Consistent Order Estimation and Minimal Penalties

Elisabeth Gassiat and Ramon van Handel

***Abstract*—Consider an i.i.d. sequence of random variables whose distribution $f^\star$ lies in one of the nested families of models $\mathcal{M}_q$, $q \geq 1$. The smallest index $q^\star$ such that $\mathcal{M}_{q^\star}$ contains $f^\star$ is called the model order. The aim of this paper is to explore the consistency properties of penalized likelihood model order estimators such as Bayesian information criterion. We show in a general setting that the minimal strongly consistent penalty is of order $\eta(q)\log\log n$, where $\eta(q)$ is a dimensional quantity. In contrast to previous work, an *a priori* upper bound on the model order is not assumed. The results rely on a sharp characterization of the pathwise fluctuations of the generalized likelihood ratio statistic under entropy assumptions on the model classes. Our results are applied to the geometrically complex problem of location mixture order estimation, which is widely used but poorly understood.**

***Index Terms*—Consistent order estimation, location mixtures, penalized likelihood, uniform law of iterated logarithm.**

I. Introduction

LET $(X_k)_{k\geq 1}$ be a sequence of random variables whose distribution $f^\star$ lies in one of the nested families of models $(\mathcal{M}_q)_{q\geq 1}$, indexed (and ordered) by the integers. We define the model order as the smallest index $q^\star$ such that the true distribution $f^\star$ lies in the corresponding model class. The model order typically determines the most parsimonious representation of the true distribution of the underlying model (for example, it might determine the parametrization of the model which has the smallest possible dimension). On the other hand, the model order often has a concrete interpretation in terms of the modeling of the underlying phenomenon (for example, the estimation of the number of clusters in a dataset, or the number of regimes in an economic time series). Therefore, the problem of estimating the model order from observed data is of significant practical, as well as theoretical, interest.

Of course, a satisfactory solution to this problem must provide an estimation method that does not assume prior knowledge on the unknown distribution $f^\star$. In particular, prior bounds on model order and on parameter sets should be avoided. Yet, in this light, even one of the most widely used model selection criteria—the Bayesian information criterion (BIC) of Schwarz—is poorly understood. The chief motivation for the use of BIC (as opposed to other model selection criteria, such as Akaike's information criterion) is that it is expected to yield a strongly consistent estimator of the model order. However, almost all existing consistency proofs assume a prior upper bound on the order as well as compactness of the parameter set. As is emphasized by Csiszár and Shields [1], this is hardly satisfactory from the theoretical point of view and provides little confidence in the basic motivation for this method. More delicate questions, such as the minimal penalty that yields a consistent order estimator in absence of a prior bound on the order, remain open (the problem of identifying the minimal penalty, which minimizes the probability of underestimating the order, is also raised in [1]).

In this paper, we consider a general class of penalized likelihood order estimators of the form

$$\hat{q}_n = \operatorname*{argmax}_{q\geq 1} \left\{ \sup_{f\in\mathcal{M}_q} \ell_n(f) - \mathrm{pen}(n,q) \right\}$$

where $\mathrm{pen}(n,q)$ is a penalty function and $\ell_n(f)$ is the likelihood of $(X_k)_{1\leq k\leq n}$ under the distribution f. Our aim is to understand what penalties yield strong consistency of the order estimator, i.e., $\hat{q}_n \to q^\star$ as $n\to\infty$ a.s. Characterizing strong consistency hinges on a precise understanding of the pathwise fluctuations of the likelihood ratio statistic

$$\sup_{f\in\mathcal{M}_q} \ell_n(f) - \sup_{f\in\mathcal{M}_{q^\star}} \ell_n(f)$$

as $n\to\infty$, *uniformly* in the model order $q > q^\star$. When there is a known upper bound on the order $q^\star \leq q_{\max} < \infty$ and the model classes $\mathcal{M}_q$ are regularly parametrized by a compact subset of Euclidean space, an upper bound on the pathwise fluctuations can be obtained by classical parametric methods: Taylor expansion of the likelihood and an application of a law of iterated logarithm. This approach forms the basis for most consistency proofs for penalized likelihood order estimators in the literature (see, for example, [2]–[6]). However, such techniques fail in the absence of a prior upper bound: even though each model class $\mathcal{M}_q$ is finite dimensional, the full model $\mathcal{M} = \bigcup_q \mathcal{M}_q$ is infinite dimensional and, as such, the problem in the absence of a prior upper bound is inherently nonparametric. When the classes $\mathcal{M}_q$ are noncompact, one must introduce sieves $\mathcal{M}_q^n \subset \mathcal{M}_q^{n+1} \subset \cdots \subset \mathcal{M}_q$, complicating the problem further (in this case, even the parametric theory remains poorly understood [7]–[9]). An entirely different approach based on universal coding theory [4], [10]–[14] yields bounds on the pathwise fluctuations that do not require prior bounds on the order or compactness of the models. However, these bounds are far from tight and cannot even establish consistency of BIC, let alone smaller penalties (this appears to be a fundamental limitation of this approach due to Rissanen's theorem; see [15] and [16]).

The problem area that is investigated in this paper was initiated in the work of Csiszár and Shields [1], [17], who proved consistency of BIC for Markov chain order estimation in absence of a prior bound on the order (see also [18]). To our

Manuscript received February 21, 2012; revised August 06, 2012; accepted September 23, 2012. Date of publication October 02, 2012; date of current version January 16, 2013.

E. Gassiat is with the Laboratoire de Mathématiques d'Orsay, Université Paris-Sud, 91405 Orsay Cedex, France (e-mail: elisabeth.gassiat@math.u-psud.fr).

R. van Handel is with the Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ 08544 USA (e-mail: rvan@princeton.edu).

Communicated by G. Moustakides, Associate Editor for Detection and Estimation.

Digital Object Identifier 10.1109/TIT.2012.2221122

knowledge, little progress has been made on this subject beyond their work. The proofs in [1] and [17] rely heavily on the availability of an explicit expression for the maximum likelihood for Markov chains and employ delicate estimates specific to that setting. Their techniques are, therefore, not well suited to investigating such problems in other settings. Moreover, the methods in [1] and [17] do not yield minimal penalties. However, the Markov chain case was recently reconsidered in [19] using very different techniques based on empirical process theory, which are potentially much more generally applicable and which shed light on minimal penalties.

The main results of this paper provide generally applicable upper and lower bounds on the pathwise fluctuations of the likelihood ratio statistic uniformly in the model order $q > q^\star$, for the case of i.i.d. observations $(X_k)_{k\geq 1}$, without a prior bound on the model order and in possibly noncompact parameter spaces. These results are then used to investigate strong consistency of penalized likelihood order estimators. We use empirical process methods as in [19], but the difficulties to be surmounted in the present setting are of a different nature. The main difficulty for Markov chain models in [1], [17], and [19] is their dependence structure; in this paper, we assume i.i.d. models. On the other hand, the geometric structure of Markov chains is exceedingly simple: the family of qth-order Markov chains in the Hellinger distance is simply a Euclidean ball when viewed in the appropriate parametrization. In contrast, in general order estimation problems, one is often faced with model classes that are geometrically very complex. An important case study that will be considered in this paper is location mixture models (widely used in practice for clustering), which possess a notoriously complicated nonregular geometry. We will be able, for example, to establish strong consistency of BIC for mixture order estimation in absence of a prior bound on the order or on the parameter set, providing a counterpart to the results of Csiszár and Shields [1] in a setting very different than that of Markov chains.

The techniques developed here originate in our attempts to understand the order estimation problem for hidden Markov models (HMM) [12]. In that setting, consistency of BIC (even with a prior bound on the order) remains unknown. The two cases considered here and in [19]—Markov chains and i.i.d. mixtures—can be viewed as two extreme cases of HMM. While our approach provides a substantial step toward understanding the HMM setting, a striking and as of yet poorly understood breakdown in the ergodicity of HMM [20] has so far prohibited further progress in this direction.

The remainder of this paper is organized as follows. Section II introduces the general model under consideration and states our results on the pathwise fluctuations of the likelihood ratio statistic. Section III derives the consequences for order estimation and considers also the special case of location mixture models. Proofs are given in the appendixes.

II. Pathwise Fluctuations of the Likelihood

A. Basic Setting and Notation

Let $(E, \mathcal{E}, \mu)$ be a measure space. For each $q, n \geq 1$, let $\mathcal{M}_q^n$ be a given family of strictly positive probability densities with respect to μ (that is, we assume that $\int f d\mu = 1$ and that $f > 0$ μ-a.e. for every $f \in \mathcal{M}_q^n$). Moreover, we assume that $(\mathcal{M}_q^n)_{q,n\geq 1}$ is a nested family of models in the sense that $\mathcal{M}_q^n \subseteq \mathcal{M}_{q+1}^n$ and $\mathcal{M}_q^n \subseteq \mathcal{M}_q^{n+1}$ for all $q, n \geq 1$. Let $\mathcal{M}_q = \bigcup_n \mathcal{M}_q^n$, $\mathcal{M}^n = \bigcup_q \mathcal{M}_q^n$, and $\mathcal{M} = \bigcup_{q,n} \mathcal{M}_q^n$.

Consider an i.i.d. sequence of E-valued random variables $(X_k)_{k\geq 1}$ whose common distribution under the measure $\mathbf{P}^\star$ is $f^\star d\mu$, where $f^\star \in \mathcal{M}_{q^\star} \setminus \mathrm{cl}\, \mathcal{M}_{q^\star - 1}$ for some $q^\star \geq 1$ (here $\mathrm{cl}\, \mathcal{M}_q$ denotes the $L^1(d\mu)$-closure of $\mathcal{M}_q$). The index $q^\star$ is called the *model order*. Let us define

$$\ell_n(f) = \sum_{i=1}^n \log f(X_i), \qquad f \in \mathcal{M}.$$

Evidently, $\ell_n(f)$ is the log-likelihood of the i.i.d. sequence $(X_k)_{k\leq n}$ when $X_k \sim f d\mu$. Our aim is to study the pathwise fluctuations of the likelihood ratio statistic

$$\sup_{f\in\mathcal{M}_q^n} \ell_n(f) - \sup_{f\in\mathcal{M}_{q^\star}^n} \ell_n(f)$$

as $n \to \infty$, *uniformly* over the order parameter $q \geq q^\star$. Pathwise upper and lower bounds on the likelihood ratio statistic are the key ingredient in the study of strong consistency of penalized likelihood order estimators (see Section III).

Example II.1 (Location Mixtures): The guiding example for our theory, the case of location mixtures, will be studied in detail in Section III-B. We presently introduce this example in order to clarify our basic setup.

Let $E = \mathbb{R}^d$ (with its Borel σ-field $\mathcal{E}$) and let μ be the Lebesgue measure on $\mathbb{R}^d$. We fix a strictly positive probability density f_0 with respect to μ, and define $f_\theta(x) = f_0(x - \theta)$ for $x, \theta \in \mathbb{R}^d$. Fix a sequence $T(n) \uparrow \infty$ and define

$$\mathcal{M}_q^n = \left\{ \sum_{i=1}^q \pi_i f_{\theta_i} : \pi_i \geq 0,\ \sum_{i=1}^q \pi_i = 1,\ \|\theta_i\| \leq T(n) \right\}.$$

Then, $\mathcal{M}_q$ is the family of all q-component mixtures of translates of the density f_0, while $\mathcal{M}_q^n$ is the subset of the mixtures $\mathcal{M}_q$ whose translation parameters $(\theta_i)_{i=1,\ldots,q}$ are restricted to a ball of radius $T(n)$. The number of components $q^\star$ of the true mixture $f^\star \in \mathcal{M}$ can be estimated from observations using the order estimator

$$\hat{q}_n = \operatorname*{argmax}_{q\geq 1} \left\{ \sup_{f\in\mathcal{M}_q^n} \ell_n(f) - \mathrm{pen}(n, q) \right\}.$$

Pathwise control of the likelihood ratio statistic allows us to identify what penalties $\mathrm{pen}(n, q)$ and cutoff sequences $T(n)$ yield strong consistency of $\hat{q}_n$ (cf., Section III-B).

Remark II.2: To avoid measurability problems and other technical complications, we employ throughout this paper the simplifying convention that all uncountable suprema (such as $\sup_{f\in\mathcal{M}_q^n} \ell_n(f)$) are interpreted as essential suprema with respect to the measure $\mathbf{P}^\star$. In the majority of applications, the model classes $\mathcal{M}_q^n$ will be separable, in which case the supremum and essential supremum coincide.

In the sequel, we will denote by $\|\cdot\|_p$ the $L^p(f^\star d\mu)$-norm, i.e., $\|g\|_p^p = \int |g(x)|^p f^\star(x)\mu(dx)$, and we denote by

$\langle f, g\rangle = \int f(x)g(x)f^\star(x)\mu(dx)$ the Hilbert space inner product in $L^2(f^\star d\mu)$. Define the Hellinger distance

$$h(f,g)^2 = \int (\sqrt{f} - \sqrt{g})^2 d\mu, \qquad f, g \in \mathcal{M}.$$

It is easily seen that $h(f, f^\star) = \|\sqrt{f/f^\star} - 1\|_2$. Finally, we will denote by $\mathcal{N}(\mathcal{Q}, \delta)$ for any class of functions $\mathcal{Q}$ and $\delta > 0$ the minimal number of brackets of $L^2(f^\star d\mu)$-width δ needed to cover $\mathcal{Q}$: that is, $\mathcal{N}(\mathcal{Q}, \delta)$ is the smallest cardinality N of a collection of pairs of functions $\{g_i^L, g_i^U\}_{i=1,\ldots,N}$ such that $\max_{i\le N} \|g_i^U - g_i^L\|_2 \le \delta$ and for every $g \in \mathcal{Q}$, we have $g_i^L \le g \le g_i^U$ pointwise for some $i \le N$.

B. Upper Bound

We aim to obtain a pathwise upper bound on the likelihood ratio statistic that holds *uniformly* in $q > q^\star$. To this end, define for $q, n \ge 1$ and $\varepsilon > 0$ the Hellinger ball

$$\mathcal{H}_q^n(\varepsilon) = \{\sqrt{f/f^\star} : f \in \mathcal{M}_q^n,\ h(f, f^\star) \le \varepsilon\}.$$

Note that the definition of $\mathcal{H}_q^n(\varepsilon)$ depends on $f^\star$ (which is fixed throughout this paper). The following result shows that the geometry of the Hellinger balls $\mathcal{H}_q^n(\varepsilon)$ controls the pathwise fluctuations of the likelihood ratio statistic.

Theorem II.3: Suppose that for all n sufficiently large

$$\mathcal{N}(\mathcal{H}_q^n(\varepsilon), \delta) \le \left(\frac{K(n)\varepsilon}{\delta}\right)^{\eta(q)}$$

for all $q \ge q^\star$ and $\delta \le \varepsilon$, where $K(n) \ge 1$ and $\eta(q) \ge q$ are increasing functions. Then

$$\overline{\lim_{n\to\infty}} \frac{1}{\log K(2n) \vee \log\log n} \times \sup_{q\ge q^\star} \frac{1}{\eta(q)} \left\{ \sup_{f\in\mathcal{M}_q^n} \ell_n(f) - \sup_{f\in\mathcal{M}_{q^\star}^n} \ell_n(f) \right\} \le C$$

$\mathbf{P}^\star$-a.s., where $C > 0$ is a universal constant.

The proof of Theorem II.3 is given in Appendix A.

The assumption of Theorem II.3 on the entropy of the Hellinger balls $\mathcal{H}_q^n(\varepsilon)$ states, roughly speaking, that the class of densities $\mathcal{M}_q^n$ endowed with the Hellinger distance has the same metric structure as a Euclidean ball of dimension $\eta(q)$ and radius of order $K(n)$, at least locally in a neighborhood of the true density $f^\star$. The effective dimension $\eta(q)$ controls the fluctuations of the likelihood ratio statistic as a function of the model order, while the effective radius $K(n)$ controls the fluctuations as a function of time up to a minimal rate of order $\log\log n$. In the following section, we will see that the minimal $\log\log n$ rate is indeed optimal.

Remark II.4: A bound on $\mathcal{N}(\mathcal{H}_q^n(\varepsilon), \delta)$ of the form required by Theorem II.3 is easily obtained if $\mathcal{M}_q^n$ are regularly parametrized classes. That is, suppose that we can write

$$\mathcal{M}_q^n = \{f_\theta : \theta \in \Theta_q^n\}, \qquad \Theta_q^n \subset \mathbb{R}^{\eta(q)}$$

where we have a pointwise Lipschitz estimate of the form

$$|\sqrt{f_\theta(x)/f^\star(x)} - \sqrt{f_{\theta'}(x)/f^\star(x)}| \le F(x)\,|||\theta - \theta'|||$$

for some function F in L^2 and norm $|||\cdot|||$ on $\mathbb{R}^{\eta(q)}$, and

$$h(f_\theta, f^\star) \ge c\,|||\theta - \theta^\star|||$$

with $c > 0$. Then, the requisite bound on $\mathcal{N}(\mathcal{H}_q^n(\varepsilon), \delta)$ follows easily (cf., [21, Example 19.7]). This covers many cases of practical interest. However, geometrically complex models such as finite mixtures do not admit a regular parametrization, while our results are nonetheless sufficiently general to apply to such models (see Section III-B). In nonregular models, the entropy bound required by Theorem II.3 is far from obvious, and the requisite geometric analysis is of independent interest. Such problems are investigated by the authors in [22] and form the basis for the results in Section III-B.

C. Lower Bound

Throughout this section, we specialize to the case that $\mathcal{M}_q^n = \mathcal{M}_q$ does not depend on n (this implies essentially that $\mathcal{M}_q$ is compact). In this setting, Theorem II.3 yields an upper bound of order $\log\log n$ on the pathwise fluctuations of the likelihood ratio statistic. The aim of this section is to obtain a matching lower bound of order $\log\log n$, which shows that the minimal rate in Theorem II.3 is essentially optimal. For the purposes of a lower bound, uniformity in q is irrelevant, so it suffices to restrict attention to some fixed $q > q^\star$. We will in fact obtain a much stronger result in this case that completely characterizes the pathwise asymptotics of the likelihood ratio statistic for fixed q in sufficiently smooth families.

The geometric structure required in this section is somewhat different than that of Theorem II.3. Instead of Hellinger balls, we consider the classes of weighted densities $\mathcal{D}_q = \{d_f : f \in \mathcal{M}_q,\ f \ne f^\star\}$ and $\mathcal{D} = \bigcup_q \mathcal{D}_q$, where

$$d_f = \frac{\sqrt{f/f^\star} - 1}{h(f, f^\star)}, \quad f \in \mathcal{M}, \quad f \ne f^\star.$$

Define for $\varepsilon > 0$ and $q \ge 1$ the local weighted classes

$$\mathcal{D}_q(\varepsilon) = \{d_f : f \in \mathcal{M}_q,\ 0 < h(f, f^\star) \le \varepsilon\}$$
$$\bar{\mathcal{D}}_q = \bigcap_{\varepsilon>0} \mathrm{cl}\,\mathcal{D}_q(\varepsilon)$$

where the closure $\mathrm{cl}\,\mathcal{D}_q(\varepsilon)$ is in $L^2(f^\star d\mu)$. Clearly, $\bar{\mathcal{D}}_q$ is the set of all possible limit points of d_f as $h(f, f^\star) \to 0$ in $\mathcal{M}_q$. If the neighborhoods of $\bar{\mathcal{D}}_q$ are sufficiently rich, such limits can be taken along a continuous path in the following sense.

Definition II.5: A point $d \in \bar{\mathcal{D}}_q$ is called continuously accessible if there is a path $(f_t)_{t\in]0,1]} \subset \mathcal{M}_q \backslash \{f^\star\}$ such that the map $t \mapsto h(f_t, f^\star)$ is continuous, $h(f_t, f^\star) \to 0$ as $t \to 0$, and $d_{f_t} \to d$ in $L^2(f^\star d\mu)$ as $t \to 0$. The subset of all continuously accessible points in $\bar{\mathcal{D}}_q$ is denoted as $\bar{\mathcal{D}}_q^c$.

We can now formulate the main result of this section.

Theorem II.6: Let $q^\star \leq p < q$. Assume that

$$\int_0^1 \sqrt{\log \mathcal{N}(\mathcal{D}_q, u)}\, du < \infty$$

and that $|d| \leq D$ for all $d \in \mathcal{D}_q$ with $D \in L^{2+\alpha}(f^\star d\mu)$ for some $\alpha > 0$. Then

$$\overline{\lim_{n\to\infty}} \frac{1}{\log\log n} \left\{ \sup_{f\in\mathcal{M}_q} \ell_n(f) - \sup_{f\in\mathcal{M}_p} \ell_n(f) \right\} \geq \sup_{g\in L_0^2(f^\star d\mu)} \left\{ \sup_{f\in\bar{\mathcal{D}}_q^c} (\langle f,g\rangle)_+^2 - \sup_{f\in\bar{\mathcal{D}}_p} (\langle f,g\rangle)_+^2 \right\} \quad \mathbf{P}^\star\text{-a.s.}$$

as well as

$$\overline{\lim_{n\to\infty}} \frac{1}{\log\log n} \left\{ \sup_{f\in\mathcal{M}_q} \ell_n(f) - \sup_{f\in\mathcal{M}_p} \ell_n(f) \right\} \leq \sup_{g\in L_0^2(f^\star d\mu)} \left\{ \sup_{f\in\bar{\mathcal{D}}_q} (\langle f,g\rangle)_+^2 - \sup_{f\in\bar{\mathcal{D}}_p^c} (\langle f,g\rangle)_+^2 \right\} \quad \mathbf{P}^\star\text{-a.s.}$$

where $L_0^2(f^\star d\mu) = \{g \in L^2(f^\star d\mu) : \|g\|_2 \leq 1,\ \langle 1, g\rangle = 0\}$.

Only the first (lower bound) part of the theorem is needed to conclude optimality of the minimal $\log\log n$ rate in Theorem II.3. Indeed, we will obtain as a corollary the following lower bound counterpart to Theorem II.3.

Corollary II.7: Suppose there exists $q > q^\star$ such that the following hold.

1) There is an envelope function $D : E \to \mathbb{R}$ such that $|d| \leq D$ for all $d \in \mathcal{D}_q$ and $D \in L^{2+\alpha}(f^\star d\mu)$ for some $\alpha > 0$. Moreover, $\int_0^1 \sqrt{\log \mathcal{N}(\mathcal{D}_q, u)}\, du < \infty$.
2) $\bar{\mathcal{D}}_q^c \backslash \bar{\mathcal{D}}_{q^\star}$ is nonempty.

Let $\eta(q) > 0$ be an arbitrary positive function. Then

$$\overline{\lim_{n\to\infty}} \frac{1}{\log\log n} \sup_{q\geq q^\star} \frac{1}{\eta(q)} \left\{ \sup_{f\in\mathcal{M}_q} \ell_n(f) - \sup_{f\in\mathcal{M}_{q^\star}} \ell_n(f) \right\} \geq C$$

$\mathbf{P}^\star$-a.s., where $C > 0$ is nonrandom but may depend on $f^\star, \eta$.

The proofs of Theorem II.6 and Corollary II.7 are given in Appendix B.

The fact that the geometric assumptions in Theorem II.6 and Corollary II.7 are expressed in terms of weighted classes is not surprising, as the sharp asymptotic expression provided by Theorem II.6 for the pathwise fluctuations of the likelihood are expressed in terms of a variational problem on the weighted classes. Nonetheless, we are naturally led to ask whether there is any relation between the geometric assumptions imposed in the upper bound Theorem II.3 and the lower bound Theorem II.6, which appear to be quite different at first sight. In [22], we show that the global entropy of the weighted class is closely related to local entropy, so that the geometric assumptions for the upper and lower bounds are not too far apart.

Remark II.8: When $\bar{\mathcal{D}}_q$ and $\bar{\mathcal{D}}_p$ each contain an $L^2(f^\star d\mu)$-dense subset of continuously accessible points (which is typically the case in sufficiently smooth models), Theorem II.6 provides the exact characterization

$$\overline{\lim_{n\to\infty}} \frac{1}{\log\log n} \left\{ \sup_{f\in\mathcal{M}_q} \ell_n(f) - \sup_{f\in\mathcal{M}_p} \ell_n(f) \right\} = \sup_{g\in L_0^2(f^\star d\mu)} \left\{ \sup_{f\in\bar{\mathcal{D}}_q} (\langle f,g\rangle)_+^2 - \sup_{f\in\bar{\mathcal{D}}_p} (\langle f,g\rangle)_+^2 \right\} \quad \mathbf{P}^\star\text{-a.s..}$$

Besides its intrinsic interest, this result has a surprising consequence. In the case that $\mathcal{M}_q$ and $\mathcal{M}_p$ are regular parametric models with $\dim(\mathcal{M}_q) > \dim(\mathcal{M}_p)$, one can choose $g \in \bar{\mathcal{D}}_q$ which is orthogonal to $\bar{\mathcal{D}}_p$. As $\bar{\mathcal{D}}_q, \bar{\mathcal{D}}_p \subseteq L_0^2(f^\star d\mu)$ (see the proof of Corollary II.7), it follows easily that in this case, the right-hand side of the previous equation display is precisely equal to 1. In particular, we obtain the curious conclusion that in regular parametric models, the magnitude of the fluctuations of the likelihood ratio statistic does not depend on the dimensions $\dim(\mathcal{M}_q)$ and $\dim(\mathcal{M}_p)$. In contrast, it is well known that in regular parametric models, the likelihood ratio statistic itself converges weakly to a chi-square distribution with $\dim(\mathcal{M}_q) - \dim(\mathcal{M}_p)$ degrees of freedom, so the tails of the distribution of the likelihood ratio statistic do in fact depend strongly on the dimensions $\dim(\mathcal{M}_q)$ and $\dim(\mathcal{M}_p)$. Of course, the dimension independence of the pathwise fluctuations will also cease to hold if we are interested in a result that is uniform in the order q, as in Theorem II.3. This highlights the fact that the problems investigated in this paper are fundamentally different depending on whether or not one assumes a prior upper bound on the model order.

III. Strongly Consistent Order Estimation

The goal of this section is to apply the results of Section II to identify what penalties and cutoffs yield strongly consistent order estimators. We first develop some general consistency and inconsistency results and then consider specifically the challenging problem of mixture order estimation.

A. Consistency and Minimal Penalties

In this section, we consider the general setting introduced in Section II-A. We now suppose, however, that the true model order $q^\star$ (as well as the true density $f^\star$) is not known, so that we must estimate $q^\star$ from an observation sequence $(X_k)_{k\geq 1}$. To this end, define the penalized likelihood order estimator

$$\hat{q}_n = \operatorname*{argmax}_{q\geq 1} \left\{ \sup_{f\in\mathcal{M}_q^n} \ell_n(f) - \mathrm{pen}(n,q) \right\}$$

where $\mathrm{pen}(n,q)$ is a penalty function. Our goal is to show that the penalized likelihood order estimator is strongly consistent, i.e., $\hat{q}_n \to q^\star$ as $n \to \infty$ $\mathbf{P}^\star$-a.s., for a suitable choice of the penalty (that does not depend on $q^\star$ or $f^\star$). Let us emphasize that the maximum in the definition of $\hat{q}_n$ is taken over *all* model orders $q \geq 1$, that is, we do not assume that an *a priori* upper bound on the order is available, in contrast to most previous work on this topic.

We obtain the following general result.

Theorem III.1: Suppose that for all n sufficiently large

$$\mathcal{N}(\mathcal{H}_q^n(\varepsilon),\delta) \le \left(\frac{K(n)\varepsilon}{\delta}\right)^{\eta(q)}$$

for all $q \ge q^\star$ and $\delta \le \varepsilon$, where $K(n) \ge 1$ and $\eta(q) \ge q$ are increasing functions and we assume that $\log K(n) = o(n)$. Let $\mathrm{pen}(n,q)$ be a penalty that is increasing in q and

$$\lim_{n\to\infty} \sup_{q>q^\star} \frac{\eta(q)\{\log K(2n) \vee \log\log n\}}{\mathrm{pen}(n,q) - \mathrm{pen}(n,q^\star)} = 0$$
$$\lim_{n\to\infty} \max_{q<q^\star} \frac{\mathrm{pen}(n,q)}{n} = 0.$$

Then, $\hat{q}_n \to q^\star$ as $n \to \infty$ $\mathbf{P}^\star$-a.s.

Theorem III.1 is proved in Appendix C.

Let us now specialize to the case that $\mathcal{M}_q^n = \mathcal{M}_q$ does not depend on n, as in Section II-C. In this case, Theorem III.1 immediately yields the following corollary.

Corollary III.2: Suppose that for all $q \ge q^\star$ and $\delta \le \varepsilon$

$$\mathcal{N}(\mathcal{H}_q(\varepsilon),\delta) \le \left(\frac{K\varepsilon}{\delta}\right)^{\eta(q)}$$

where $K \ge 1$ and $\eta(q) \ge q$ is a strictly increasing function. Define the penalty

$$\mathrm{pen}(n,q) = \eta(q)\,\varpi(n)$$

where $\varpi(n)$ is any function such that

$$\lim_{n\to\infty} \frac{\log\log n}{\varpi(n)} = 0, \qquad \lim_{n\to\infty} \frac{\varpi(n)}{n} = 0.$$

Then, $\hat{q}_n \to q^\star$ as $n \to \infty$ $\mathbf{P}^\star$-a.s.

Corollary III.2 states that, when $\mathcal{M}_q^n = \mathcal{M}_q$ does not depend on n, the penalized likelihood order estimator is strongly consistent provided the penalty grows faster than $\log\log n$ and slower than n. Clearly, the $\log\log n$ rate is the minimal one attainable by applying Theorem III.1. This raises the question whether the $\log\log n$ rate is indeed minimal, in the sense that smaller penalties yield inconsistent estimators. The following result shows that this is indeed the case, so that the result of Corollary III.2 is essentially optimal.

Corollary III.3: Suppose there exists $q > q^\star$ such that

1) there is an envelope function $D : E \to \mathbb{R}$ such that $|d| \le D$ for all $d \in \mathcal{D}_q$, $D \in L^{2+\alpha}(f^\star d\mu)$ for some $\alpha > 0$, and $\int_0^1 \sqrt{\log \mathcal{N}(\mathcal{D}_q, u)}\, du < \infty$;
2) $\bar{\mathcal{D}}_q^c \backslash \bar{\mathcal{D}}_{q^\star}$ is nonempty.

Let $\eta(q) > 0$ be any strictly increasing function, and let

$$\mathrm{pen}(n,q) = C\,\eta(q)\,\log\log n.$$

If $C > 0$ is sufficiently small, $\hat{q}_n \ne q^\star$ infinitely often $\mathbf{P}^\star$-a.s.

The proof of Corollary III.3 is given in Appendix C. Let us note that the proof of Corollary III.3 actually shows that $\sup_{f\in\mathcal{M}_q} \ell_n(f) - \mathrm{pen}(n,q) > \sup_{f\in\mathcal{M}_{q^\star}} \ell_n(f) - \mathrm{pen}(n,q^\star)$ infinitely often $\mathbf{P}^\star$-a.s., so Corollary III.3 is not altered even if we were to impose a prior upper bound on the order.

In conclusion, we have shown that when $\mathcal{M}_q^n = \mathcal{M}_q$ does not depend on n, penalties growing faster than $\log\log n$ are consistent, while the penalty $C\,\eta(q)\log\log n$ is inconsistent when the constant C is sufficiently small. From the proof of Theorem III.1, we can also see that the penalty $C\,\eta(q)\log\log n$ is consistent when C is sufficiently large. However, the critical value of C may depend on the unknown parameter $f^\star$, so that this *minimal* penalty may not be implementable. On the other hand, assuming that $\eta(q)$ does not depend on $f^\star$ (as is typically the case), penalties satisfying the assumptions of Theorem III.1 obviously do not depend on the unknown parameter $f^\star$ and therefore define admissible estimators. When $\mathcal{M}_q^n$ depends on n, larger penalties may be required to ensure consistency, depending on the growth rate of $K(n)$.

B. Location Mixture Order Estimation

We finally apply our general results to location mixture order estimation. Throughout this section, let $E = \mathbb{R}^d$ and let μ be the Lebesgue measure on $\mathbb{R}^d$. Fix a strictly positive probability density f_0 with respect to μ, and define

$$\mathcal{M}_q^n = \left\{ \sum_{i=1}^q \pi_i f_{\theta_i} : \pi_i \ge 0,\ \sum_{i=1}^q \pi_i = 1,\ \theta_i \in \Theta(n) \right\}$$

where $f_\theta(x) = f_0(x-\theta)$ and $\cdots \subseteq \Theta(n) \subseteq \Theta(n+1) \subseteq \cdots \subset \mathbb{R}^d$ is an increasing family of bounded subsets of $\mathbb{R}^d$. We fix $f^\star \in \mathcal{M}$ throughout this section. Let

$$H_0(x) = \sup_{\theta\in\Theta} f_\theta(x)/f^\star(x)$$
$$H_1(x) = \sup_{\theta\in\Theta} \max_{i=1,\ldots,d} |\partial f_\theta(x)/\partial\theta^i|/f^\star(x)$$
$$H_2(x) = \sup_{\theta\in\Theta} \max_{i,j=1,\ldots,d} |\partial^2 f_\theta(x)/\partial\theta^i\partial\theta^j|/f^\star(x)$$
$$H_3(x) = \sup_{\theta\in\Theta} \max_{i,j,k=1,\ldots,d} |\partial^3 f_\theta(x)/\partial\theta^i\partial\theta^j\partial\theta^k|/f^\star(x)$$

when f_0 is sufficiently differentiable, and let

Assumption A: The following hold:

1) $f_0 \in C^3$ and $f_0(x)$, $(\partial f_0/\partial\theta^i)(x)$ vanish as $\|x\| \to \infty$.
2) $H_k \in L^4(f^\star d\mu)$ for $k = 0,1,2$ and $H_3 \in L^2(f^\star d\mu)$.

In the following, we consider two separate cases. The first case is that of a compact parameter set, where $\Theta(n) = \Theta$ does not depend on n. In this setting, we obtain a general result. Then, we consider the noncompact case in the setting of Gaussian mixtures and illustrate how Theorem III.1 can be used to obtain consistency results in this case. To be able to use Theorem III.1, we need suitable estimates on the local entropy of mixtures. The following result is given in [22].

Theorem III.4: Suppose that Assumption A holds. Then, if $\Theta(n) = \Theta$ is a bounded subset of $\mathbb{R}^d$ with diameter $2T$

$$\mathcal{N}(\mathcal{H}_q^n(\varepsilon),\delta) \le \left(\frac{C_\Theta\,\varepsilon}{\delta}\right)^{10(d+1)q+1}$$

for all $q \ge q^\star$ and $\delta/\varepsilon \le 1$, where

$$C_\Theta = L^\star\,(T \vee 1)^{1/3}\,(\|H_0\|_4^4 \vee \|H_1\|_4^4 \vee \|H_2\|_4^4 \vee \|H_3\|_2^2)^{5/4}$$

and $L^\star$ is a constant that depends only on d, $q^\star$, and $f^\star$.

Example III.5 (Gaussian Mixtures): Consider mixtures of Gaussian densities $f_0(x) = (2\pi)^{-d/2}|\Sigma|^{-1/2}e^{-x^T\Sigma^{-1}x/2}$, where Σ is a positive-definite $d \times d$ matrix, $|\Sigma|$ denotes its determinant, and x^T the transpose vector of x, and let $\Theta(T) = \{\theta \in \mathbb{R}^d : \|\theta\| \leq T\}$. Fix a nondegenerate mixture $f^\star$, and define $T^\star = \max_{i=1,\ldots,q^\star} \|\theta_i^\star\|$. Denote by $\mathcal{H}_q(\varepsilon, T)$ the Hellinger ball associated with the parameter set $\Theta(T)$. Then

$$\mathcal{N}(\mathcal{H}_q(\varepsilon, T), \delta) \leq \left(\frac{C_1^\star e^{C_2^\star T^2}\varepsilon}{\delta}\right)^{10(d+1)q+1}$$

for all $q \geq q^\star$, $T \geq T^\star$, and $\delta/\varepsilon \leq 1$, where $C_1^\star \text{and} C_2^\star$ are constants that depend on d, $q^\star$, and $f^\star$ only. To prove this, it suffices to show that Assumption A holds and that $\|H_k\|_4$ for $k = 0, 1, 2$ and $\|H_3\|_2$ are of order e^{CT^2}. These facts are readily verified by a straightforward computation.

Let us first consider the case of a compact parameter set. We obtain a general consistency result under Assumption A.

Proposition III.6: Suppose that the parameter set $\Theta(n) = \Theta$ is a bounded subset of $\mathbb{R}^d$ independent of n, and that Assumption A holds. If we choose a penalty of the form

$$\text{pen}(n, q) = q\,\omega(n), \qquad \lim_{n\to\infty} \frac{\log\log n}{\omega(n)} = \lim_{n\to\infty} \frac{\omega(n)}{n} = 0,$$

then $\hat{q}_n \to q^\star$ as $n \to \infty$ $\mathbf{P}^\star$-a.s. On the other hand, if

$$\text{pen}(n, q) = C\,q\,\log\log n$$

where $C > 0$ is a sufficiently small constant, then we have $\hat{q}_n \neq q^\star$ infinitely often $\mathbf{P}^\star$-a.s.

We, therefore, find that in the setting of location mixtures with a compact parameter set, the minimal penalty is of order $\log\log n$. Moreover, the popular BIC penalty

$$\text{pen}(n, q) = \frac{dq + q - 1}{2}\,\log n \tag{III.1}$$

yields a strongly consistent mixture order estimator in this setting, without a prior upper bound on the order. The requisite Assumption A is mild, which highlights the broad applicability of this result. However, the assumption of a compact parameter space can be quite restrictive in practice.

Let us, therefore, consider a case where the parameter space is noncompact. For simplicity, we restrict our attention to Gaussian mixtures, that is, we choose $f_0(x) = (2\pi)^{-d/2}e^{-\|x\|^2/2}$, and we choose the restricted parameter sets $\Theta(n) = \{\theta \in \mathbb{R}^d : \|\theta\| \leq T(n)\}$ for some sequence $T(n) \uparrow \infty$. Our aim is to choose the penalty $\text{pen}(n, q)$ and cutoff $T(n)$ so that the penalized likelihood order estimator is strongly consistent.

In this setting, we obtain the following result.

Proposition III.7: For the case $f_0(x) = (2\pi)^{-d/2}|\Sigma|^{-1/2}e^{-x^T\Sigma^{-1}x/2}$ and $\Theta(n) = \{\theta \in \mathbb{R}^d : \|\theta\| \leq T(n)\}$, consider a penalty of the form $\text{pen}(n, q) = q\,\omega(n)$. If

$$\lim_{n\to\infty} \frac{\log\log n}{\omega(n)} = \lim_{n\to\infty} \frac{\omega(n)}{n} = 0, \quad T(n) = O(\sqrt{\log\log n})$$

then $\hat{q}_n \to q^\star$ as $n \to \infty$ $\mathbf{P}^\star$-a.s.

On the other hand, the BIC penalty (III.1) yields a strongly consistent order estimator if $T(n) = o(\sqrt{\log n})$.

This result illustrates that our theory can establish consistency of the penalized likelihood mixture order estimator without any prior upper bounds on the model order or the magnitude of the true parameters. Let us note that there is nothing particularly special about the Gaussian case: a similar result can be obtained, in principle, for any mixture distribution f_0, as long as one can obtain suitable estimates on the quantities $\|H_i\|_4$ that appear in Theorem III.4 (see Example III.5 for the Gaussian case).

The proofs of Propositions III.6 and III.7 are given in Appendix D.

C. Remarks

The above results provide a fairly sharp theoretical understanding of the consistency properties of penalized likelihood order estimators. We have seen that penalties of order $\log\log n$ are minimal and that penalties of order n are maximal (the latter is easily seen in the proof of Theorem III.1) without a prior upper bound on the order. This rather broad range leaves significant freedom in the choice of a penalty in practice.

There can be no optimal penalty for all situations. The choice of penalty allows us to tradeoff between underestimation and overestimation of the order: a larger penalty increases the probability of underestimation but decreases the probability of overestimation. Which effect is more undesirable will depend on the application at hand. Moreover, we have only addressed the consistency properties of order estimators: other considerations, such as finite-sample performance, may be of significant importance. Finite-sample bounds that appear in our proofs (as in the proof of Theorem II.3) suggest that the rate of convergence of $\log\log n$ penalties can be very slow.

Many practitioners use the BIC penalty, whose performance is supported by numerous empirical studies in the literature. The aim of this paper is not to suggest an alternative methodology, but to explore the fundamental limitations to a broad class of penalized likelihood order estimators in the absence of a prior upper bound on the order. When the parameter space is not compact, our results can also provide some guidance to the practical implementation of order estimators such as BIC.

Appendix A
Proof of Theorem II.3

The proof of Theorem II.3 is based on the following deviation bound for the log-likelihood ratio. This bound is essentially from [23, Corollary 7.5], but the additional maximum inside the probability is essential for our purposes.

Theorem A.1: Let $\mathcal{M}$ be a family of strictly positive probability densities with respect to a reference measure μ, fix some $f^\star \in \mathcal{M}$, and define the Hellinger ball $\mathcal{H}(\varepsilon) = \{\sqrt{f/f^\star} : f \in \mathcal{M},\ h(f, f^\star) \leq \varepsilon\}$ where $h(f, g)^2 = \int(\sqrt{f} - \sqrt{g})^2 d\mu$. Suppose that for some constants $K \geq 1$, $p \geq 1$ and all $\delta \leq \varepsilon$

$$\mathcal{N}(\mathcal{H}(\varepsilon), \delta) \leq \left(\frac{K\varepsilon}{\delta}\right)^p$$

where $\mathcal{N}(\mathcal{H}(\varepsilon), \delta)$ is the minimal number of brackets of $L^2(f^\star d\mu)$-width δ needed to cover $\mathcal{H}(\varepsilon)$. Let $(X_i)_{i\in\mathbb{N}}$ be i.i.d. with distribution $f^\star d\mu$. Then

$$\mathbf{P}\left[\max_{n\leq k\leq 2n} \sup_{f\in\mathcal{M}} \sum_{j=1}^{k} \log\left(\frac{f(X_j)}{f^\star(X_j)}\right) \geq \alpha\right] \leq C\,e^{-\alpha/C}$$

for all $\alpha \geq Cp(1 + \log K)$, $n \geq 1$ [C is a universal constant].

Proof: Define $\bar{f} = (f + f^\star)/2$ for any $f \in \mathcal{M}$, and define the empirical process $\nu_n(g) = n^{-1/2} \sum_{k=1}^n \{g(X_k) - \mathbf{E}[g(X_k)]\}$. Using concavity of $\log x$, we have

$$\sum_{j=1}^k \log\left(\frac{f(X_j)}{f^\star(X_j)}\right) \le 2k^{1/2}\nu_k(\log(\bar{f}/f^\star)) - 2kD(f^\star\|\bar{f})$$

where $D(f^\star\|f) = \int \log(f^\star/f) f^\star d\mu$ is relative entropy. As $D(f^\star\|f) \ge h(f,f^\star)^2$, we can estimate

$$\begin{aligned}
&\mathbf{P}\left[\max_{n\le k\le 2n}\sup_{f\in\mathcal{M}}\sum_{j=1}^k \log\left(\frac{f(X_j)}{f^\star(X_j)}\right) \ge \alpha\right] \\
&\le \mathbf{P}\left[\max_{n\le k\le 2n}\sup_{f\in\mathcal{M}}\{\sqrt{k}\,\nu_k(\log(\bar{f}/f^\star)) - kh(\bar{f},f^\star)^2\} \ge \frac{\alpha}{2}\right] \\
&\le \sum_{s=0}^S \mathbf{P}\left[\max_{n\le k\le 2n}\sup_{f\in\mathcal{M}^s}|\sqrt{k}\,\nu_k(\log(\bar{f}/f^\star))| \ge \alpha 2^{s-1}\right] \\
&\le 3\sum_{s=0}^S \max_{n\le k\le 2n}\mathbf{P}\left[\sup_{f\in\mathcal{M}^s}|\nu_k(\log(\sqrt{\bar{f}/f^\star}))| \ge \alpha 2^{s-5}/\sqrt{n}\right]
\end{aligned}$$

where $\mathcal{M}^s = \{f \in \mathcal{M} : \alpha 2^{s-1} < nh(\bar{f},f^\star)^2 \le \alpha 2^s\}$, $1 \le s \le S$, $S = \min\{s : \alpha 2^s n^{-1} > 2\}$, $\mathcal{M}^0 = \{f \in \mathcal{M} : nh(\bar{f},f^\star)^2 \le \alpha\}$ and we have used Lemma A.2 below for the last inequality. The remainder of the proof is identical to that of [23, Th. 7.4], provided we show that for $\bar{\mathcal{H}}(\varepsilon) = \{\sqrt{\bar{f}/f^\star} : f \in \mathcal{M},\ h(\bar{f},f^\star) \le \varepsilon\}$

$$\mathcal{N}(\bar{\mathcal{H}}(\varepsilon),\delta) \le \left(\frac{2\sqrt{2}K\varepsilon}{\delta}\right)^p.$$

To this end, fix $\delta \le \varepsilon$, and note that $h(f,f^\star) \le 4h(\bar{f},f^\star)$ by [23, Lemma 4.2], so that $\{f \in \mathcal{M} : h(\bar{f},f^\star) \le \varepsilon\} \subseteq \{f \in \mathcal{M} : h(f,f^\star) \le 4\varepsilon\}$. By assumption, there exist $N \le (2\sqrt{2}K\varepsilon/\delta)^p$ and functions $g_1,\ldots,g_N,h_1,\ldots,h_N$ such that $\|h_i - g_i\|_2 \le \delta\sqrt{2}$ for every i, and for every $u \in \mathcal{H}(4\varepsilon)$ there is an i such that $g_i \le u \le h_i$. But for every $f \in \mathcal{M}$ such that $h(\bar{f},f^\star) \le \varepsilon$, we then have for some i

$$2^{-1/2}\sqrt{g_i^2+1} \le \sqrt{\bar{f}/f^\star} \le 2^{-1/2}\sqrt{h_i^2+1}.$$

Using $|\sqrt{a+c} - \sqrt{b+c}| \le |\sqrt{a} - \sqrt{b}|$ for $a,b,c \ge 0$, we have

$$\left\|2^{-1/2}\sqrt{h_i^2+1} - 2^{-1/2}\sqrt{g_i^2+1}\right\|_2 \le 2^{-1/2}\|h_i - g_i\|_2 \le \delta.$$

The result now follows directly. ∎

The following variant of Etemadi's inequality was used in the proof. The proof follows closely that of the classical Etemadi inequality (see [24, Appendix M19]).

Lemma A.2: Let $\mathcal{Q}$ be a family of measurable functions $f : E \to \mathbb{R}$. Then, we have for every $\alpha > 0$ and $m,n \in \mathbb{N}$, $m \le n$

$$\begin{aligned}
&\mathbf{P}^\star\left[\max_{k=m,\ldots,n}\sup_{f\in\mathcal{Q}}|S_k(f)| \ge 3\alpha\right] \\
&\qquad\qquad \le 3\max_{k=m,\ldots,n}\mathbf{P}^\star\left[\sup_{f\in\mathcal{Q}}|S_k(f)| \ge \alpha\right]
\end{aligned}$$

where $S_n(f) = n^{1/2}\nu_n(f)$.

Proof: Define the stopping time

$$\tau = \inf\left\{k \ge m : \sup_{f\in\mathcal{Q}}|S_k(f)| \ge 3\alpha\right\}.$$

Then

$$\begin{aligned}
&\mathbf{P}^\star\left[\max_{k=m,\ldots,n}\sup_{f\in\mathcal{Q}}|S_k(f)| \ge 3\alpha\right] = \mathbf{P}^\star[\tau \le n] \\
&\le \mathbf{P}^\star\left[\sup_{f\in\mathcal{Q}}|S_n(f)| \ge \alpha\right] + \\
&\qquad \sum_{k=m}^n \mathbf{P}^\star\left[\tau = k \text{ and } \sup_{f\in\mathcal{Q}}|S_n(f)| < \alpha\right].
\end{aligned}$$

But on the event $\{\tau = k \text{ and } \sup_{f\in\mathcal{Q}}|S_n(f)| < \alpha\}$, we have

$$2\alpha \le \sup_{f\in\mathcal{Q}}|S_k(f)| - \sup_{f\in\mathcal{Q}}|S_n(f)| \le \sup_{f\in\mathcal{Q}}|S_k(f) - S_n(f)|.$$

Therefore, we can estimate

$$\begin{aligned}
&\mathbf{P}^\star\left[\max_{k=m,\ldots,n}\sup_{f\in\mathcal{Q}}|S_k(f)| \ge 3\alpha\right] \\
&\le \mathbf{P}^\star\left[\sup_{f\in\mathcal{Q}}|S_n(f)| \ge \alpha\right] + \\
&\qquad \sum_{k=m}^n \mathbf{P}^\star\left[\tau = k \text{ and } \sup_{f\in\mathcal{Q}}|S_n(f) - S_k(f)| \ge 2\alpha\right] \\
&\le \mathbf{P}^\star\left[\sup_{f\in\mathcal{Q}}|S_n(f)| \ge \alpha\right] + \\
&\qquad \max_{k=m,\ldots,n}\mathbf{P}^\star\left[\sup_{f\in\mathcal{Q}}|S_n(f) - S_k(f)| \ge 2\alpha\right]
\end{aligned}$$

where we have used that $\sup_{f\in\mathcal{Q}}|S_n(f) - S_k(f)|$ and $\{\tau = k\}$ are independent to obtain the last inequality. The remainder of the proof is now easily completed. ∎

We can now complete the proof of Theorem II.3.

Proof of Theorem II.3: By assumption, we have $f^\star \in \mathcal{M}_q^n$ for all $q \ge q^\star$ when n is sufficiently large. Then, by Theorem A.1, we have for n sufficiently large

$$\mathbf{P}^\star\left[\max_{n\le k\le 2n}\sup_{f\in\mathcal{M}_q^{2n}}\{\ell_k(f) - \ell_k(f^\star)\} \ge \alpha\right] \le C\,e^{-\alpha/C}$$

for all $\alpha \ge C\eta(q)(1 + \log K(2n))$ and $q \ge q^\star$. Define

$$\Delta_k(q,q^\star) = \sup_{f\in\mathcal{M}_q^k}\ell_k(f) - \sup_{f\in\mathcal{M}_{q^\star}^k}\ell_k(f).$$

Using that $\mathcal{M}_q^k \subseteq \mathcal{M}_q^{2n}$ for $n \le k \le 2n$ and $\ell_k(f^\star) \le \sup_{f\in\mathcal{M}_{q^\star}^k}\ell_k(f)$, we have for n sufficiently large

$$\mathbf{P}^\star\left[\max_{n\le k\le 2n}\sup_{q\ge q^\star}\frac{1}{\eta(q)}\Delta_k(q,q^\star) \ge \alpha\right] \le \sum_{q=q^\star}^\infty C\,e^{-\alpha\eta(q)/C}$$

for all $\alpha \geq C(1+\log K(2n))$. Let $\beta(n)$ be an increasing function. Then, for all n sufficiently large

$$\mathbf{P}^\star\left[\max_{2^n\leq k\leq 2^{n+1}}\frac{1}{\beta(k)}\sup_{q\geq q^\star}\frac{1}{\eta(q)}\Delta_k(q,q^\star)\geq 2C\right]\leq\frac{2C}{n^2}$$

provided that $\beta(2^n)\geq\log K(2^{n+1})\vee\log\log 2^n$. The proof is now easily completed using the Borel–Cantelli lemma. ■

Appendix B
Proof of Theorem II.6

The proof of Theorem II.6 is based on a sequence of auxiliary results. First, we will need a compact law of iterated logarithm for the Strassen functional

$$I_n(g)=\frac{1}{\sqrt{2n\log\log n}}\sum_{i=1}^n\{g(X_i)-\mathbf{E}^\star(g(X_1))\}.$$

We state the requisite result for future reference.

Theorem B.1: Let $\mathcal{Q}$ be a family of measurable functions from E to $\mathbb{R}$ such that

$$\int_0^1\sqrt{\log\mathcal{N}(\mathcal{Q},u)}\,du<\infty.$$

Then, $\mathbf{P}^\star$-a.s., the sequence $(I_n)_{n\geq 0}$ is relatively compact in $\ell_\infty(\mathcal{Q})$, and its set of cluster points coincides precisely with the set $\mathcal{K}=\{f\mapsto\langle f,g\rangle : g\in L_0^2(f^\star d\mu)\}$.

Proofs of this result can be found in [25, Th. 4.2] or in [26, Th. 9]. We will also need the following simple well-known fact, whose proof is omitted.

Lemma B.2: Let $(X_i)_{i\geq 1}$ be an i.i.d. sequence such that $\mathbf{E}[|X_1|^p]<\infty$. Then, $n^{-1/p}\max_{i=1,\ldots,n}|X_i|\to 0$ a.s.

Finally, we will need the following likelihood inequality that relates the log-likelihood ratio $\ell_n(f)-\ell_n(f^\star)$ to the empirical process. Related inequalities appear in [6], [27], and [28], but the following form is perhaps the most natural.

Lemma B.3: For any probability density $f\neq f^\star$

$$\ell_n(f)-\ell_n(f^\star)\leq|\nu_n(d_f)|^2$$

where $\nu_n(g)=n^{-1/2}\sum_{k=1}^n\{g(X_k)-\mathbf{E}^\star[g(X_k)]\}$.

Proof: Note that

$$h(f,f^\star)^2=2-\int 2\sqrt{ff^\star}\,d\mu=-2\,h(f,f^\star)\,\mathbf{E}^\star(d_f(X_1)).$$

Using $\log(1+x)\leq x$, we can estimate

$$\begin{aligned}\ell_n(f)-\ell_n(f^\star)&=\sum_{i=1}^n 2\,\log(1+h(f,f^\star)\,d_f(X_i))\\&\leq\sum_{i=1}^n 2\,h(f,f^\star)\,d_f(X_i)\\&=2\,\nu_n(d_f)\,h(f,f^\star)\,\sqrt{n}-h(f,f^\star)^2\,n\\&\leq\sup_{p\in\mathbb{R}}\left\{2\,\nu_n(d_f)\,p-p^2\right\}.\end{aligned}$$

The proof is easily completed. ■

We can now obtain the following asymptotic expansion of the log-likelihood, which provides a pathwise counterpart to the weak convergence theory in [27] and [28].

Proposition B.4: Let $q\geq q^\star$. Assume that

$$\int_0^1\sqrt{\log\mathcal{N}(\mathcal{D}_q,u)}\,du<\infty.$$

Moreover, suppose that $|d|\leq D$ for all $d\in\mathcal{D}_q$ with $D\in L^{2+\alpha}(f^\star d\mu)$ for some $\alpha>0$. Then

$$\begin{aligned}\sup_{f\in\mathcal{M}_q(4\sqrt{\log\log n/n})}&\left\{2\,I_n(d_f)\,h(f,f^\star)\sqrt{\frac{2n}{\log\log n}}\right.\\&\left.-h(f,f^\star)^2\,\frac{2n}{\log\log n}\right\}\\-\frac{1}{\log\log n}&\left\{\sup_{f\in\mathcal{M}_q}\ell_n(f)-\ell_n(f^\star)\right\}\overset{n\to\infty}{\to}0\quad\mathbf{P}^\star\text{-a.s.}\end{aligned}$$

where we have defined $\mathcal{M}_q(\varepsilon)=\{f\in\mathcal{M}_q : h(f,f^\star)\leq\varepsilon\}$.

Proof: We proceed in several steps.

Step 1 (Localization): As $q\geq q^\star$ (hence $f^\star\in\mathcal{M}_q$), clearly

$$\sup_{f\in\mathcal{M}_q}\ell_n(f)-\ell_n(f^\star)=\sup_{f\in\mathcal{M}_q:\ell_n(f)-\ell_n(f^\star)\geq 0}\{\ell_n(f)-\ell_n(f^\star)\}.$$

Now note that, as in the proof of Lemma B.3,

$$\ell_n(f)-\ell_n(f^\star)\leq 2\,\nu_n(d_f)\,h(f,f^\star)\,\sqrt{n}-h(f,f^\star)^2\,n.$$

Therefore, we can estimate

$$\begin{aligned}&\sup_{f\in\mathcal{M}_q:\ell_n(f)-\ell_n(f^\star)\geq 0}h(f,f^\star)\\&\quad\leq\sup_{f\in\mathcal{M}_q:\ell_n(f)-\ell_n(f^\star)\geq 0}\left\{h(f,f^\star)+\frac{\ell_n(f)-\ell_n(f^\star)}{n\,h(f,f^\star)}\right\}\\&\quad\leq\frac{2}{\sqrt{n}}\sup_{f\in\mathcal{M}_q:\ell_n(f)-\ell_n(f^\star)\geq 0}\nu_n(d_f)\\&\quad\leq\sqrt{\frac{8\log\log n}{n}}\sup_{d\in\mathcal{D}_q}I_n(d).\end{aligned}$$

Now note that we can estimate

$$\begin{aligned}\sup_{d\in\mathcal{D}_q}I_n(d)\leq\inf_{g\in L_0^2(f^\star d\mu)}&\sup_{d\in\mathcal{D}_q}|I_n(d)-\langle d,g\rangle|\\&+\sup_{d\in\mathcal{D}_q}\sup_{g\in L_0^2(f^\star d\mu)}\langle d,g\rangle.\end{aligned}$$

The first term on the right converges to zero $\mathbf{P}^\star$-a.s. as $n\to\infty$ by Theorem B.1, while the second term is easily seen to equal $\sup_{d\in\mathcal{D}_q}\|d-\langle 1,d\rangle\|_2\leq 1$. Therefore

$$\sup_{f\in\mathcal{M}_q:\ell_n(f)-\ell_n(f^\star)\geq 0}h(f,f^\star)\leq(1+\varepsilon)\sqrt{\frac{8\log\log n}{n}}$$

eventually as $n\to\infty$ $\mathbf{P}^\star$-a.s. for any $\varepsilon>0$. In particular

$$\begin{aligned}\{f\in\mathcal{M}_q:\ell_n(f)-\ell_n(f^\star)\geq 0\}\subseteq&\\\left\{f\in\mathcal{M}_q : h(f,f^\star)\leq 4\sqrt{\log\log n/n}\right\}&\end{aligned}$$

eventually as $n \to \infty$ $\mathbf{P}^\star$-a.s. This implies that

$$\sup_{f\in\mathcal{M}_q} \ell_n(f) - \ell_n(f^\star) \leq \sup_{f\in\mathcal{M}_q : h(f,f^\star)\leq 4\sqrt{\log\log n/n}} \{\ell_n(f) - \ell_n(f^\star)\}$$

eventually as $n \to \infty$ $\mathbf{P}^\star$-a.s. But the reverse inequality clearly holds for all $n \geq 0$, so that in fact

$$\sup_{f\in\mathcal{M}_q} \ell_n(f) - \ell_n(f^\star) = \sup_{f\in\mathcal{M}_q(4\sqrt{\log\log n/n})} \{\ell_n(f) - \ell_n(f^\star)\}$$

eventually as $n \to \infty$ $\mathbf{P}^\star$-a.s.

Step 2 (Taylor Expansion): Taylor expansion gives $2\log(1+x) = 2x - x^2 + x^2 R(x)$, where $R(x) \to 0$ as $x \to 0$. Thus, we can write, for any $f \in \mathcal{M}_q$

$$\begin{aligned}\ell_n(f) - \ell_n(f^\star) &= \sum_{i=1}^n 2\,\log(1 + h(f,f^\star)\,d_f(X_i)) = \\ &\quad 2\,h(f,f^\star)\sum_{i=1}^n \left\{ d_f(X_i) + \frac{1}{2}\,h(f,f^\star)\right\} \\ &\quad - h(f,f^\star)^2 \sum_{i=1}^n (d_f(X_i))^2 - n\,h(f,f^\star)^2 \\ &\quad + h(f,f^\star)^2 \sum_{i=1}^n (d_f(X_i))^2 R(h(f,f^\star)\,d_f(X_i)).\end{aligned}$$

Using that $\mathbf{E}^\star(d_f(X_1)) = -h(f,f^\star)/2$, we therefore have

$$\begin{aligned}\frac{1}{\log\log n}\,\{\ell_n(f) - \ell_n(f^\star)\} &= R_{f,n}\,\frac{n\,h(f,f^\star)^2}{\log\log n} \\ &\quad + 2\,I_n(d_f)\,h(f,f^\star)\sqrt{\frac{2n}{\log\log n}} - h(f,f^\star)^2\,\frac{2n}{\log\log n}\end{aligned}$$

where we have defined

$$\begin{aligned}R_{f,n} &= \frac{1}{n}\sum_{i=1}^n \{1 - (d_f(X_i))^2\} \\ &\quad + \frac{1}{n}\sum_{i=1}^n (d_f(X_i))^2 R(h(f,f^\star)\,d_f(X_i)).\end{aligned}$$

It follows easily that

$$\begin{aligned}&\left| \sup_{f\in\mathcal{M}_q(4\sqrt{\log\log n/n})} \left\{ 2\,I_n(d_f)\,h(f,f^\star)\sqrt{\frac{2n}{\log\log n}} \right.\right. \\ &\qquad \left. - h(f,f^\star)^2\,\frac{2n}{\log\log n}\right\} \\ &\qquad \left. - \frac{1}{\log\log n}\left\{ \sup_{f\in\mathcal{M}_q} \ell_n(f) - \ell_n(f^\star)\right\}\right| \\ &\leq \sup_{f\in\mathcal{M}_q(4\sqrt{\log\log n/n})} |R_{f,n}|\,\frac{n\,h(f,f^\star)^2}{\log\log n} \\ &\leq 16 \sup_{f\in\mathcal{M}_q(4\sqrt{\log\log n/n})} |R_{f,n}|\end{aligned}$$

eventually as $n \to \infty$ $\mathbf{P}^\star$-a.s.

Step 3 (End of Proof): We can easily estimate

$$\begin{aligned}\sup_{f\in\mathcal{M}_q(4\sqrt{\log\log n/n})} |R_{f,n}| &\leq \sup_{f\in\mathcal{M}_q} \left| \frac{1}{n}\sum_{i=1}^n \{d_f(X_i)^2 - 1\}\right| \\ &\quad + \left(\sup_{|x|\leq 4\sqrt{\log\log n/n}\,\max_{i\leq n} D(X_i)} |R(x)|\right) \frac{1}{n}\sum_{i=1}^n (D(X_i))^2.\end{aligned}$$

As $\mathcal{N}(\mathcal{D}_q,\delta) < \infty$ for every $\delta > 0$, the class $\{d^2 : d \in \mathcal{D}_q\}$ can be covered by a finite number of brackets with arbitrary small $L^1(f^\star d\mu)$-norm and is therefore $\mathbf{P}^\star$-Glivenko–Cantelli. Moreover, by construction, $\mathbf{E}^\star[(d_f(X_i))^2] = 1$ for all $f \in \mathcal{M}_q$. Therefore, the first term in this expression converges to zero as $n \to \infty$ $\mathbf{P}^\star$-a.s. On the other hand, by Lemma B.2 and the fact that $D \in L^{2+\alpha}(f^\star d\mu)$, we have $\mathbf{P}^\star$-a.s.

$$\begin{aligned}\sqrt{\log\log n/n}\,\max_{i=1,\ldots,n} D(X_i) &= \\ \frac{\sqrt{\log\log n}}{n^{\alpha/2(2+\alpha)}}\,n^{-1/(2+\alpha)}\,\max_{i=1,\ldots,n} D(X_i) &\overset{n\to\infty}{\to} 0.\end{aligned}$$

Therefore, the second term converges to zero also, and the proof is evidently complete. ■

Proposition B.5: Let $q \geq q^\star$. Assume that

$$\int_0^1 \sqrt{\log \mathcal{N}(\mathcal{D}_q,u)}\,du < \infty.$$

Moreover, suppose that $|d| \leq D$ for all $d \in \mathcal{D}_q$ with $D \in L^{2+\alpha}(f^\star d\mu)$ for some $\alpha > 0$. Then

$$\varliminf_{n\to\infty} \left\{ \sup_{d\in\bar{\mathcal{D}}_q} (I_n(d))_+^2 - \frac{1}{\log\log n}\left\{ \sup_{f\in\mathcal{M}_q} \ell_n(f) - \ell_n(f^\star)\right\}\right\}$$

is nonnegative $\mathbf{P}^\star$-a.s.

Proof: By Proposition B.4, we have

$$\begin{aligned}&\varliminf_{n\to\infty} \left\{ \sup_{d\in\bar{\mathcal{D}}_q} (I_n(d))_+^2 - \frac{1}{\log\log n}\left\{ \sup_{f\in\mathcal{M}_q} \ell_n(f) - \ell_n(f^\star)\right\}\right\} \\ &\geq \varliminf_{n\to\infty} \left\{ \sup_{d\in\bar{\mathcal{D}}_q} (I_n(d))_+^2 \right. \\ &\qquad \left. - \sup_{f\in\mathcal{M}_q(4\sqrt{\log\log n/n})} \sup_{p\geq 0} \left\{ 2\,I_n(d_f)\,p - p^2\right\}\right\} \\ &= \varliminf_{n\to\infty} \left\{ \sup_{d\in\bar{\mathcal{D}}_q} (I_n(d))_+^2 - \sup_{f\in\mathcal{M}_q(4\sqrt{\log\log n/n})} (I_n(d_f))_+^2 \right\}.\end{aligned}$$

Suppose that the right-hand side is negative with positive probability. Then, there exists $\varepsilon > 0$ and a sequence $\tau_n \uparrow \infty$ of random times such that

$$\sup_{d\in\bar{\mathcal{D}}_q} (I_{\tau_n}(d))_+^2 - \sup_{f\in\mathcal{M}_q(4\sqrt{\log\log \tau_n/\tau_n})} (I_{\tau_n}(d_f))_+^2 \leq -\varepsilon \tag{B1}$$

for all n with positive probability. We will show that this entails a contradiction.

By Theorem B.1 (which can be applied here as $\mathcal{N}(\mathcal{D}_q,\delta) = \mathcal{N}(\mathrm{cl}\,\mathcal{D}_q,\delta)$ for all $\delta>0$), the process $(I_{\tau_n})_{n\geq 0}$ is $\mathbf{P}^\star$-a.s. relatively compact in $\ell_\infty(\mathrm{cl}\,\mathcal{D}_q)$ with

$$\inf_{g\in L_0^2(f^\star d\mu)} \sup_{d\in \mathrm{cl}\,\mathcal{D}_q} |I_{\tau_n}(d) - \langle d,g\rangle| \stackrel{n\to\infty}{\to} 0 \quad \mathbf{P}^\star\text{-a.s.} \qquad \text{(B2)}$$

Then, there is a set of positive probability on which (B1) and (B2) hold simultaneously. We now concentrate our attention on a single sample path in this set. For any such path, we can clearly find a further subsequence $\sigma_n \uparrow \infty$ such that $\sup_{d\in\mathrm{cl}\,\mathcal{D}_q}|I_{\sigma_n}(d)-\langle d,g\rangle|\to 0$ as $n\to\infty$ for some element $g\in L_0^2(f^\star d\mu)$. Therefore, we obtain

$$\sup_{d\in\mathrm{cl}\,\mathcal{D}_q} |(I_{\sigma_n}(d))_+^2 - (\langle d,g\rangle)_+^2| \leq \sup_{d\in\mathrm{cl}\,\mathcal{D}_q} |I_{\sigma_n}(d)-\langle d,g\rangle|^2 + 2\sup_{d\in\mathrm{cl}\,\mathcal{D}_q}|I_{\sigma_n}(d)-\langle d,g\rangle| \sup_{d\in\mathrm{cl}\,\mathcal{D}_q}|\langle d,g\rangle| \stackrel{n\to\infty}{\to} 0$$

where we have used the elementary estimate

$$\begin{aligned}|a_+^2-b_+^2| &= |a_+-b_+|(a_++b_+)\\ &\leq |a_+-b_+|(|a_+-b_+|+2b_+) \leq |a-b|(|a-b|+2|b|)\end{aligned}$$

for any $a,b\in\mathbb{R}$, and the fact that $\sup_{d\in\mathrm{cl}\,\mathcal{D}_q}|\langle d,g\rangle| \leq \sup_{d\in\mathrm{cl}\,\mathcal{D}_q}\|d\|_2\|g\|_2\leq 1$. Thus, (B1) gives

$$\varliminf_{n\to\infty}\left\{\sup_{d\in\bar{\mathcal{D}}_q}(\langle d,g\rangle)_+^2 - \sup_{f\in\mathcal{M}_q(4\sqrt{\log\log\sigma_n/\sigma_n})}(\langle d_f,g\rangle)_+^2\right\} = \varliminf_{n\to\infty}\left\{\sup_{d\in\bar{\mathcal{D}}_q}(I_{\sigma_n}(d))_+^2 - \sup_{f\in\mathcal{M}_q(4\sqrt{\log\log\sigma_n/\sigma_n})}(I_{\sigma_n}(d_f))_+^2\right\} \leq -\varepsilon.$$

But as the map $d\mapsto\langle d,g\rangle$ is continuous in $L^2(f^\star d\mu)$ and $\mathrm{cl}\,\mathcal{D}_q(4\sqrt{\log\log\sigma_n/\sigma_n})$ is compact in $L^2(f^\star d\mu)$ (this follows from $\mathcal{N}(\mathcal{D}_q,\delta)<\infty$ for all $\delta>0$), we have

$$\sup_{f\in\mathcal{M}_q(4\sqrt{\log\log\sigma_n/\sigma_n})}(\langle d_f,g\rangle)_+^2 = \sup_{d\in\mathrm{cl}\,\mathcal{D}_q(4\sqrt{\log\log\sigma_n/\sigma_n})}(\langle d,g\rangle)_+^2 \stackrel{n\to\infty}{\to} \sup_{d\in\bigcap_{n\geq 0}\mathrm{cl}\,\mathcal{D}_q(4\sqrt{\log\log\sigma_n/\sigma_n})}(\langle d,g\rangle)_+^2 = \sup_{d\in\bar{\mathcal{D}}_q}(\langle d,g\rangle)_+^2.$$

Thus, we have a contradiction, completing the proof. ■

We now obtain a converse to the previous result.

Proposition B.6: Let $q\geq q^\star$. Assume that

$$\int_0^1\sqrt{\log\mathcal{N}(\mathcal{D}_q,u)}\,du<\infty.$$

Moreover, suppose that $|d|\leq D$ for all $d\in\mathcal{D}_q$ with $D\in L^{2+\alpha}(f^\star d\mu)$ for some $\alpha>0$. Then

$$\overline{\lim_{n\to\infty}}\left\{\sup_{d\in\bar{\mathcal{D}}_q^c}(I_n(d))_+^2 - \frac{1}{\log\log n}\left\{\sup_{f\in\mathcal{M}_q}\ell_n(f)-\ell_n(f^\star)\right\}\right\}$$

is nonpositive $\mathbf{P}^\star$-a.s.

Proof: Suppose the result is false. By Proposition B.4, there is $\varepsilon>0$ and a sequence $\tau_n\uparrow\infty$ of random times so that

$$\sup_{d\in\bar{\mathcal{D}}_q^c}(I_{\tau_n}(d))_+^2 - \sup_{f\in\mathcal{M}_q(4\sqrt{\log\log\tau_n/\tau_n})}\left\{-h(f,f^\star)^2\frac{2\tau_n}{\log\log\tau_n} + 2\,I_{\tau_n}(d_f)\,h(f,f^\star)\sqrt{\frac{2\tau_n}{\log\log\tau_n}}\right\}\geq\varepsilon$$

for all n with positive probability. Proceeding as in the proof of Proposition B.5, we can then show that there is a sequence of times $\sigma_n\uparrow\infty$ and some $g\in L_0^2(f^\star d\mu)$ such that

$$\overline{\lim_{n\to\infty}}\left\{\sup_{d\in\bar{\mathcal{D}}_q^c}(\langle d,g\rangle)_+^2 - \sup_{f\in\mathcal{M}_q(4\sqrt{\log\log\sigma_n/\sigma_n})}\left\{-h(f,f^\star)^2\frac{2\sigma_n}{\log\log\sigma_n} + 2\,\langle d_f,g\rangle\,h(f,f^\star)\sqrt{\frac{2\sigma_n}{\log\log\sigma_n}}\right\}\right\}\geq\varepsilon.$$

We will show that this entails a contradiction.

Let $d_0\in\bar{\mathcal{D}}_q$ be a continuously accessible point. Then, there exists an $\alpha_0>0$ (depending on d_0) and a path $(f_\alpha)_{\alpha\in]0,\alpha_0]}$ such that $h(f_\alpha,f^\star)=\alpha$ for all $\alpha\in]0,\alpha_0]$ and $d_{f_\alpha}\to d_0$ in $L^2(f^\star d\mu)$ as $\alpha\to 0$. Now choose the sequence

$$\alpha_n = \{(\langle d_0,g\rangle)_+ + \sigma_n^{-1}\}\sqrt{\frac{\log\log\sigma_n}{2\sigma_n}}.$$

As $(\langle d_0,g\rangle)_+\leq\|d_0\|_2\|g\|_2\leq 1$, we clearly have

$$0<\alpha_n<\alpha_0\wedge 4\sqrt{\log\log\sigma_n/\sigma_n}$$

for all n sufficiently large. In particular, it follows that $f_{\alpha_n}\in\mathcal{M}_q(4\sqrt{\log\log\sigma_n/\sigma_n})$, so that

$$\sup_{f\in\mathcal{M}_q(4\sqrt{\log\log\sigma_n/\sigma_n})}\left\{2\,\langle d_f,g\rangle\,h(f,f^\star)\sqrt{\frac{2\sigma_n}{\log\log\sigma_n}} - h(f,f^\star)^2\frac{2\sigma_n}{\log\log\sigma_n}\right\} \geq 2\,\langle d_{f_{\alpha_n}},g\rangle\,\{(\langle d_0,g\rangle)_+ + \sigma_n^{-1}\} - \{(\langle d_0,g\rangle)_+ + \sigma_n^{-1}\}^2.$$

Therefore, we have

$$\overline{\lim_{n\to\infty}}\left\{\sup_{d\in\bar{\mathcal{D}}_q^c}(\langle d,g\rangle)_+^2 - \sup_{f\in\mathcal{M}_q(4\sqrt{\log\log\sigma_n/\sigma_n})}\left\{-h(f,f^\star)^2\frac{2\sigma_n}{\log\log\sigma_n} + 2\,\langle d_f,g\rangle\,h(f,f^\star)\sqrt{\frac{2\sigma_n}{\log\log\sigma_n}}\right\}\right\} \leq \sup_{d\in\bar{\mathcal{D}}_q^c}(\langle d,g\rangle)_+^2 - (\langle d_0,g\rangle)_+^2$$

for any continuously accessible element $d_0 \in \bar{\mathcal{D}}_q$. But clearly, we can choose d_0 to make the right-hand side of this expression arbitrarily small. Thus, we have the desired contradiction. ■

We can now complete the proof of Theorem II.6.

Proof of Theorem II.6: We obtain separately the lower and upper bounds.

Lower Bound: By Propositions B.5 and B.6, we have

$$\varlimsup_{n\to\infty} \frac{1}{\log\log n}\left\{\sup_{f\in\mathcal{M}_q} \ell_n(f) - \sup_{f\in\mathcal{M}_p} \ell_n(f)\right\} \geq \varlimsup_{n\to\infty}\left\{\sup_{d\in\bar{\mathcal{D}}_q^c} (I_n(d))_+^2 - \sup_{d\in\bar{\mathcal{D}}_p} (I_n(d))_+^2\right\} \quad \mathbf{P}^\star\text{-a.s.}$$

Now fix any $g \in L_0^2(f^\star d\mu)$. By Theorem B.1 (which applies here as $\mathcal{N}(\mathcal{D}_q,\delta) = \mathcal{N}(\operatorname{cl}\mathcal{D}_q,\delta) \geq \mathcal{N}(\bar{\mathcal{D}}_q,\delta)$ for all $\delta > 0$), there is a sequence $\tau_n \uparrow \infty$ of random times such that $I_{\tau_n} \to \langle\cdot, g\rangle$ in $\ell_\infty(\bar{\mathcal{D}}_q)$ $\mathbf{P}^\star$-a.s. Therefore

$$\sup_{d\in\bar{\mathcal{D}}_q^c} (I_{\tau_n}(d))_+^2 - \sup_{d\in\bar{\mathcal{D}}_p} (I_{\tau_n}(d))_+^2 \overset{n\to\infty}{\to} \sup_{d\in\bar{\mathcal{D}}_q^c} (\langle d,g\rangle)_+^2 - \sup_{d\in\bar{\mathcal{D}}_p} (\langle d,g\rangle)_+^2 \quad \mathbf{P}^\star\text{-a.s.}$$

so that certainly

$$\varlimsup_{n\to\infty} \frac{1}{\log\log n}\left\{\sup_{f\in\mathcal{M}_q} \ell_n(f) - \sup_{f\in\mathcal{M}_p} \ell_n(f)\right\} \geq \sup_{d\in\bar{\mathcal{D}}_q^c} (\langle d,g\rangle)_+^2 - \sup_{d\in\bar{\mathcal{D}}_p} (\langle d,g\rangle)_+^2$$

$\mathbf{P}^\star$-a.s. But as this inequality holds for every $g \in L_0^2(f^\star d\mu)$, taking the supremum over g gives the requisite lower bound.

Upper Bound: By Propositions B.5 and B.6, we have

$$\varlimsup_{n\to\infty} \frac{1}{\log\log n}\left\{\sup_{f\in\mathcal{M}_q} \ell_n(f) - \sup_{f\in\mathcal{M}_p} \ell_n(f)\right\} \leq \varlimsup_{n\to\infty}\left\{\sup_{d\in\bar{\mathcal{D}}_q} (I_n(d))_+^2 - \sup_{d\in\bar{\mathcal{D}}_p^c} (I_n(d))_+^2\right\} \quad \mathbf{P}^\star\text{-a.s.}$$

It is elementary that for any $d, d' \in \bar{\mathcal{D}}_q$ and $g \in L_0^2(f^\star d\mu)$

$$\begin{aligned}(I_n(d))_+^2 - (I_n(d'))_+^2 &\leq |(I_n(d))_+^2 - (\langle d,g\rangle)_+^2| + |(I_n(d'))_+^2 - (\langle d',g\rangle)_+^2| \\ &\quad + (\langle d,g\rangle)_+^2 - (\langle d',g\rangle)_+^2 \\ &\leq 2\sup_{d\in\bar{\mathcal{D}}_q} |(I_n(d))_+^2 - (\langle d,g\rangle)_+^2| + (\langle d,g\rangle)_+^2 - (\langle d',g\rangle)_+^2.\end{aligned}$$

Taking the supremum over $d \in \bar{\mathcal{D}}_q$ and the infimum over $d' \in \bar{\mathcal{D}}_p^c$, we find that

$$\begin{aligned}&\sup_{d\in\bar{\mathcal{D}}_q} (I_n(d))_+^2 - \sup_{d\in\bar{\mathcal{D}}_p^c} (I_n(d))_+^2 \\ &\quad\leq 2\sup_{d\in\bar{\mathcal{D}}_q} |(I_n(d))_+^2 - (\langle d,g\rangle)_+^2| + \sup_{d\in\bar{\mathcal{D}}_q} (\langle d,g\rangle)_+^2 - \sup_{d\in\bar{\mathcal{D}}_p^c} (\langle d,g\rangle)_+^2 \\ &\quad\leq 2\sup_{d\in\bar{\mathcal{D}}_q} |(I_n(d))_+^2 - (\langle d,g\rangle)_+^2| \\ &\qquad + \sup_{g\in L_0^2(f^\star d\mu)}\left\{\sup_{d\in\bar{\mathcal{D}}_q} (\langle d,g\rangle)_+^2 - \sup_{d\in\bar{\mathcal{D}}_p^c} (\langle d,g\rangle)_+^2\right\}.\end{aligned}$$

But as this holds for any $g \in L_0^2(f^\star d\mu)$, we finally obtain

$$\begin{aligned}&\sup_{d\in\bar{\mathcal{D}}_q} (I_n(d))_+^2 - \sup_{d\in\bar{\mathcal{D}}_p^c} (I_n(d))_+^2 \\ &\quad\leq 2\inf_{g\in L_0^2(f^\star d\mu)}\sup_{d\in\bar{\mathcal{D}}_q} |(I_n(d))_+^2 - (\langle d,g\rangle)_+^2| \\ &\qquad + \sup_{g\in L_0^2(f^\star d\mu)}\left\{\sup_{d\in\bar{\mathcal{D}}_q} (\langle d,g\rangle)_+^2 - \sup_{d\in\bar{\mathcal{D}}_p^c} (\langle d,g\rangle)_+^2\right\}.\end{aligned}$$

It follows as in the proof of Proposition B.5 that the first term in this expression converges to zero $\mathbf{P}^\star$-a.s. The requisite upper bound follows immediately. ■

Finally, we now complete the proof of Corollary II.7.

Proof of Corollary II.7: It evidently suffices to prove that

$$\Gamma := \sup_{g\in L_0^2(f^\star d\mu)}\left\{\sup_{d\in\bar{\mathcal{D}}_q^c} (\langle d,g\rangle)_+^2 - \sup_{d\in\bar{\mathcal{D}}_{q^\star}} (\langle d,g\rangle)_+^2\right\} > 0. \tag{B3}$$

To this end, note that by direct computation

$$\langle 1, d_f\rangle = \frac{\int\sqrt{ff^\star}\,d\mu - 1}{h(f,f^\star)} = -\frac{h(f,f^\star)}{2}.$$

Choose $(f_n)_{n\geq 0} \subset \mathcal{M}_q\backslash\{f^\star\}$ such that $h(f_n,f^\star) \to 0$ and $d_{f_n} \to d_0 \in \bar{\mathcal{D}}_q$; then

$$\langle 1, d_0\rangle = \lim_{n\to\infty}\langle 1, d_{f_n}\rangle = -\lim_{n\to\infty}\frac{h(f_n,f^\star)}{2} = 0.$$

Moreover, it is immediate that $\|d_0\|_2 \leq 1$. We have, therefore, shown that $\bar{\mathcal{D}}_q \subset L_0^2(f^\star d\mu)$. Now choose $g \in \bar{\mathcal{D}}_q^c\backslash\bar{\mathcal{D}}_{q^\star}$. As $\bar{\mathcal{D}}_{q^\star}$ is closed, it follows directly that

$$\sup_{d\in\bar{\mathcal{D}}_q^c} (\langle d,g\rangle)_+^2 = 1, \qquad \sup_{d\in\bar{\mathcal{D}}_{q^\star}} (\langle d,g\rangle)_+^2 < 1.$$

Therefore, (B3) holds, and the proof is complete. ■

Appendix C
Proof of Theorem III.1

The proof of Theorem III.1 is based on Theorem II.3 and the following result.

Proposition C.1: Let $\mathcal{M}^n$ for $n \geq 1$ be a family of strictly positive probability densities with respect to a reference measure μ such that $\mathcal{M}^n \subseteq \mathcal{M}^{n+1}$ for all n. Define $\mathcal{M} = \bigcup_n \mathcal{M}^n$, and let $f^\star$ be another probability density with respect to μ such that $f^\star \notin \operatorname{cl}\mathcal{M}$, where $\operatorname{cl}\mathcal{M}$ denotes the $L^1(d\mu)$-closure of $\mathcal{M}$. Let $\mathcal{H}^n = \{\sqrt{f/f^\star} : f \in \mathcal{M}^n\}$, and suppose there exist $K(n) \geq 1$ and $p \geq 1$ so that

$$\mathcal{N}(\mathcal{H}^n,\delta) \leq \left(\frac{K(n)}{\delta}\right)^p$$

for all $\delta \leq 1$ and $n \geq 1$, where $\mathcal{N}(\mathcal{H}^n, \delta)$ is the minimal number of brackets of $L^2(f^\star d\mu)$-width δ needed to cover $\mathcal{H}^n$. Let $(X_i)_{i\in\mathbb{N}}$ be i.i.d. with distribution $f^\star d\mu$. If in addition $\log K(n) = o(n)$, then we have

$$\overline{\lim_{n\to\infty}} \sup_{f\in\mathcal{M}^n} \frac{1}{n}\sum_{j=1}^n \log\left(\frac{f(X_j)}{f^\star(X_j)}\right) < 0 \quad \text{a.s.}$$

Proof: As in the proof of Theorem A.1, we have

$$\frac{1}{n}\sum_{j=1}^n \log\left(\frac{f(X_j)}{f^\star(X_j)}\right) \leq 4n^{-1/2}\nu_n(\log(\{\bar{f}/f^\star\}^{1/2})) - 2D(f^\star\|\bar{f}).$$

The following claim will be proved below:

$$\lim_{n\to\infty} \sup_{f\in\mathcal{M}^n} n^{-1/2}\nu_n(\log(\{\bar{f}/f^\star\}^{1/2})) = 0 \quad \text{a.s.}$$

Using the claim, the proof is easily completed: indeed, if the claim holds, then we have a.s.

$$\overline{\lim_{n\to\infty}} \sup_{f\in\mathcal{M}^n} \frac{1}{n}\sum_{j=1}^n \log\left(\frac{f(X_j)}{f^\star(X_j)}\right) \leq -2 \inf_{f\in\mathcal{M}} D(f^\star\|\bar{f}) < 0$$

where the last inequality follows from Pinsker's inequality and $f^\star \notin \operatorname{cl}\mathcal{M}$.

It, therefore, remains to prove the claim. To this end, we apply [23, Th. 5.11] as in the proof of [23, Th. 7.4] (cf., Theorem A.1), which yields

$$\mathbf{P}\left[\sup_{f\in\mathcal{M}^n} |n^{-1/2}\nu_n(\log(\{\bar{f}/f^\star\}^{1/2}))| \geq \alpha\right] \leq C\, e^{-n\alpha^2/C}$$

for every $\alpha > 0$ such that $C\sqrt{p}\,(1+\sqrt{\log K(n)}) \leq \alpha\sqrt{n} \leq 32\sqrt{n}$ and $n \geq 1$, where C is a universal constant. As $\log K(n) = o(n)$, we have

$$\sum_{n\geq 1} \mathbf{P}\left[\sup_{f\in\mathcal{M}^n} |n^{-1/2}\nu_n(\log(\{\bar{f}/f^\star\}^{1/2}))| \geq \alpha\right] < \infty$$

for $0 < \alpha \leq 32$, so the claim follows from Borel–Cantelli. ■

We can now complete the proof of Theorem III.1.

Proof of Theorem III.1: Define

$$\Delta_n(q,q^\star) = \sup_{f\in\mathcal{M}_q^n} \ell_n(f) - \sup_{f\in\mathcal{M}_{q^\star}^n} \ell_n(f).$$

By Theorem II.3 and easy manipulations, $\mathbf{P}^\star$-a.s.

$$\overline{\lim_{n\to\infty}} \sup_{q>q^\star} \frac{1}{\operatorname{pen}(n,q)-\operatorname{pen}(n,q^\star)}\Delta_n(q,q^\star) \leq \lim_{n\to\infty}\sup_{q>q^\star}\frac{\eta(q)\{\log K(2n)\vee\log\log n\}}{\operatorname{pen}(n,q)-\operatorname{pen}(n,q^\star)} \times \overline{\lim_{n\to\infty}}\frac{1}{\log K(2n)\vee\log\log n}\sup_{q>q^\star}\frac{1}{\eta(q)}\Delta_n(q,q^\star) = 0.$$

Therefore, $\mathbf{P}^\star$-a.s. eventually as $n\to\infty$

$$\sup_{f\in\mathcal{M}_q^n} \ell_n(f) - \operatorname{pen}(n,q) < \sup_{f\in\mathcal{M}_{q^\star}^n} \ell_n(f) - \operatorname{pen}(n,q^\star)$$

for all $q > q^\star$. It follows that $\overline{\lim}_{n\to\infty}\hat{q}_n \leq q^\star$ $\mathbf{P}^\star$-a.s., that is, the penalized likelihood order estimator does not asymptotically overestimate the order.

On the other hand, we note that for every $q < q^\star$

$$\overline{\lim_{n\to\infty}}\frac{1}{n}\left\{\sup_{f\in\mathcal{M}_q^n}\ell_n(f) - \sup_{f\in\mathcal{M}_{q^\star}^n}\ell_n(f)\right\} \leq \overline{\lim_{n\to\infty}}\sup_{f\in\mathcal{M}_q^n}\frac{1}{n}\sum_{j=1}^n\log\left(\frac{f(X_j)}{f^\star(X_j)}\right)$$

which is strictly negative $\mathbf{P}^\star$-a.s. by Proposition C.1, where we have used that $\log K(n) = o(n)$ and that $\mathcal{N}(\mathcal{H}_q^n(2),\delta) \leq \mathcal{N}(\mathcal{H}_{q^\star}^n(2),\delta) \leq (2K(n)/\delta)^{\eta(q^\star)}$ for all $\delta \leq 2$ and n sufficiently large. As $\operatorname{pen}(n,q)/n \to 0$ as $n\to\infty$ for $q < q^\star$

$$\overline{\lim_{n\to\infty}}\max_{q<q^\star}\frac{1}{n}\{\Delta_n(q,q^\star) - \operatorname{pen}(n,q) + \operatorname{pen}(n,q^\star)\} < 0$$

$\mathbf{P}^\star$-a.s. In particular, we find that $\mathbf{P}^\star$-a.s. eventually as $n\to\infty$

$$\sup_{f\in\mathcal{M}_q^n}\ell_n(f) - \operatorname{pen}(n,q) < \sup_{f\in\mathcal{M}_{q^\star}^n}\ell_n(f) - \operatorname{pen}(n,q^\star)$$

for all $q < q^\star$. It follows that $\underline{\lim}_{n\to\infty}\hat{q}_n \geq q^\star$ $\mathbf{P}^\star$-a.s., that is, the penalized likelihood order estimator does not asymptotically underestimate the order. ■

Finally, let us prove Corollary III.3.

Proof of Corollary III.3: It is shown in the proof of Corollary II.7 that

$$\Gamma := \sup_{g\in L_0^2(f^\star d\mu)}\left\{\sup_{d\in\bar{\mathcal{D}}_q^c}(\langle d,g\rangle)_+^2 - \sup_{d\in\bar{\mathcal{D}}_{q^\star}}(\langle d,g\rangle)_+^2\right\} > 0.$$

By Theorem II.6, we have $\mathbf{P}^\star$-a.s.

$$\overline{\lim_{n\to\infty}}\frac{1}{\operatorname{pen}(n,q)-\operatorname{pen}(n,q^\star)}\left\{\sup_{f\in\mathcal{M}_q}\ell_n(f) - \sup_{f\in\mathcal{M}_{q^\star}}\ell_n(f)\right\} \geq \frac{1}{C\{\eta(q)-\eta(q^\star)\}} \times \sup_{g\in L_0^2(f^\star d\mu)}\left\{\sup_{d\in\bar{\mathcal{D}}_q^c}(\langle d,g\rangle)_+^2 - \sup_{d\in\bar{\mathcal{D}}_{q^\star}}(\langle d,g\rangle)_+^2\right\}.$$

Therefore, choosing $C < \Gamma/\{\eta(q)-\eta(q^\star)\}$, we find that

$$\sup_{f\in\mathcal{M}_q}\ell_n(f) - \operatorname{pen}(n,q) > \sup_{f\in\mathcal{M}_{q^\star}}\ell_n(f) - \operatorname{pen}(n,q^\star)$$

infinitely often $\mathbf{P}^\star$-a.s., so $\hat{q}_n \neq q^\star$ infinitely often $\mathbf{P}^\star$-a.s. ■

Appendix D
Proof of Proposition III.6

The proofs of consistency in Propositions III.6 and III.7 follow almost immediately from Theorem III.1, Theorem III.4, and Example III.5. Let us begin with Proposition III.7.

Proof of Proposition III.7: By Example III.5, the assumption of Theorem III.1 holds with $\eta(q) = 10(d+1)q+1$ and $\log K(n) = \log C_1^\star + C_2^\star T(n)^2$. The desired consistency results now follow immediately from Theorem III.1. ■

The consistency part of Proposition III.6 follows similarly. The main difficulty here is to establish the condition $\bar{\mathcal{D}}_q^c\backslash\bar{\mathcal{D}}_{q^\star} \neq$

$\emptyset$ of Corollary III.3, which is needed to prove the inconsistency part of Proposition III.6. In the proof of the latter condition, we rely on the geometric results on mixtures established in [22]. In the remainder of this section, we always assume that we are in the setting of Proposition III.6.

Lemma D.1: Suppose that Assumption A holds. Then, we have

$$\bar{\mathcal{D}}_{q^\star} = \left\{ \frac{L}{\|L\|_2} : L = \sum_{i=1}^{q^\star} \left\{ \eta_i \frac{f_{\theta_i^\star}}{f^\star} + \beta_i^* \frac{D_1 f_{\theta_i^\star}}{f^\star} \right\}, \; \eta_i \in \mathbb{R},\ \beta_i \in \mathbb{R}^d,\ \sum_{i=1}^{q^\star} \eta_i = 0 \right\}.$$

Proof: Let $(f_n)_{n\geq 1} \subset \mathcal{M}_{q^\star}$ be such that $h(f_n, f^\star) \to 0$ and $d_{f_n} \to d_0 \in \bar{\mathcal{D}}_{q^\star}$. By [22, Th. 3.10], we may assume without loss of generality that $f_n = \sum_{i=1}^{q^\star} \pi_i^n f_{\theta_i^n}$ with $\theta_i^n \to \theta_i^\star$ and $\pi_i^n \to \pi_i^\star$ for $i = 1, \ldots, q^\star$. Taylor expansion gives

$$\frac{f_n - f^\star}{f^\star} = L_n + R_n, \qquad |R_n| \leq \frac{d}{2} H_2 \sum_{i=1}^{q^\star} \pi_i^n \|\theta_i^n - \theta_i^\star\|^2$$

where

$$L_n = \sum_{i=1}^{q^\star} \left\{ (\pi_i^n - \pi_i^\star) \frac{f_{\theta_i^\star}}{f^\star} + \pi_i^n (\theta_i^n - \theta_i^\star)^* \frac{D_1 f_{\theta_i^\star}}{f^\star} \right\}.$$

Proceeding as in [22, Lemmas 3.15–3.16], we can estimate

$$\left\| d_{f_n} - \frac{L_n}{\|L_n\|_2} \right\|_2 \leq 2\|S\|_4^2 \{2\|S\|_2 + 1\}\, h(f_n, f^\star) + \{\|S\|_2 + 1\} \frac{\|R_n\|_2}{\|L_n\|_2}.$$

But using [22, Th. 3.10], we have for n sufficiently large

$$\|L_n\|_2 \geq \|L_n\|_1 \geq c^\star \sum_{i=1}^{q^\star} \pi_i^n \|\theta_i^n - \theta_i^\star\|.$$

Thus, we have

$$\frac{\|R_n\|_2}{\|L_n\|_2} \leq \frac{d\|H_2\|_2}{2c^\star} \frac{\sum_{i=1}^{q^\star} \pi_i^n \|\theta_i^n - \theta_i^\star\|^2}{\sum_{i=1}^{q^\star} \pi_i^n \|\theta_i^n - \theta_i^\star\|} \leq \frac{d\|H_2\|_2}{2c^\star} \max_{i=1,\ldots,q^\star} \|\theta_i^n - \theta_i^\star\| \overset{n\to\infty}{\to} 0.$$

Therefore, $L_n/\|L_n\|_2 \to d_0$ in $L^2(f^\star d\mu)$. Now define

$$\eta_i^n = \frac{\pi_i^n - \pi_i^\star}{Z_n}, \quad \beta_i^n = \frac{\pi_i^n(\theta_i^n - \theta_i^\star)}{Z_n}$$

$$Z_n = \sum_{i=1}^{q^\star} \{|\pi_i^n - \pi_i^\star| + \|\pi_i^n(\theta_i^n - \theta_i^\star)\|\}.$$

As $\sum_{i=1}^{q^\star} \{|\eta_i^n| + \|\beta_i^n\|\} = 1$ for all n, we may extract a subsequence such that $\eta_i^n \to \eta_i$, $\beta_i^n \to \beta_i$, and $\sum_{i=1}^{q^\star} \{|\eta_i| + \|\beta_i\|\} = 1$. We obtain immediately

$$d_0 = \frac{L}{\|L\|_2}, \qquad L = \sum_{i=1}^{q^\star} \left\{ \eta_i \frac{f_{\theta_i^\star}}{f^\star} + \beta_i^* \frac{D_1 f_{\theta_i^\star}}{f^\star} \right\}.$$

Clearly, $\sum_{i=1}^{q^\star} \eta_i = 0$. Thus, we have shown that any $d_0 \in \bar{\mathcal{D}}_{q^\star}$ has the desired form.

It remains to show that any function of the desired form is in fact an element of $\bar{\mathcal{D}}_{q^\star}$. To this end, fix $\eta_i \in \mathbb{R}$, $\beta_i \in \mathbb{R}^d$ with $\sum_{i=1}^{q^\star} \eta_i = 0$, and define f_t for $t > 0$ as

$$f_t = \sum_{i=1}^{q^\star} (\pi_i^\star + t\eta_i)\, f_{\theta_i^\star + \beta_i t/\pi_i^\star}.$$

Clearly, $f_t \in \mathcal{M}_{q^\star}$ for all t sufficiently small, and $f_t \to f^\star$ as $t \to 0$. But

$$\frac{f_t - f^\star}{t} = \sum_{i=1}^{q^\star} \pi_i^\star \frac{f_{\theta_i^\star + \beta_i t/\pi_i^\star} - f_{\theta_i^\star}}{t} + \sum_{i=1}^{q^\star} \eta_i\, f_{\theta_i^\star + \beta_i t/\pi_i^\star}.$$

Therefore, clearly

$$\frac{1}{t} \frac{f_t - f^\star}{f^\star} \overset{t\to 0}{\to} \sum_{i=1}^{q^\star} \left\{ \eta_i \frac{f_{\theta_i^\star}}{f^\star} + \beta_i^* \frac{D_1 f_{\theta_i^\star}}{f^\star} \right\} = L.$$

Using [22, Lemma 3.15], we obtain

$$\lim_{t\to 0} d_{f_t} = \lim_{t\to 0} \frac{(f_t - f^\star)/t f^\star}{\|(f_t - f^\star)/t f^\star\|_2} = \frac{L}{\|L\|_2}.$$

Thus, any function of the desired form is in $\bar{\mathcal{D}}_{q^\star}$. ∎

Remark D.2: The above proof in fact shows that $\bar{\mathcal{D}}_{q^\star} = \bar{\mathcal{D}}^c_{q^\star}$.

We can now complete the proof of Proposition III.6.

Proof of Proposition III.6: We begin by proving consistency of the penalty $\mathrm{pen}(n, q) = q\,\omega(n)$. Note that by Theorem III.4, the assumption of Corollary III.2 holds with $\eta(q) = 10(d+1)q + 1 \leq 11(d+1)q$. Thus, consistency of $\mathrm{pen}(n, q) = q\,\omega(n)$ follows directly from Corollary III.2 using $\varpi(n) = \omega(n)/11(d+1)$.

To prove that the penalty $\mathrm{pen}(n, q) = C\, q \log\log n$ is inconsistent for $C > 0$ sufficiently small, it suffices to show that $\bar{\mathcal{D}}^c_{q^\star+1} \backslash \bar{\mathcal{D}}_{q^\star}$ is nonempty. Indeed, if this is the case, then we can apply Corollary III.3 with $q = q^\star + 1$, where the requisite entropy assumption follows from Theorem III.1.

Fix $v \in \mathbb{R}^d$, and consider f_t defined for $t > 0$ as follows:

$$f_t = \frac{\pi_1^\star}{2} (f_{\theta_1^\star + vt} + f_{\theta_1^\star - vt}) + \sum_{i=2}^{q^\star} \pi_i^\star f_{\theta_i^\star}.$$

Clearly, $f_t \in \mathcal{M}_{q^\star+1}$ for all t sufficiently small, $f_t \to f^\star$ as $t \to 0$, and

$$\frac{f_t - f^\star}{t^2} = \frac{\pi_1^\star}{2} \frac{f_{\theta_1^\star + vt} - 2 f_{\theta_1^\star} + f_{\theta_1^\star - vt}}{t^2} \overset{t\to 0}{\to} \frac{\pi_1^\star}{2} v^* D_2 f_{\theta_1^\star} v.$$

As in the proof of Lemma D.1, we find that

$$\lim_{t\to 0} d_{f_t} = \lim_{t\to 0} \frac{(f_t - f^\star)/t^2 f^\star}{\|(f_t - f^\star)/t^2 f^\star\|_2} = \frac{v^* D_2 f_{\theta_1^\star} v}{\|v^* D_2 f_{\theta_1^\star} v\|_2} = d_0.$$

By construction, $d_0 \in \bar{\mathcal{D}}^c_{q^\star+1}$. But by [22, Th. 3.10], the functions $f_{\theta_i^\star}$, $D_1 f_{\theta_i^\star}$, and $v^* D_2 f_{\theta_i^\star} v$ ($i = 1, \ldots, q^\star$) are all linearly independent. Together with Lemma D.1, this shows that $d_0 \notin \bar{\mathcal{D}}_{q^\star}$. Thus, $d_0 \in \bar{\mathcal{D}}^c_{q^\star+1} \backslash \bar{\mathcal{D}}_{q^\star}$. ∎

ACKNOWLEDGMENT

The authors would like to thank Michel Ledoux for suggesting some helpful references.

REFERENCES

[1] I. Csiszár and P. C. Shields, "The consistency of BIC Markov order estimator," *Ann. Stat.*, vol. 28, pp. 1601–1619, 2000.
[2] E. J. Hannan and B. G. Quinn, "The determination of the order of an autoregression," *J. Roy. Statist. Soc. Ser. B*, vol. 41, pp. 190–195, 1979.
[3] R. Nishii, "Maximum likelihood principle and model selection when the true model is unspecified," *J. Multivariate Anal.*, vol. 27, pp. 392–403, 1988.
[4] L. Finesso, "Consistent estimation of the order for Markov and hidden Markov chains," Ph.D. dissertation, Univ. Maryland, College Park, 1990.
[5] C. Keribin, "Consistent estimation of the order of mixture models," *Sankhya Ser. A*, vol. 62, pp. 49–66, 2000.
[6] A. Chambaz, "Testing the order of a model," *Ann. Statist.*, vol. 34, pp. 1166–1203, 2006.
[7] J. A. Hartigan, "A failure of likelihood asymptotics for normal mixtures," in *Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer, Vol. II (Berkeley, Calif., 1983)*. Belmont, CA: Wadsworth, 1985, pp. 807–810.
[8] P. Bickel and H. Chernoff, "Asymptotic Distribution of the Likelihood Ratio Statistic in a Prototypical Non Regular Problem," in *Statistics and Probability: A Raghu Raj Bahadur Festschrift*. New Delhi, India: Wiley, 1993, pp. 83–96.
[9] X. Liu and Y. Shao, "Asymptotics for the likelihood ratio test in a two-component normal mixture model," *J. Statist. Plann. Infer.*, vol. 123, no. 1, pp. 61–81, 2004.
[10] J. C. Kieffer, "Strongly consistent code-based identification and order estimation for constrained finite-state model classes," *IEEE Trans. Inf. Theory*, vol. 39, no. 3, pp. 893–902, May 1993.
[11] C.-C. Liu and P. Narayan, "Order estimation and sequential universal data compression of a hidden Markov source by the method of mixtures," *IEEE Trans. Inf. Theory*, vol. 40, no. 4, pp. 1167–1180, Jul. 1994.
[12] E. Gassiat and S. Boucheron, "Optimal error exponents in hidden Markov model order estimation," *IEEE Trans. Inf. Theory*, vol. 48, no. 4, pp. 964–980, Apr. 2003.
[13] O. Cappé, E. Moulines, and T. Rydén, *Inference in Hidden Markov Models*. New York: Springer, 2005.
[14] A. Chambaz, A. Garivier, and E. Gassiat, "A MDL approach to HMM with Poisson and Gaussian emissions. Application to order indentification," *J. Stat. Plann. Inf.*, vol. 139, pp. 962–977, 2009.
[15] J. Rissanen, "Stochastic complexity and modeling," *Ann. Statist.*, vol. 14, pp. 1080–1100, 1986.
[16] A. Barron, J. Rissanen, and B. Yu, "The minimum description length principle in coding and modeling," *IEEE Trans. Inf. Theory*, vol. 44, no. 6, pp. 2743–2760, Oct. 1998.
[17] I. Csiszar, "Large-scale typicality of Markov sample paths and consistency of MDL order estimators," *IEEE Trans. Inf. Theory*, vol. 48, no. 6, pp. 1616–1628, Jun. 2002.
[18] I. Csiszar and Z. Talata, "Context tree estimation for not necessarily finite memory processes, via BIC and MDL," *IEEE Trans. Inf. Theory*, vol. 52, no. 3, pp. 1007–1616, Mar. 2006.
[19] R. van Handel, "On the minimal penalty for Markov order estimation," *Probab. Th. Rel. Fields*, vol. 150, pp. 709–738, 2011.
[20] E. Gassiat and C. Keribin, "The likelihood ratio test for the number of components in a mixture with Markov regime," *ESAIM Probab. Statist.*, vol. 4, pp. 25–52, 2000.
[21] A. W. van der Vaart, *Asymptotic Statistics*. Cambridge, U.K.: Cambridge Univ. Press, 1998.
[22] E. Gassiat and R. van Handel, Preprint "The local geometry of finite mixtures," 2012.
[23] S. A. van de Geer, *Applications of Empirical Process Theory*. Cambridge, U.K.: Cambridge Univ. Press, 2000.
[24] P. Billingsley, *Convergence of Probability Measures*, 2nd ed. New York: Wiley, 1999.
[25] M. Ossiander, "A central limit theorem under metric entropy with L_2 bracketing," *Ann. Probab.*, vol. 15, pp. 897–919, 1987.
[26] M. Ledoux and M. Talagrand, "Comparison theorems, random geometry and some limit theorems for empirical processes," *Ann. Probab.*, vol. 17, pp. 596–631, 1989.
[27] E. Gassiat, "Likelihood ratio inequalities with applications to various mixtures," *Ann. Inst. H. Poincaré Probab. Statist.*, vol. 38, pp. 897–906, 2002.
[28] X. Liu and Y. Shao, "Asymptotics for likelihood ratio tests under loss of identifiability," *Ann. Statist.*, vol. 31, pp. 807–832, 2003.

Elisabeth Gassiat received the Ph.D. degree from Université Paris-Sud in 1988, and is currently Professor in the Mathematic Department at Université Paris-Sud. Her main research interest is in statistical theory, including semi-parametric statistics, mixture and hidden Markov modeling, Bayesian inference.

Ramon van Handel received the Ph.D. degree from the California Institute of Technology in 2007, and is currently on the faculty of the School of Engineering and Applied Science at Princeton University. His main research interests are in probability theory and related fields.