n provided $m_n = o(\log n)$.

From the proof of Theorem 6 (cf. (21)) it is easy to see that the analogue of Theorem 3 holds for MPVC criterion under the same conditions.

3.3 Greedy mPVC and MPVC criteria and their consistency

Optimization of a criterion function over all subsets of $\{1, 2, \dots, m_n\}$ involves a considerable computational cost for large m_n . This is a drawback of all criterion based procedures. We discuss here a two-step modification of p -value criteria introduced in Pokarowski and Mielniczuk (2011) which involves only $O(m_n)$ calculations of criterion instead of $O(2^{m_n})$. The approach is motivated by Zheng and Loh (1997) who used such an approach for linear models.

Greedy mPVC In the first step for every $j \in \{1, \dots, m_n\}$ we consider testing hypotheses

$$\begin{aligned} H_0 &: f \in \mathcal{M}_0 \\ H_1 &: f \in \mathcal{M}_{\bar{j}} \setminus \mathcal{M}_0, \end{aligned}$$

where $\bar{j} = \{1, \dots, m_n\} \setminus \{j\}$. Then we order the variables with respect to the obtained p -values:

$$p(\Lambda_{n,0,\bar{j}_1} | m_n - 1) \geq \dots \geq p(\Lambda_{n,0,\bar{j}_m} | m_n - 1)$$

and apply mPVC method considering only nested list of m models indexed by the elements of $M = \{\emptyset, \{j_1\}, \{j_1, j_2\}, \dots, \{j_1, \dots, j_m\}\}$. Thus the greedy minimal p -value criterion takes the form:

$$\hat{l}_{\text{mPVC}}^{\text{greedy}} = \arg \min_{\underline{i} \in M} p(\Lambda_{n,0,\underline{i}} | |\underline{i}|) e^{a_n |\underline{i}|}.$$

Greedy MPVC In the first step for every $j \in \{1, \dots, m_n\}$ we consider test of the form

$$\begin{aligned} H_0 &: f \in \mathcal{M}_{\bar{j}} \\ H_1 &: f \in \mathcal{M}_{\text{full}} \setminus \mathcal{M}_{\bar{j}}, \end{aligned}$$

where $\bar{j} = \{1, \dots, m_n\} \setminus \{j\}$. Then we order the variables with respect to the obtained p -values:

$$p(\Lambda_{n,\bar{j}_1,\text{full}} | 1) \leq \dots \leq p(\Lambda_{n,\bar{j}_m,\text{full}} | 1)$$

and apply MPVC method considering only nested list of m_n models indexed by the elements of $M = \{\emptyset, \{j_1\}, \{j_1, j_2\}, \dots, \{j_1, \dots, j_m\}\}$. Thus the greedy maximal p -value criterion takes the form:

$$\hat{l}_{\text{MPVC}}^{\text{greedy}} = \arg \max_{\underline{i} \in M} p(\Lambda_{n,\underline{i},\text{full}} | m_n - |\underline{i}|) e^{-a_n |\underline{i}|}.$$

Observe that in both greedy methods all m likelihood ratio test statistics LRT have under the null hypothesis the same asymptotic distribution, namely χ_{m-1}^2 in the case of

Greedy mPVC and χ_1^2 in the case of Greedy MPVC. Hence in both cases ordering the p -values coincides with ordering the values of test statistics monotonically. Moreover, it follows that the sets M yielded by the both procedures are equal.

It is also of interest to note that selection methods based on truncation $\hat{\underline{t}} = \{j : |\hat{\theta}_j^{\text{full}}| > C_n\}$ can be also viewed as two-step procedures similar to these defined above. Namely, in the first step estimates of the parameters in the full model are ordered: $|\hat{\theta}_{R_1}^{\text{full}}| \geq |\hat{\theta}_{R_2}^{\text{full}}| \cdots \geq |\hat{\theta}_{R_m}^{\text{full}}|$. Then we choose a model $\{R_1, R_2, \dots, R_{k_0}\}$ such that $k_0 = \operatorname{argmax}_k \{\sum_{i=1}^k (\hat{\theta}_{R_i}^{\text{full}})^2 - kC_n^2\}$. It is easy to see that $\hat{\underline{t}} = \{R_1, R_2, \dots, R_{k_0}\}$. Observe that the criterion function in the second step is a penalized score statistic $\sum_{i=1}^k (\hat{\theta}_{R_i}^{\text{full}})^2$ for the model indexed by $\{R_1, R_2, \dots, R_k\}$. We refer to [Wojtyś \(2011\)](#) for discussion of such rules.

Corollary 1 *Under conditions of Theorem 2 (b) and Theorem 6 (b), respectively, greedy mPVC and MPVC methods are consistent.*

Proof For every $j \notin \underline{t}$ we have $\Lambda_{n,0,\underline{t}} \leq \Lambda_{n,0,\bar{j}}$. Thus as for $k \in \underline{t}$ we have $\bar{k} \not\supset \underline{t}$ (13) implies that for every such k

$$\lim_{n \rightarrow \infty} P \left(\min_{j \notin \underline{t}} \frac{1}{2n} \Lambda_{n,0,\bar{j}} > \frac{1}{2n} \Lambda_{n,0,\bar{k}} \right) = 1.$$

Thus with probability tending to 1 after the initial ordering all indices belonging to set $\underline{t}$ will precede the remaining indices. Hence the consistency of both greedy methods follows from the consistency of their respective full search method applied to M . $\square$

Remark 5 It can be shown using Lemma 3(i) that under its assumptions probability of incorrect ordering in the first step of the procedure is bounded by

$$P(\min_{i \notin \underline{t}} \Lambda_{n,0,\bar{i}} < \max_{j \in \underline{t}} \Lambda_{n,0,\bar{j}}) \leq C \left(\frac{m^{1+2\omega}}{n} \right) (\lambda^{-1} + \lambda^{-2}),$$

where $\lambda = \min_{i \notin \underline{t}, j \in \underline{t}} (\lambda_{\bar{i}} - \lambda_{\bar{j}})$ and C is an absolute constant. Note that the bound becomes larger when the selection problem becomes more difficult, i.e. λ decreases.

4 Simulation study

We conducted numerical experiments to check how the considered selectors behave in practice for moderate sample sizes. The considered sample size was $n = 300$ with $m_n = 5$. Random samples were generated either from the uniform density or belonged to one of the following four families:

- (i) exponential family $\mathcal{M}_1$ with $\underline{t} = \{1\}$ and $\theta_1 = 0.3, 0.4, \dots, 1$;
- (ii) exponential family $\mathcal{M}_3$ with $\underline{t} = \{3\}$ and $\theta_3 = 0.3, 0.4, \dots, 1$;
- (iii) exponential family $\mathcal{M}_5$ with $\underline{t} = \{1, 5\}$, $\theta_1 = 0.1$ and $\theta_5 = 0.1, \dots, 0.5$;
- (iv) beta densities with parameters $(a, b) = (1, 1.2), (1.1, 1.3), (1.5, 1.5), (2, 2)$.

Fig. 1 Theoretical densities

Recall that e.g. $\underline{t} = \{1, 5\}$ means that only the first and the fifth coefficient is allowed to be nonzero. In all cases we considered Legendre polynomials basis. The densities are plotted in Fig. 1. The values of the parameters were chosen to obtain typical shapes in the considered family. Number of Legendre polynomials appearing in the definitions of densities (i)–(iii) corresponds to shape complexity. Observe e.g. that typical density in example (ii) has two modes in $(0,1)$ in contrast to four modes in the case (iii). Example (iv) corresponds to misspecification case, i.e. the situation when the underlying density does not belong to any model in the considered family.

Six selection criteria were taken into account: BIC (Bayesian Information Criterion, Schwarz (1978)), mPVC, MPVC and their greedy counterparts. BIC is defined as

$$\text{BIC}(\underline{i}) = 2\Lambda_{n,0,\underline{i}} - |\underline{i}| \log n.$$

Several values of scaling constants a_n for p -value based criteria were considered. We report the results for constants which performed the best on average: $a_n = 0$ for mPVC and $a_n = \log n/2$ for MPVC, i.e. an unscaled $p(\Lambda_{n,0,\underline{l}}||\underline{l})$ for mPVC and $p(\Lambda_{n,\underline{l},\text{full}}|m - |\underline{l}|)$ scaled down by $n^{-1/2}$ for MPVC.

For three main selectors ML estimators had to be calculated for any of 2^m models or, after initial preordering, for m models in the case of their greedy counterparts. As ML estimators are calculated using iterative Newton-Raphson procedure some cases of non-convergence occur. In such a case corresponding model was excluded from the list and only the models for which ML estimators were obtained were considered for optimization. Number of cases when lack of convergence occurs increases with complexity of the model, however it does not exceed 3 % of the number of the models considered.

Two main measures of performance were taken into account: fraction of correct model specifications averaged over 10^4 repetitions of the experiment and averaged empirical Integrated Squared Error ISE defined as

$$\text{ISE}(\hat{f}) = \int_0^1 (f(x) - \hat{f}(x))^2 dx,$$

where f is the theoretical density of random sample and $\hat{f}$ is its post-model-selection estimator. Mean of ISE is denoted as MISE in the plots. It turned out that (see Figs. 2, 3, 4, 5) these two measures are approximately concordant in the sense that a large probability of correct specification corresponds to a small ISE and the ranking of selectors with respect to both measures coincide in general. The only exception are the densities close to $\mathcal{M}_0$ (e.g. members of $\mathcal{M}_1$ with small θ_1) for which a method may err by choosing the uniform density but still has a small MISE. For misspecification case the accuracy percent was replaced by averaged Kullback–Leibler distance $D(f||\hat{f})$ from f to post-model-selection estimator. Both integrals, ISE and Kullback–Leibler

Fig. 2 Percentage of correct model selections (*left graph*) and estimated MISE (*right graph*) for data pertaining to $f_\theta(x) = c(\theta) \exp\{\theta_1 b_1(x)\}$, where $\theta_1 = 0.3, 0.2, \dots, 1$

Fig. 3 Percentage of correct model selections (*left graph*) and estimated MISE (*right graph*) for data pertaining to $f_\theta(x) = c(\theta) \exp\{\theta_3 b_3(x)\}$, where $\theta_3 = 0.3, 0.4, \dots, 1$

Fig. 4 Percentage of correct model selections (*left graph*) and estimated MISE (*right graph*) for data pertaining to $f_\theta(x) = c(\theta) \exp\{\theta_1 b_1(x) + \theta_5 b_5(x)\}$, where $\theta_1 = 0.1, \theta_5 = 0.1, \dots, 0.5$

distance, were calculated numerically using the Gaussian quadrature for Legendre polynomials. In order not to obscure the clarity of pictures in each example we plotted the best overall performing selector and depending on whether it was the main selector or one of the greedy modification we supplemented it by the two remaining selectors from the same group.

We start with the comparison of methods for data pertaining to the uniform density presented in Table 1. In this case the best performance was achieved by greedy MPVC, which selected the correct model in more than 99 % of Monte Carlo repetitions and yielded the smallest MISE. Also MPVC had comparably good properties as it attained nearly 99 % accuracy of model selection and had MISE three times as large as the winner. BIC criterion achieved more than 90 % of correct model selections, however

Fig. 5 Estimated Kullback–Leibler distance from theoretical density $D(f||\hat{f})$ (left graph) and MISE (right graph) and for data pertaining to Beta distribution with parameters a and b

the estimated MISE was 10 times larger than for greedy MPVC. The methods mPVC and greedy mPVC behaved considerably worse.

This may come as a surprise that the greedy method may actually perform the best even when optimization over all subsets is taken into account. It happens for the simplest densities considered (case (i), cf. Fig. 2) where greedy MPVC performs most favorably. For the other case with only one nonzero coefficient (example (ii), cf. Fig. 3) its non-greedy counterpart is the best. The similar situation occurs in the misspecification case (example (iv)). However, for the most complex shapes of densities (example (iii), cf. Fig. 4) minimal p -value criterion mPVC was a clear winner. This case is clearly the most difficult among cases considered as accuracy percent falls below 0.4 and it is here where the improvement is most desirable. Note that in this case the gain of the winner over its BIC competitor (greedy or non-greedy) was more pronounced in the terms identification of the correct model selection than for accuracy of estimator measured by MISE. This happened especially for the larger values of the considered parameters. Thus it seems that the main advantage of the introduced methods is in terms of increase of probability of model specification rather than increase of estimation accuracy of density estimator measured by MISE.

Figure 6 shows the probability of correct ordering in the first step of greedy methods for model (i) as a function of m . In this case in order to reduce computational

Table 1 Estimated $\text{MISE}(\hat{f})$ and probabilities of correct model selections for the data pertaining to the uniform distribution $U(0, 1)$

$\hat{t}$	$\widehat{\text{MISE}}(\hat{f})$	$\hat{P}(\hat{t} = t)$
BIC	22×10^{-4}	0.913
$\hat{t}_{\text{mPVC}}$	64×10^{-4}	0.719
$\hat{t}_{\text{MPVC}}$	6×10^{-4}	0.985
$\hat{t}_{\text{mPVC}}^{\text{greedy}}$	50×10^{-4}	0.814
$\hat{t}_{\text{MPVC}}^{\text{greedy}}$	2×10^{-4}	0.993

Fig. 6 Estimated probability of correct ordering of variables in the first step of greedy methods (dashed line) compared to percentages of correct model selections for these methods (solid lines) for data pertaining to $f_\theta(x) = c(\theta) \exp\{\theta_1 b_1(x)\}$ with $\theta_1 = 0.5$

cost of simulations the sample size was taken as $n = 100$. The plotted curves indicate that probability of correct ordering depends heavily on the size of the list of models, however, once the correct order is established it is relatively unlikely to commit errors in the second step using p -value based methods or BIC. This underlines the importance of choosing a small subfamily of models before optimizing the criterion.

We also recorded the cases when a supermodel, that is a subset $\underline{i} \supset \underline{t}$ instead of $\underline{t}$, is selected. It turns out that greedy methods in general are more conservative than their full-search counterparts, i.e. much more prone to choosing supermodel. Moreover, mPVC is much more conservative than MPVC; for the first method in some cases fraction of supermodels chosen is around 30 %, whereas in the case of MPVC it is always below 3 %. This is possibly due to penalty a_n which equals 0 for mPVC.

Finally, note that in all the cases considered one of the proposed method performed better than its BIC-like competitor. The natural question being a topic of current research is whether some combined version of p -value based criteria would exhibit better performance overall.

Acknowledgments Comments of an anonymous referee which helped to improve the presentation in the paper are gratefully acknowledged. In numerical experiments we used a computer program for calculating ML estimators in exponential families which is an extension of the routine kindly supplied by T. Ledwina.

Open Access This article is distributed under the terms of the Creative Commons Attribution License which permits any use, distribution, and reproduction in any medium, provided the original author(s) and the source are credited.

Appendix

Proof of Lemma 1 Lemma 1 in [Barron and Sheu \(1991\)](#) implies

$$D(f||f_{\underline{i}}^*) \geq \frac{1}{2} e^{-||\log(f/f_{\underline{i}}^*)||_{\infty}} \int_0^1 f(x) (\log(f(x)/f_{\underline{i}}^*(x)))^2 dx.$$

It follows from (A2) and (A5) that there exists a constant $C_1 > 0$ such that $\max_{\underline{i} \subset \{1, \dots, m_n\}} e^{-||\log(f/f_{\underline{i}}^*)||_{\infty}} \geq C_1$ for all n . Let $|| \cdot ||$ denote the norm in l^2 . Finite dimensional vectors are embedded in l^2 by augmentation with zeros. Orthonormality of the system $(b_i)_{i=0}^{\infty}$ in $L^2([0, 1], \lambda)$ yields for $C_2 = \frac{1}{2} C_1$

$$\begin{aligned} D(f||f_{\underline{i}}^*) &\geq C_2 \inf_{x \in [0, 1]} f(x) \int_0^1 (\log(f(x)/f_{\underline{i}}^*(x)))^2 dx \\ &= C(||\theta - \theta_{\underline{i}}^*||^2 + (\psi(\theta) - \psi_{\underline{i}}(\theta_{\underline{i}}^*))^2), \end{aligned} \quad (23)$$

where $C = C_2 \inf_{x \in [0, 1]} f(x)$. Assumption (A5) implies $C > 0$. Now the assertion follows from the fact that the left hand side is finite as $D(f||f_{\underline{i}}^*)$ is not greater than $D(f||f_0) = \int_0^1 f(x) \log f(x) dx < \infty$ and from the fact that in view of (A6) we have $||\theta|| < \infty$. $\square$

Proof of Lemma 2 For every $\underline{i} \subset \{1, \dots, m_n\}$ we apply Lemma 7 with $q = \lambda$, $A_{m_n}(\lambda) = \sqrt{m_n + 1} V_{m_n}$, $\theta_0 = \theta_{\underline{i}}^*$, $\alpha_0 = \mathbb{E}_{f_{\underline{i}}^*} b^{\underline{i}}(X) = \mathbb{E}_f b^{\underline{i}}(X)$, $\alpha = \bar{b}_n^{\underline{i}}$ and $c = e^{||\log f_{\underline{i}}^*||_{\infty}}$. Let $\bar{b}_n^{m_n} = \bar{b}_n^{\{1, \dots, m_n\}}$ and $b^{m_n} = b^{\{1, \dots, m_n\}} = (b_1, \dots, b_{m_n})$. Markov's inequality yields $||\bar{b}_n^{m_n} - \mathbb{E}_f b^{m_n}(X)|| = O_P((m_n^{1+2\omega}/n)^{1/2})$. Thus equality

$$\max_{\underline{i} \subset \{1, \dots, m_n\}} ||\bar{b}_n^{\underline{i}} - \mathbb{E}_f b^{\underline{i}}(X)|| = ||\bar{b}_n^{m_n} - \mathbb{E}_f b^{m_n}(X)|| \quad (24)$$

implies

$$\max_{\underline{i} \subset \{1, \dots, m_n\}} ||\bar{b}_n^{\underline{i}} - \mathbb{E}_f b^{\underline{i}}(X)|| = O_P\left((m_n^{1+2\omega}/n)^{1/2}\right). \quad (25)$$

Now assumptions (A2) and (A4) yield (26) for all $\underline{i} \subset \{1, \dots, m_n\}$ on the set of probability tending to 1 as $n \rightarrow \infty$.

Hence Lemma 7 and conditions (A2) and (24) imply existence of a constant $C > 0$ such that

$$\max_{\underline{i} \subset \{1, \dots, m_n\}} D(f_{\hat{\theta}_{\underline{i}}^*} || f_{\hat{\theta}_{\text{NW}}^*}) \leq C ||\bar{b}_n^{m_n} - \mathbb{E}_f b^{m_n}(X)||^2$$

and

$$\max_{\underline{i} \subset \{1, \dots, m_n\}} D(f_{\hat{\theta}_{\text{NW}}^*} || f_{\hat{\theta}_{\underline{i}}^*}) \leq C ||\bar{b}_n^{m_n} - \mathbb{E}_f b^{m_n}(X)||^2.$$

Thus in view of (25) the conclusion follows. $\square$

Proof of Lemma 3 In view of (6) we have

$$\begin{aligned} \left| \frac{1}{2n} \Lambda_{n,0,\underline{i}} - \lambda_{\underline{i}} \right| &= \left| (\hat{\theta}_{\text{ML}}^{\underline{i}} - \theta_{\underline{i}}^*) \circ \bar{b}_{\underline{n}}^{\underline{i}} - \psi_{\underline{i}}(\hat{\theta}_{\text{ML}}^{\underline{i}}) + \psi_{\underline{i}}(\theta_{\underline{i}}^*) + \theta_{\underline{i}}^* \circ (\bar{b}_{\underline{n}}^{\underline{i}} - E_f(b^{\underline{i}}(X))) \right| \\ &= \left| D(\hat{f}_{\underline{i}} \| f_{\underline{i}}^*) + \theta_{\underline{i}}^* \circ (\bar{b}_{\underline{n}}^{\underline{i}} - E_f(b^{\underline{i}}(X))) \right| \\ &\leq D(\hat{f}_{\underline{i}} \| f_{\underline{i}}^*) + \|\theta_{\underline{i}}^*\| \|\bar{b}_{\underline{n}}^{\underline{i}} - E_f(b^{\underline{i}}(X))\|. \end{aligned}$$

Lemma 2 yields $\max_{\underline{i} \subset \{1, \dots, m_n\}} D(\hat{f}_{\underline{i}} \| f_{\underline{i}}^*) = O_P(m_n^{1+2\omega}/n)$. If $\underline{i} \subset \{1, \dots, m_n\}$ then in view of (25) and Lemma 1 $\max_{\underline{i} \subset \{1, \dots, m_n\}} \left| \frac{1}{2n} \Lambda_{n,0,\underline{i}} - \lambda_{\underline{i}} \right| = O_P((m_n^{1+2\omega}/n)^{1/2})$.

Properties (ii) and (iii) follow from the uniqueness of information projection and the fact that $\lambda_{\underline{i}} = -D(f \| f_{\underline{i}}^*) + \int f \log f$. $\square$

Proof of Lemma 4 Recall that $m = m_n$. For $\theta \in \mathbb{R}^m$ define $\Lambda_n(\theta) = 2 \sum_{j=1}^n \log f_{\theta}(X_j)$. Note that if $\underline{t}_m^* \subset \underline{i} \subset \{1, \dots, m_n\}$ then $f_{\underline{i}}^* = f_{\underline{t}_m^*}^*$. Thus $\Lambda_n(\theta_{\underline{i}}^*) = \Lambda_n(\theta_{\underline{t}_m^*}^*)$ where the vectors $\theta_{\underline{i}}^*$ and $\theta_{\underline{t}_m^*}^*$ are completed by zeros at indices $\{1, \dots, m_n\} \setminus \underline{i}$ and $\{1, \dots, m_n\} \setminus \underline{t}_m^*$, respectively.

Then

$$\Lambda_{n,0,\underline{i}} - \Lambda_{n,0,\underline{t}_m^*} = (\Lambda_{n,0,\underline{i}} - \Lambda_n(\theta_{\underline{i}}^*)) - (\Lambda_{n,0,\underline{t}_m^*} - \Lambda_n(\theta_{\underline{t}_m^*}^*)).$$

We have

$$\Lambda_{n,0,\text{full}} - \Lambda_n(\theta_{\text{full}}^*) = 2n D(\hat{f}_{\text{full}} \| f_{\text{full}}^*)$$

and

$$\Lambda_{n,0,\underline{t}_m^*} - \Lambda_n(\theta_{\underline{t}_m^*}^*) = 2n D(\hat{f}_{\underline{t}_m^*} \| f_{\underline{t}_m^*}^*).$$

Let $(\tilde{b}_j)_{j=0}^m$ with $\tilde{b}_0 \equiv 1$ be an orthonormal system in $L^2([0, 1], f)$ such that a linear space $sp\{b_j, j \in \underline{t}_m^*\} = sp\{\tilde{b}_j, j \in \underline{t}_m^*\}$ and $sp\{b_j, j \in \{0, 1, \dots, m\}\} = sp\{\tilde{b}_j, j \in \{0, 1, \dots, m\}\}$. Such system is constructed by performing orthonormalisation of $sp\{b_j, j \in \underline{t}_m^*\}$ w.r.t. f first which ensures $sp\{b_j, j \in \underline{t}_m^*\} = sp\{\tilde{b}_j, j \in \underline{t}_m^*\}$ and then continuing the process for b_i with $i \in \{0, 1, \dots, m\} \setminus \underline{t}_m^*$, which yields the second equality. Note that it follows from the Markov inequality that $\|\tilde{b}_n^m - \mathbb{E}_f \tilde{b}^m(X)\| = O_P((m/n)^{1/2})$ as $E_f \tilde{b}_j(X) = 0$ and $E_f \tilde{b}_j^2(X) = 1$ for $j = 1, 2, \dots$. Then in the view of this and using Lemma 7 analogously as in Lemma 2 we have that $D(\hat{f}_m \| f_m^*) = O_P(m/n)$ and $D(\hat{f}_{\underline{t}_m^*} \| f_{\underline{t}_m^*}^*) = O_P(m/n)$. Thus

$$\begin{aligned} \max_{\underline{i}: \underline{t}_m^* \subset \underline{i} \subset \{1, \dots, m_n\}} |\Lambda_{n,0,\underline{i}} - \Lambda_{n,0,\underline{t}_m^*}| &= \Lambda_{n,0,\text{full}} - \Lambda_{n,0,\underline{t}_m^*} \\ &\leq 2n \left(D(\hat{f}_{\text{full}} \| f_{\text{full}}^*) + D(\hat{f}_{\underline{t}_m^*} \| f_{\underline{t}_m^*}^*) \right) \\ &= O_P(m_n). \end{aligned}$$

$\square$

Proof of Lemma 5 Lemma 1 in [Barron and Sheu \(1991\)](#) yields for any $\underline{i} \subset \mathbb{N}$

$$D(f||f_{\underline{i}}^*) \geq \frac{1}{2} e^{-||\log(f/f_{\underline{i}}^*)||_{\infty}} \int_0^1 f(\log(f/f_{\underline{i}}^*))^2.$$

Assumptions (A2) and (A5) imply that there exists a constant $C > 0$ such that $D(f||f_{\underline{i}}^*) \geq C \int_0^1 (\log(f/f_{\underline{i}}^*))^2$. Note that orthogonality of $(b_j)_{j=0}^{\infty}$ and $\theta \in l^2$ yield

$$\int_0^1 (\log(f/f_{\underline{i}}^*))^2 = \sum_{j \in \underline{i}} (\theta_j - \theta_{i,j}^*)^2 + \sum_{j \notin \underline{i}} \theta_j^2 + (\psi(\theta) - \psi_{\underline{i}}(\theta_{\underline{i}}^*))^2 \geq \min_{j \in \underline{i}} \theta_j^2.$$

Thus $D(f||f_{\underline{i}}^*) > C \min_{j \in \underline{i}} \theta_j^2$ which proves the lemma. $\square$

Lemma 7 Let $(b_j(x))_{j=0}^m$ be an orthonormal system in $L^2([0, 1], q)$, where q is some measure on $[0, 1]$, and let $A_m(q)$ be a constant such that $||\log g||_{\infty} \leq A_m(q) ||\log g||_{L^2(q)}$ for all $g \in \mathcal{M}_m$. Let $\theta_0 \in \mathbb{R}^m$, $\alpha_0 = \int b f_{\theta_0}$ and $\alpha \in \mathbb{R}^m$ be given. Let $c = e^{||\log q/f_{\theta_0}||_{\infty}}$. If

$$||\alpha - \alpha_0|| \leq \frac{1}{4ecA_m(q)}, \quad (26)$$

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

$$||\theta(\alpha) - \theta_0|| \leq 2ec||\alpha - \alpha_0||, \quad (27)$$

$$||\log f_{\theta_0}/f_{\theta(\alpha)}||_{\infty} \leq 4ecA_m(q)||\alpha - \alpha_0|| \leq 1, \quad (28)$$

$$D(f_{\theta_0}||f_{\theta(\alpha)}) \leq 2ec||\alpha - \alpha_0||^2 \quad (29)$$

and

$$D(f_{\theta(\alpha)}||f_{\theta_0}) \leq 2ec||\alpha - \alpha_0||^2. \quad (30)$$

Proof The existence of $\theta(\alpha)$ and inequalities (27), (28) and (29) are proved in Lemma 5 in [Barron and Sheu \(1991\)](#) (see also Lemma 2 in [Mielniczuk and Wojtyś \(2010\)](#)). As to inequality (30) we have

$$\begin{aligned} D(f_{\theta(\alpha)}||f_{\theta_0}) &= (\theta(\alpha) - \theta_0)^\circ (\alpha - \alpha_0) - D(f_{\theta_0}||f_{\theta(\alpha)}) \\ &\leq (\theta(\alpha) - \theta_0)^\circ (\alpha - \alpha_0) \leq ||\theta(\alpha) - \theta_0|| ||\alpha - \alpha_0|| \leq 2ec||\alpha - \alpha_0||^2, \end{aligned}$$

where the last inequality is implied by (27). $\square$

Let us note that if the measure q in the above lemma is the Lebesgue measure λ on $[0, 1]$ then $A_m(\lambda) = \sqrt{m} V_m$, where $V_m = \max_{j=1, \dots, m} ||b_j||_{\infty}$. If q is induced by a density f then $A_m(f) = A_m(\lambda)/\inf_x |f(x)|^{1/2}$.

References

- Barron AR, Sheu CH (1991) Approximation of density functions by sequences of exponential families. *Ann Stat* 19:1347–1369
- Cover T, Thomas J (2006) Elements of information theory, 2nd edn. Wiley, New York
- Gordon RD (1941) Values of Mills' ratio of area to bounding ordinate and of the normal probability integral for large values of the argument. *Ann Math Stat* 12:364–366
- Ingolot T, Ledwina T (1996) Asymptotic optimality of data-driven Neyman's tests for uniformity. *Ann Stat* 24:1982–2019
- Ingolot T, Ledwina T (2006) Asymptotic optimality of new adaptive test in regression model. *Annales de l'Institut Henri Poincaré. Probab Stat* 42:579–590
- Kallenberg WCM, Ledwina T (1997) Data driven smooth test when the hypothesis is composite. *J Am Stat Assoc* 2:1094–1104
- Ledwina T (1994) Data-driven version of Neyman's smooth test of fit. *J Am Stat Assoc* 89:1000–1005
- Mielniczuk J, Wojtyś M (2010) Estimation of Fisher information using model selection. *Metrika* 72:163–187
- Pokarowski P, Mielniczuk J (2011) P-values of likelihood ratio statistics for consistent model selection and testing. Unpublished manuscript
- Schwarz G (1978) Estimating the dimension of a model. *Ann Stat* 6:461–464
- Sen PK, Singer JM (1993) Large sample methods in statistics. An introduction with applications. Chapman and Hall, London
- van der Vaart AW (2000) Asymptotic statistics. Cambridge University Press, Cambridge
- Wainwright MJ, Jordan MI (2008) Graphical models, exponential families and variational inference. *Found Trends Mach Learn* 1:1–305
- Wojtyś M (2011) Post-model-selection method for density estimation. *Commun Stat Theory Methods* 40:3082–3098
- Yurinski VV (1976) Exponential inequalities for sums of random vectors. *J Multivar Anal* 6:473–499
- Zheng X, Loh W-Y (1997) A consistent variable selection criterion for linear models with high dimensional covariates. *Stat Sin* 7:311–325