Frequentist validity of Bayesian limits

Kleijn, B.J.K.

On the frequentist validity of Bayesian limits

B. J. K. Kleijn*

Korteweg-de Vries Institute for Mathematics, University of Amsterdam

November 2017

Abstract

To the frequentist who computes posteriors, not all priors are useful asymptotically: in this paper Schwartz's 1965 Kullback-Leibler condition is generalised to enable frequentist interpretation of convergence of posterior distributions with the complex models and often dependent datasets in present-day statistical applications. We prove four simple and fully general frequentist theorems, for posterior consistency; for posterior rates of convergence; for consistency of the Bayes factor in hypothesis testing or model selection; and a theorem to obtain confidence sets from credible sets. The latter has a significant methodological consequence in frequentist uncertainty quantification: use of a suitable prior allows one to convert credible sets of a calculated, simulated or approximated posterior into asymptotically consistent confidence sets, in full generality. This extends the main inferential implication of the Bernstein-von Mises theorem to non-parametric models without smoothness conditions. Proofs require the existence of a Bayesian type of test sequence and priors giving rise to local prior predictive distributions that satisfy a weakened form of Le Cam's contiguity with respect to the data distribution. Results are applied in a wide range 0f examples and counterexamples.

1 Introduction

In this paper (following [5]: *"Statisticians should readily use both Bayesian and frequentist ideas."*) we examine for which priors Bayesian asymptotic conclusions extend to conclusions valid in the frequentist sense: how Doob's prior-almost-sure consistency is strengthened to reach Schwartz's frequentist conclusion that the posterior is consistent, or how a test that is consistent prior-almost-surely becomes a test that is consistent in *all* points of the model, or how a Bayesian credible set can serve as a frequentist confidence set asymptotically.

The central property to enable frequentist interpretation of posterior asymptotics is defined as *remote contiguity* in section 3. It expresses a weakened form of Le Cam's contiguity, relating the true distribution of the data to localized prior predictive distributions. Where Schwartz's Kullback-Leibler neighbourhoods represent a choice for the localization appropriate when the

---

*email: B.Kleijn@uva.nl, web: https://staff.fnwi.uva.nl/b.j.k.kleijn

sample is *i.i.d.*, remote contiguity generalises the notion to include non-*i.i.d.* samples, priors that change with the sample size, weak consistency with the Dirichlet process prior, *etcetera*.

Although firstly aimed at enhancing insight into asymptotic relations by simplification and generalisation, this paper also has a significant methodological consequence: theorem 4.12 demonstrates that if the prior is such that remote contiguity applies, credible sets can be converted to asymptotically consistent confidence sets in full generality. So the asymptotic validity of credible sets as confidence sets in smooth parametric models [53] extends much further: in practice, the frequentist can simulate the posterior in any model, construct his preferred type of credible sets and 'enlarge' them to obtain asymptotic confidence sets, provided his prior induces remote contiguity. This extends the main inferential implication of the Bernstein-von Mises theorem to non-parametric models.

In the remainder of this section we discuss posterior consistency. In section 2 we concentrate on an inequality that relates testing to posterior concentration and indicates the relation with Le Cam's inequality. Section 3 introduces remote contiguity and the analogue of Le Cam's First Lemma. In section 4, frequentist theorems on the asymptotic behaviour of posterior distributions are proved, on posterior consistency, on posterior rates of convergence, on consistent testing and model selection with Bayes factors and on the conversion of credible sets to confidence sets. Section 5 formulates the conclusions.

Definitions, notation, conventions roughly follow those of [51] and are collected in appendix A with some other preliminaries. All applications, illustrations, examples and counterexamples have been collected in appendix B. Proofs are found in appendix C.

1.1 Posterior consistency and inconsistency

For a statistical procedure to be *consistent*, it must infer the truth with arbitrarily large accuracy and probability, if we gather enough data. For example, when using sequential data $X^n \sim P_{\theta_0,n}$ to estimate the value θ_0, a consistent estimator sequence θ_n converges to θ_0 in $P_{\theta_0,n}$-probability. For a posterior $\Pi(\cdot|X^n)$ to be consistent, it must concentrate mass arbitrarily close to one in any neighbourhood of θ_0 as $n \to \infty$ (see definition 4.1).

Consider a model $\mathscr{P}$ for *i.i.d.* data with single-observation distribution P_0. Give $\mathscr{P}$ a Polish topology with Borel prior Π so that the posterior is well-defined (see definition A.3). The first general consistency theorem for posteriors is due to Doob.

THEOREM 1.1 *(Doob (1949))*
For all $n \geq 1$, let $(X_1, X_2, \ldots, X_n) \in \mathscr{X}^n$ be $i.i.d. - P_0$, where P_0 lies in a model $\mathscr{P}$. Suppose $\mathscr{X}$ and $\mathscr{P}$ are Polish spaces. Assume that $P \mapsto P(A)$ is Borel measurable for every Borel set $A \subset \mathscr{X}$. Then for any Borel prior Π on $\mathscr{P}$ the posterior is consistent, for Π-almost-all P.

In parametric applications Doob's Π-null-set of potential inconsistency can be considered small

(for example, when the prior dominates Lebesgue measure). But in non-parametric context these null-sets can become very large (or not, see [54]): the first examples of unexpected posterior inconsistency are due to Schwartz [61], but it was Freedman [29] who made the point famous with a simple non-parametric counterexample (discussed in detail as example B.1). In [30] it was even shown that inconsistency is generic in a topological sense: the set of pairs (P_0, Π) for which the posterior is consistent is *meagre*: posteriors that only wander around, placing and re-placing mass aimlessly, are the rule rather than the exception. (For a discussion, see example B.2.)

These and subsequent examples of posterior inconsistency established a widespread conviction that Bayesian methods were wholly unfit for frequentist purposes, at least in non-parametric context. The only justifiable conclusion from Freedman's meagreness, however, is that a condition is missing: Doob's assertion may be all that a Bayesian requires, a frequentist demands strictly more, thus restricting the class of possible choices for his prior. Strangely, a condition representing this restriction had already been found when Freedman's meagreness result was published.

THEOREM 1.2 *(Schwartz (1965))*
For all $n \geq 1$, let $(X_1, X_2, \ldots, X_n) \in \mathcal{X}^n$ be i.i.d. $-P_0$, where P_0 lies in a model $\mathcal{P}$. Let U denote an open neighbourhood of P_0 in $\mathcal{P}$. If,

(i) there exist measurable $\phi_n : \mathcal{X}^n \to [0,1]$, such that,

$$P_0^n \phi_n = o(1), \quad \sup_{Q \in U^c} Q^n (1 - \phi_n) = o(1), \tag{1}$$

(ii) and Π is a Kullback-Leibler prior, i.e. for all $\delta > 0$,

$$\Pi\left(P \in \mathcal{P} \, : \, -P_0 \log \frac{dP}{dP_0} < \delta \right) > 0, \tag{2}$$

then $\Pi(U|X^n) \xrightarrow{P_0\text{-}a.s.} 1$.

Over the decades, examples of problematic posterior behaviour in non-parametric setting continued to captivate [20, 21, 17, 22, 23, 31, 32], while Schwartz's theorem received initially limited but steadily growing amounts of attention: subsequent frequentist theorems (*e.g.* by Barron [3], Barron-Schervish-Wasserman [4], Ghosal-Ghosh-van der Vaart [34], Shen-Wasserman [63], Walker [70] and Walker-Lijoi-Prünster [72], Kleijn-Zhao [46] and many others) have extended the applicability of theorem 1.2 but not its essence, condition (2) for the prior. The following example illustrates that Schwartz's condition cannot be the whole truth, though.

EXAMPLE 1.3 Consider $X_1, X_2, \ldots$ that are *i.i.d.-*P_0 with Lebesgue density $p_0 : \mathbb{R} \to \mathbb{R}$ supported on an interval of known width (say, 1) but unknown location. Parametrize in terms of a continuous density η on $[0,1]$ with $\eta(x) > 0$ for all $x \in [0,1]$ and a location $\theta \in \mathbb{R}$: $p_{\theta,\eta}(x) = \eta(x - \theta) 1_{[\theta, \theta+1]}(x)$. A moment's thought makes clear that if $\theta \neq \theta'$,

$$-P_{\theta,\eta} \log \frac{p_{\theta',\eta'}}{p_{\theta,\eta}} = \infty,$$

3

for all η, η'. Therefore Kullback-Leibler neighbourhoods do not have any extent in the θ-direction and *no prior is a Kullback-Leibler prior in this model.* Nonetheless the posterior is consistent (see examples B.14 and B.15). □

Similar counterexamples exist [46] for the type of prior that is proposed in the analyses of posterior rates of convergence in (Hellinger) metric setting [34, 63]. Although methods in [46] avoid this type of problem, the essential nature of condition (2) in *i.i.d.* setting becomes apparent there as well.

This raises the central question of this paper: is Schwartz's Kullback-Leibler condition perhaps a manifestation of a more general notion? The argument leads to other questions for which insightful answers have been elusive: why is Doob's theorem completely different from Schwartz's? The accepted explanation views the lack of congruence as an indistinct symptom of differing philosophies, but is this justified? Why does weak consistency in the full non-parametric model (*e.g.* with the Dirichlet process prior [28], or more modern variations [19]) reside in a corner of its own (with *tailfreeness* [30] as sufficient property of the prior), apparently unrelated to posterior consistency in either Doob's or Schwartz's views? Indeed, what would Schwartz's theorem look like without the assumption that the sample is *i.i.d.* (*e.g.* with data that form a Markov chain or realize some other stochastic process) or with growing parameter spaces and changing priors? And to extend the scope further, what can be said about hypothesis testing, classification, model selection, *etcetera*? Given that the Bernstein-von Mises theorem cannot be expected to hold in any generality outside parametric setting [17, 32], what relationship exists between credible sets and confidence sets? This paper aims to shed more light on these questions in a general sense, by providing a prior condition that enables strengthening Bayesian asymptotic conclusions to frequentist ones, illustrated with a variety of examples and counterexamples.

2   Posterior concentration and asymptotic tests

In this section, we consider a lemma that relates concentration of posterior mass in certain model subsets to the existence of test sequences that distinguish between those subsets. More precisely, it is shown that the expected posterior mass outside a model subset V with respect to the local prior predictive distribution over a model subset B, is upper bounded (roughly) by the testing power of *any* statistical test for the hypotheses B versus V: if a test sequence exists, the posterior will concentrate its mass appropriately.

2.1 Bayesian test sequences

Since the work of Schwartz [62], test sequences and posterior convergence have been linked intimately. Here we follow Schwartz and consider asymptotic testing; however, we define test sequences immediately in Bayesian context by involving priors from the outset.

DEFINITION 2.1 *Given priors* (Π_n), *measurable model subsets* $(B_n), (V_n) \subset \mathscr{G}$ *and* $a_n \downarrow 0$, *a sequence of* $\mathscr{B}_n$-*measurable maps* $\phi_n : \mathscr{X}_n \to [0,1]$ *is called a* Bayesian test sequence *for* B_n *versus* V_n (*under* Π_n) *of power* a_n, *if,*

$$\int_{B_n} P_{\theta,n} \phi_n \, d\Pi_n(\theta) + \int_{V_n} P_{\theta,n}(1 - \phi_n) \, d\Pi_n(\theta) = o(a_n). \tag{3}$$

We say that (ϕ_n) *is a* Bayesian test sequence *for* B_n *versus* V_n (*under* Π_n) *if (3) holds for some* $a_n \downarrow 0$.

Note that if we have sequences (C_n) and (W_n) such that $C_n \subset B_n$ and $W_n \subset V_n$ for all $n \geq 1$, then a Bayesian test sequence for (B_n) versus (V_n) of power a_n is a Bayesian test sequence for (C_n) versus (W_n) of power (at least) a_n.

LEMMA 2.2 *For any* $B, V \in \mathscr{G}$ *with* $\Pi(B) > 0$ *and any measurable* $\phi : \mathscr{X} \to [0,1]$,

$$\int P_\theta \Pi(V|X) \, d\Pi(\theta|B) \leq \int P_\theta \phi \, d\Pi(\theta|B) + \frac{1}{\Pi(B)} \int_V P_\theta(1 - \phi) \, d\Pi(\theta). \tag{4}$$

So the mere existence of a test sequence is enough to guarantee posterior concentration, a fact expressed in n-dependent form through the following proposition.

PROPOSITION 2.3 *Assume that for given priors* Π_n, *sequences* $(B_n), (V_n) \subset \mathscr{G}$ *and* $a_n, b_n \downarrow 0$ *such that* $a_n = o(b_n)$ *with* $\Pi_n(B_n) \geq b_n > 0$, *there exists a Bayesian test sequence for* B_n *versus* V_n *of power* a_n. *Then,*

$$P_n^{\Pi_n|B_n} \Pi(V_n|X^n) = o(a_n \, b_n^{-1}), \tag{5}$$

for all $n \geq 1$.

To see how this leads to posterior consistency, consider the following: if the model subsets $V_n = V$ are all equal to the complement of a neighbourhood U of P_0, and the B_n are chosen such that the expectations of the random variables $X^n \mapsto \Pi(V|X^n)$ under $P_n^{\Pi_n|B_n}$ 'dominate' their expectations under $P_{0,n}$ in a suitable way, sufficiency of prior mass b_n given testing power $a_n \downarrow 0$, is enough to assert that $P_{0,n} \Pi(V|X^n) \to 0$, so an arbitrarily large fraction of posterior mass is found in U with high probability for n large enough.

2.2 Existence of Bayesian test sequences

Lemma 2.2 and proposition 2.3 require the existence of test sequences of the Bayesian type. That question is unfamiliar, frequentists are used to test sequences for pointwise or uniform

testing. For example, an application of Hoeffding's inequality demonstrates that, weak neighbourhoods are uniformly testable (see proposition A.6). Another well-known example concerns testability of convex model subsets. Mostly the uniform test sequences in Schwartz's theorem are constructed using convex building blocks B and V separated in Hellinger distance (see proposition B.7 and subsequent remarks).

Requiring the existence of a Bayesian test sequence *c.f.* (3) is quite different. We shall illustrate this point in various ways below. First of all the existence of a Bayesian test sequence is linked directly to behaviour of the posterior itself.

THEOREM 2.4 *Let* $(\Theta, \mathscr{G}, \Pi)$ *be given. For any* $B, V \in \mathscr{G}$ *with* $\Pi(B) > 0, \Pi(V) > 0$, *the following are equivalent,*

(i) *there are* $\mathscr{B}_n$*-measurable* $\phi_n : \mathscr{X}_n \to [0, 1]$ *such that for* Π*-almost-all* $\theta \in B, \theta' \in V$,

$$P_{\theta,n}\phi_n \to 0, \quad P_{\theta',n}(1 - \phi_n) \to 0,$$

(ii) *there are* $\mathscr{B}_n$*-measurable* $\phi_n : \mathscr{X}_n \to [0, 1]$ *such that,*

$$\int_B P_{\theta,n}\phi_n \, d\Pi(\theta) + \int_V P_{\theta,n}(1 - \phi_n) \, d\Pi(\theta) \to 0,$$

(iii) *for* Π*-almost-all* $\theta \in B, \theta' \in V$,

$$\Pi(V|X^n) \xrightarrow{P_{\theta,n}} 0, \qquad \Pi(B|X^n) \xrightarrow{P_{\theta',n}} 0.$$

The interpretation of this theorem is gratifying to supporters of the likelihood principle and pure Bayesians: distinctions between model subsets are Bayesian testable, if and only if, they are picked up by the posterior asymptotically, if and only if, there exists a pointwise test for B versus V that is Π-almost-surely consistent.

For a second, more frequentist way to illustrate how basic the existence of a Bayesian test sequences is, consider a parameter space (Θ, d) which is a metric space with fixed Borel prior Π and d-consistent estimators $\hat{\theta}_n : \mathscr{X}_n \to \Theta$ for θ. Then for every $\theta_0 \in \Theta$ and $\epsilon > 0$, there exists a pointwise test sequence (and hence, by dominated convergence, also a Bayesian test sequence) for $B = \{\theta \in \Theta : d(\theta, \theta_0) < \frac{1}{2}\epsilon\}$ versus $V = \{\theta \in \Theta : d(\theta, \theta_0) > \epsilon\}$. This approach is followed in example B.19 on random walks, see the definition of the test following inequality (B.36).

A third perspective on the existence of Bayesian tests arises from Doob's argument. From our present perspective, we note that theorem 2.4 implies an alternative proof of Doob's consistency theorem through the following existence result on Bayesian test sequences. (Note: here and elsewhere in *i.i.d.* setting, the parameter space Θ is $\mathscr{P}$, θ is the single-observation distribution P and $\theta \mapsto P_{\theta,n}$ is $P \mapsto P^n$.)

PROPOSITION 2.5 *Consider a model* $\mathscr{P}$ *of single-observation distributions* P *for i.i.d. data* $(X_1, X_2, \ldots, X_n) \sim P^n$, $(n \geq 1)$. *Assume that* $\mathscr{P}$ *is a Polish space with Borel prior* Π. *For any Borel set* V *there is a Bayesian test sequence for* V *versus* $\mathscr{P} \setminus V$ *under* Π.

Doob's theorem is recovered when we let V be the complement of any open neighbourhood U of P_0. Comparing with conditions for the existence of uniform tests, Bayesian tests are quite abundant: whereas uniform testing relies on the minimax theorem (forcing convexity, compactness and continuity requirements into the picture), Bayesian tests exist quite generally (at least, for Polish parameters with *i.i.d.* data).

The fourth perspective on the existence of Bayesian tests concerns a direct way to construct a Bayesian test sequence of optimal power, based on the fact that we are really only testing barycentres against each other: let priors (Π_n) and $\mathscr{G}$-measurable model subsets B_n, V_n be given. For given tests (ϕ_n) and power sequence a_n, write (3) as follows:

$$\Pi_n(B_n) \, P_n^{\Pi_n|B_n} \phi_n(X^n) + \Pi_n(V_n) \, P_n^{\Pi_n|V_n} \phi_n(X^n) = o(a_n),$$

and note that what is required here, is a (weighted) test of $(P_n^{\Pi_n|B_n})$ versus $(P_n^{\Pi_n|V_n})$. The likelihood-ratio test (denote the density for $P_n^{\Pi_n|B_n}$ with respect to $\mu_n = P_n^{\Pi_n|B_n} + P_n^{\Pi_n|V_n}$ by $p_{B_n,n}$, and similar for $P_n^{\Pi_n|V_n}$),

$$\phi_n(X^n) = 1_{\{\Pi_n(V_n) \, p_{V_n,n}(X^n) > \Pi_n(B_n) \, p_{B_n,n}(X^n)\}},$$

is optimal and has power $\|\Pi_n(B_n) \, P_n^{\Pi_n|B_n} \wedge \Pi_n(V_n) \, P_n^{\Pi_n|B_n}\|$. This proves the following useful proposition that re-expresses power in terms of the relevant Hellinger transform (see, *e.g.* section 16.4 in [51], particularly, Remark 1).

PROPOSITION 2.6 *Let priors (Π_n) and measurable model subsets B_n, V_n be given. There exists a test sequence $\phi_n : \mathscr{X}_n \to [0,1]$ such that,*

$$\int_{B_n} P_{\theta,n} \phi_n \, d\Pi_n(\theta) + \int_{V_n} P_{\theta,n}(1 - \phi_n) \, d\Pi_n(\theta)$$
$$\leq \int \Big(\Pi_n(B_n) \, p_{B_n,n}(x)\Big)^{\alpha} \Big(\Pi_n(V_n) \, p_{V_n,n}(x)\Big)^{1-\alpha} d\mu_n(x), \tag{6}$$

for every $n \geq 1$ and any $0 \leq \alpha \leq 1$.

Proposition 2.6 generalises proposition 2.5 and makes Bayesian tests available with a (close-to-)sharp bound on the power under fully general conditions. For the connection with minimax tests, we note the following. If $\{P_{\theta,n} : \theta \in B_n\}$ and $\{P_{\theta,n} : \theta \in V_n\}$ are *convex* sets (and the Π_n are Radon measures, *e.g.* in Polish parameter spaces), then,

$$H\big(P_n^{\Pi_n|B_n}, P_n^{\Pi_n|V_n}\big) \geq \inf\{H(P_{\theta,n}, P_{\theta',n}) : \theta \in B_n, \theta' \in V_n\}.$$

Combination with (6) for $\alpha = 1/2$, implies that the minimax upper bound in *i.i.d.* cases, *c.f.* proposition B.7 remains valid:

$$\int_{B_n} P^n \phi_n \, d\Pi_n(P) + \int_{V_n} Q^n (1 - \phi_n) \, d\Pi_n(Q) \leq \sqrt{\Pi_n(B_n) \, \Pi_n(V_n)} \, e^{-n\epsilon_n^2}, \tag{7}$$

where $\epsilon_n = \inf\{H(P,Q) : P \in B_n, Q \in V_n\}$. Given $a_n \downarrow 0$, any Bayesian test ϕ_n that satisfies (3) *for all* probability measures Π_n on Θ, is a (weighted) minimax test for B_n versus V_n of power a_n.

Note that the above enhances the role that the prior plays in the frequentist discussion of the asymptotic behaviour of the posterior: the prior is not only important in requirements like (2), but can also be of influence in the testing condition: where testing power is relatively weak, prior mass should be scarce to compensate and where testing power is strong, prior mass should be plentiful. To make use of this, one typically imposes *upper* bounds on prior mass in certain hard-to-test subsets of the model (as opposed to *lower* bounds like (2)). See example B.19 on random-walk data. In the Hellinger-geometric view, the prior determines whether the local prior predictive distributions $P_n^{\Pi_n|B_n}$ and $P_n^{\Pi_n|V_n}$ lie close together or not in Hellinger distance, and thus to the *r.h.s.* of (6) for $\alpha = 1/2$. This phenomenon plays a role in example B.17 on the estimation of a sparse vector of normal means, where it explains why the *slab*-component of a *spike-and-slab prior* must have a tail that is heavy enough.

2.3 Le Cam's inequality

Referring to the argument following proposition 2.3, one way of guaranteeing that the expectations of $X^n \mapsto \Pi(V|X^n)$ under $P_n^{\Pi|B_n}$ approximate those under $P_{0,n}$, is to choose $B_n = \{\theta \in \Theta : \|P_{\theta,n} - P_{\theta_0,n}\| \leq \delta_n\}$, for some sequence $\delta_n \to 0$, because in that case, $|P_{0,n}\psi - P_n^{\Pi|B_n}\psi| \leq \|P_{0,n} - P_n^{\Pi|B_n}\| \leq \delta_n$, for any random variable $\psi : \mathscr{X}_n \to [0,1]$. Without fixing the definition of the sets B_n, one may use this step to specify inequality (4) further:

$$
\begin{aligned}
P_{0,n}\Pi(V_n|X) \leq & \left\|P_{0,n} - P_n^{\Pi|B_n}\right\| \\
& + \int P_{\theta,n}\phi_n \, d\Pi_n(\theta|B_n) + \frac{\Pi_n(V_n)}{\Pi_n(B_n)} \int P_{\theta,n}(1 - \phi_n) \, d\Pi_n(\theta|V_n),
\end{aligned}
\tag{8}
$$

for B_n and V_n such that $\Pi_n(B_n) > 0$ and $\Pi_n(V_n) > 0$. Le Cam's inequality (8) is used, for example, in the proof of the Bernstein-von Mises theorem, see lemma 2 in section 8.4 of [53]. A less successful application pertains to non-parametric posterior rates of convergence for *i.i.d.* data, in an unpublished paper [50]. Rates of convergence obtained in this way are suboptimal: Le Cam qualifies the first term on the right-hand side of (8) as a *"considerable nuisance"* and concludes that *"it is unclear at the time of this writing what general features, besides the metric structure, could be used to refine the results"*, (see [51], end of section 16.6). In [74], Le Cam relates the posterior question to dimensionality restrictions [49, 63, 34] and reiterates, *"And for Bayes risk, I know that just the metric structure does not catch everything, but I don't know what else to look at, except calculations."*

3   Remote contiguity

Le Cam's notion of contiguity describes an asymptotic version of absolute continuity, applicable to sequences of probability measures in a limiting sense [48]. In this section we weaken the

property of contiguity in a way that is suitable to promote Π-almost-everywhere Bayesian limits to frequentist limits that hold everywhere.

3.1 Definition and criteria for remote contiguity

The notion of 'domination' left undefined in the argument following proposition 2.3 is made rigorous here.

DEFINITION 3.1 *Given measurable spaces $(\mathscr{X}_n, \mathscr{B}_n)$, $n \geq 1$ with two sequences (P_n) and (Q_n) of probability measures and a sequence $\rho_n \downarrow 0$, we say that Q_n is ρ_n-remotely contiguous with respect to P_n, notation $Q_n \lhd \rho_n^{-1} P_n$, if,*

$$P_n \phi_n(X^n) = o(\rho_n) \quad \Rightarrow \quad Q_n \phi_n(X^n) = o(1), \tag{9}$$

for every sequence of $\mathscr{B}_n$-measurable $\phi_n : \mathscr{X}_n \to [0, 1]$.

Note that for a sequence (Q_n) that is a_n-remotely contiguous with respect to (P_n), there exists no test sequence that distinguishes between P_n and Q_n with power a_n. Note also that given two sequences (P_n) and (Q_n), contiguity $P_n \lhd Q_n$ is equivalent to remote contiguity $P_n \lhd a_n^{-1} Q_n$ for all $a_n \downarrow 0$. Given sequences $a_n, b_n \downarrow 0$ with $a_n = O(b_n)$, b_n-remote contiguity implies a_n-remote contiguity of (P_n) with respect to (Q_n).

EXAMPLE 3.2 Let $\mathscr{P}$ be a model for the distribution of a single observation in *i.i.d.* samples $X^n = (X_1, \ldots, X_n)$. Let P_0, P and $\epsilon > 0$ be such that $-P_0 \log(dP/dP_0) < \epsilon^2$. The law of large numbers implies that for large enough n,

$$\frac{dP^n}{dP_0^n}(X^n) \geq e^{-\frac{n}{2}\epsilon^2}, \tag{10}$$

with P_0^n-probability one. Consequently, for large enough n and for any $\mathscr{B}_n$-measurable sequence $\psi_n : \mathscr{X}_n \to [0, 1]$,

$$P^n \psi_n \geq e^{-\frac{1}{2}n\epsilon^2} P_0^n \psi_n. \tag{11}$$

Therefore, if $P^n \phi_n = o(\exp(-\frac{1}{2}n\epsilon^2))$ then $P_0^n \phi_n = o(1)$. Conclude that for every $\epsilon > 0$, the Kullback-Leibler neighbourhood $\{P : -P_0 \log(dP/dP_0) < \epsilon^2\}$ consists of model distributions for which the sequence (P_0^n) of product distributions are $\exp(-\frac{1}{2}n\epsilon^2)$-remotely contiguous with respect to (P^n). $\square$

Criteria for remote contiguity are given in the lemma below; note that, here, we give sufficient conditions, rather than necessary and sufficient, as in Le Cam's First Lemma. (For the definition of $(dP_n/dQ_n)^{-1}$, see appendix A, *notation and conventions*.)

LEMMA 3.3 *Given (P_n), (Q_n), $a_n \downarrow 0$, $Q_n \lhd a_n^{-1} P_n$, if any of the following hold:*

(i) for any $\mathscr{B}_n$-measurable $\phi_n : \mathscr{X}_n \to [0, 1]$, $a_n^{-1} \phi_n \xrightarrow{P_n} 0$ implies $\phi_n \xrightarrow{Q_n} 0$,

(ii) given $\epsilon > 0$, there is a $\delta > 0$ such that $Q_n(dP_n/dQ_n < \delta a_n) < \epsilon$, for large enough n,

(iii) there is a $b > 0$ such that $\liminf_n b\, a_n^{-1} P_n(dQ_n/dP_n > b\, a_n^{-1}) = 1$,

(iv) for any $\epsilon > 0$, there is a constant $c > 0$ such that $\|Q_n - Q_n \wedge c\, a_n^{-1} P_n\| < \epsilon$, for large enough n,

(v) under Q_n every subsequence of $(a_n(dP_n/dQ_n)^{-1})$ has a weakly convergent subsequence.

PROOF The proof of this lemma can be found in appendix C. It actually proves that (*(i)* or *(iv)*) implies remote contiguity; that (*(ii)* or *(iii)*) implies *(iv)* and that *(v)* is equivalent to *(ii)*. □

Contiguity and its remote variation are compared in the context of (parametric and non-parametric) regression in examples B.11 and B.12. We may specify the definition of remote contiguity slightly further.

DEFINITION 3.4 *Given measurable spaces $(\mathscr{X}_n, \mathscr{B}_n)$, $(n \geq 1)$ with two sequences (P_n) and (Q_n) of probability measures and sequences $\rho_n, \sigma_n > 0$, $\rho_n, \sigma_n \to 0$, we say that Q_n is ρ_n-to-σ_n remotely contiguous with respect to P_n, notation $\sigma_n^{-1} Q_n \lhd \rho_n^{-1} P_n$, if,*

$$P_n \phi_n(X^n) = o(\rho_n) \quad \Rightarrow \quad Q_n \phi_n(X^n) = o(\sigma_n),$$

for every sequence of $\mathscr{B}_n$-measurable $\phi_n : \mathscr{X}_n \to [0, 1]$.

Like definition 3.1, definition 3.4 allows for reformulation similar to lemma 3.3, *e.g.* if for some sequences ρ_n, σ_n like in definition 3.4,

$$\left\| Q_n - Q_n \wedge \sigma_n \rho_n^{-1} P_n \right\| = o(\sigma_n),$$

then $\sigma_n^{-1} Q_n \lhd \rho_n^{-1} P_n$. We leave the formulation of other sufficient conditions to the reader. Note that inequality (11) in example 3.2 implies that $b_n^{-1} P_0^n \lhd a_n^{-1} P^n$, for any $a_n \leq \exp(-n\alpha^2)$ with $\alpha^2 > \frac{1}{2}\epsilon^2$ and $b_n = \exp(-n(\alpha^2 - \frac{1}{2}\epsilon^2))$. It is noted that this implies that $\phi_n(X^n) \xrightarrow{Q_n\text{-a.s.}} 0$ for any $\phi_n : \mathscr{X}_n \to [0, 1]$ such that $P_n \phi_n(X^n) = o(\rho_n)$ (more generally, this holds whenever $\sum_n \sigma_n < \infty$, as a consequence of the first Borel-Cantelli lemma).

3.2 Remote contiguity for Bayesian limits

The relevant applications in the context of Bayesian limit theorems concern remote contiguity of the sequence of true distributions $P_{\theta_0, n}$ with respect to local prior predictive distributions $P_n^{\Pi_n | B_n}$, where the sets $B_n \subset \Theta$ are such that,

$$P_{\theta_0, n} \lhd a_n^{-1} P_n^{\Pi_n | B_n}, \tag{12}$$

for some rate $a_n \downarrow 0$.

In the case of *i.i.d.* data, Barron [3] introduces strong and weak notions of *merging* of $P_{\theta_0, n}$ with (non-local) prior predictive distributions P_n^Π. The weak version imposes condition *(ii)* of

lemma 3.3 for all exponential rates simultaneously. Strong merging (or *matching* [2]) coincides with Schwartz's almost-sure limit, while weak matching is viewed as a limit in probability.

By contrast, if we have a specific rate a_n in mind, the relevant mode of convergence is Prohorov's weak convergence: according to lemma 3.3-*(v)*, (12) holds if inverse likelihood ratios Z_n have a weak limit Z when re-scaled by a_n,

$$Z_n = (dP_n^{\Pi_n|B_n}/dP_{\theta_0,n})^{-1}(X^n), \quad a_n Z_n \xrightarrow{P_{\theta_0,n}\text{-w.}} Z.$$

To better understand the counterexamples of section B, notice the high sensitivity of this criterion to the existence of subsets of the sample spaces assigned probability zero under some model distributions, while the true probability is non-zero. More generally, remote contiguity is sensitive to subsets E_n assigned fast decreasing probabilities under local prior predictive distributions $P_n^{\Pi_n|B_n}(E_n)$, while the probabilities $P_{\theta_0,n}(E_n)$ remain high, which is what definition 3.1 expresses. The rate $a_n \downarrow 0$ helps to control the likelihood ratio (compare to the unscaled limits of likelihood ratios that play a central role in the theory of convergence of experiments [51]), conceivably enough to force uniform tightness in many non-parametric situations.

But condition (12) can also be written out, for example to the requirement that for some constant $\delta > 0$,

$$P_{\theta_0,n}\Big(\int \frac{dP_{\theta,n}}{dP_{\theta_0,n}}(X^n)\, d\Pi_n(\theta|B_n) < \delta\, a_n \Big) \to 0,$$

with the help of lemma 3.3-*(ii)*.

EXAMPLE 3.5 Consider again the model of example 1.3. In example B.14, it is shown that if the prior Π for $\theta \in \mathbb{R}$ has a continuous and strictly positive Lebesgue density and we choose $B_n = [\theta_0, \theta_0 + 1/n]$, then for every $\delta > 0$ and all $a_n \downarrow 0$,

$$P_{\theta_0}^n\left(\int \frac{dP_{\theta,n}}{dP_{\theta_0,n}}(X^n)\, d\Pi(\theta|B_n) < \delta\, a_n \right) \le P_{\theta_0}^n\big(n(X_{(1)} - \theta_0) < 2\delta\, a_n \big),$$

for large enough $n \ge 1$, and the *r.h.s.* goes to zero for any a_n because the random variables $n(X_{(1)} - \theta_0)$ have a non-degenerate, positive weak limit under $P_{\theta_0}^n$ as $n \to \infty$. Conclude that with these choices for Π and B_n, (12) holds, for any a_n. $\square$

The following proposition should be viewed in light of [52], which considers properties like contiguity, convergence of experiments and local asymptotic normality in situations of statistical information loss. In this case, we are interested in (remote) contiguity of the probability measures that arise as marginals for the data X^n when information concerning the (Bayesian random) parameter θ is unavailable.

PROPOSITION 3.6 *Let $\theta_0 \in \Theta$ and a prior $\Pi : \mathscr{G} \to [0,1]$ be given. Let B be a measurable subset of Θ such that $\Pi(B) > 0$. Assume that for some $a_n \downarrow 0$, the family,*

$$\left\{ a_n\Big(\frac{dP_{\theta,n}}{dP_{\theta_0,n}}\Big)^{-1}(X^n) : \theta \in B, n \ge 1 \right\},$$

is uniformly tight under $P_{\theta_0,n}$. Then $P_{\theta_0,n} \lhd a_n^{-1} P_n^{\Pi|B}$.

Other sufficient conditions from lemma 3.3 may replace the uniform tightness condition. When the prior Π and subset B are n-dependent, application of lemma 3.3 requires more. (See, for instance, example B.12 and lemma B.13, where local asymptotic normality is used to prove (12).)

To re-establish contact with the notion of merging, note the following. If remote contiguity of the type (12) can be achieved for a sequence of subsets (B_n), then it also holds for any sequence of sets (*e.g.* all equal to Θ, in Barron's case) that contain the B_n but at a rate that differs proportionally to the fraction of prior masses.

LEMMA 3.7 *For all $n \geq 1$, let $B_n \subset \Theta$ be such that $\Pi_n(B_n) > 0$ and C_n such that $B_n \subset C_n$ with $c_n = \Pi_n(B_n)/\Pi_n(C_n) \downarrow 0$, then,*

$$P_n^{\Pi_n|B_n} \lhd c_n^{-1} P_n^{\Pi_n|C_n}.$$

Also, if for some sequence (P_n), $P_n \lhd a_n^{-1} P_n^{\Pi_n|B_n}$ then $P_n \lhd a_n^{-1} c_n^{-1} P_n^{\Pi_n|C_n}$.

So when considering possible choices for the sequence (B_n), smaller choices lead to slower rates a_n, rendering (9) applicable to more sequences of test functions. This advantage is to be balanced against later requirements that $\Pi_n(B_n)$ may not decrease too fast.

4 Posterior concentration

In this section new frequentist theorems are formulated involving the convergence of posterior distributions. First we give a basic proof for posterior consistency assuming existence of suitable test sequences and remote contiguity of true distributions $(P_{\theta_0,n})$ with respect to local prior predictive distributions. Then it is not difficult to extend the proof to the case of posterior rates of convergence in metric topologies. With the same methodology it is possible to address questions in Bayesian hypothesis testing and model selection: if a Bayesian test to distinguish between two hypotheses exists and remote contiguity applies, frequentist consistency of the Bayes Factor can be guaranteed. We conclude with a theorem that uses remote contiguity to describe a general relation that exists between credible sets and confidence sets, provided the prior induces remotely-contiguous local prior predictive distributions.

4.1 Consistent posteriors

First, we consider posterior consistency generalising Schwartz's theorem to sequentially observed (non-*i.i.d.*) data, non-dominated models and priors or parameter spaces that may depend on the sample size. For an early but very complete overview of literature and developments in posterior consistency, see [33].

DEFINITION 4.1 *The posteriors* $\Pi(\,\cdot\,|X^n)$ *are* consistent *at* $\theta \in \Theta$ *if for every neighbourhood* U *of* θ,

$$\Pi(U|X^n) \xrightarrow{P_{\theta,n}} 1. \tag{13}$$

The posteriors are said to be consistent *if this holds for all* $\theta \in \Theta$. *We say that the posterior is* almost-surely consistent *if convergence occurs almost-surely with respect to some coupling for the sequence* $(P_{\theta_0,n})$.

Equivalently, posterior consistency can be characterized in terms of posterior expectations of bounded and continuous functions (see proposition B.5).

THEOREM 4.2 *Assume that for all* $n \geq 1$, *the data* $X^n \sim P_{\theta_0,n}$ *for some* $\theta_0 \in \Theta$. *Fix a prior* $\Pi : \mathscr{G} \to [0,1]$ *and assume that for given* $B, V \in \mathscr{G}$ *with* $\Pi(B) > 0$ *and* $a_n \downarrow 0$,

(i) there exist Bayesian tests ϕ_n *for* B *versus* V,

$$\int_B P_{\theta,n}\phi_n \, d\Pi(\theta) + \int_V P_{\theta',n}(1 - \phi_n) \, d\Pi(\theta') = o(a_n), \tag{14}$$

(ii) the sequence $P_{\theta_0,n}$ *satisfies* $P_{\theta_0,n} \lhd a_n^{-1} P_n^{\Pi|B}$.

Then $\Pi(V|X^n) \xrightarrow{P_{\theta_0,n}} 0$.

These conditions are to be interpreted as follows: theorem 2.4 lends condition *(i)* a distinctly Bayesian interpretation: it requires a Bayesian test to set V apart from B with testing power a_n. Lemma 2.2 translates this into the (still Bayesian) statement that the posteriors for V go to zero in $P_n^{\Pi|B}$-expectation. Condition *(ii)* is there to promote this Bayesian point to a frequentist one through (9). To present this from another perspective: condition *(ii)* ensures that the $P_n^{\Pi|B}$ cannot be tested versus $P_{\theta_0,n}$ at power a_n, so the posterior for V go to zero in $P_{\theta_0,n}$-expectation as well (otherwise a sequence $\phi_n(X^n) \propto \Pi(V|X^n)$ would constitute such a test).

To illustrate theorem 4.2 and its conditions Freedman's counterexamples are considered in detail in example B.4.

A proof of a theorem very close to Schwartz's theorem is now possible. Consider condition *(i)* of theorem 1.2: a well-known argument based on Hoeffding's inequality guarantees the existence of a uniform test sequence *of exponential power* whenever a uniform test sequence test sequence exists, so Schwartz equivalently assumes that there exists a $D > 0$ such that,

$$P_0^n \phi_n + \sup_{Q \in \mathscr{P}\setminus U} Q^n(1 - \phi_n) = o(e^{-nD}).$$

We vary slightly and assume the existence of a Bayesian test sequence of exponential power. In the following theorem, let $\mathscr{P}$ denote a Hausdorff space of single-observation distributions on $(\mathscr{X}, \mathscr{B})$ with Borel prior Π.

COROLLARY 4.3 *For all $n \geq 1$, let $(X_1, X_2, \ldots, X_n) \sim P_0^n$ for some $P_0 \in \mathscr{P}$. Let U denote an open neighbourhood of P_0 and define $K(\epsilon) = \{P \in \mathscr{P} : -P_0 \log(dP/dP_0) < \epsilon^2\}$. If,*

(i) there exist $\epsilon > 0$, $D > 0$ and a sequence of measurable $\psi_n : \mathscr{X}^n \to [0,1]$, such that,

$$\int_{K(\epsilon)} P^n \psi_n \, d\Pi(P) + \int_{\mathscr{P} \setminus U} Q^n (1 - \psi_n) \, d\Pi(Q) = o(e^{-nD}),$$

(ii) and $\Pi(K(\epsilon)) > 0$ for all $\epsilon > 0$,

then $\Pi(U|X^n) \xrightarrow{P_0\text{-}a.s.} 1$.

An instance of the application of corollary 4.3 is given in example B.10. Example B.23 demonstrates posterior consistency in total variation for *i.i.d.* data from a finite sample space, for priors of full support. Extending this, example B.24 concerns consistency of posteriors for priors that have Freedman's tailfreeness property [30], like the Dirichlet process prior. Also interesting in this respect is the Neyman-Scott paradox, a classic example of inconsistency for the ML estimator, discussed in Bayesian context in [5]: whether the posterior is (in)consistent depends on the prior. The Jeffreys prior follows the ML estimate while the reference prior avoids the Neyman-Scott inconsistency. Another question in a sequence model arises when we analyse FDR-like posterior consistency for a sequence vector that is assumed to be sparse (see example B.17).

4.2   Rates of posterior concentration

A significant extension to the theory on posterior convergence is formed by results concerning posterior convergence in metric spaces *at a rate*. Minimax rates of convergence for (estimators based on) posterior distributions were considered more or less simultaneously in Ghosal-Ghosh-van der Vaart [34] and Shen-Wasserman [63]. Both propose an extension of Schwartz's theorem to posterior rates of convergence [34, 63] and apply Barron's sieve idea with a well-known entropy argument [7, 8] to a shrinking sequence of Hellinger neighbourhoods and employs a more specific, rate-related version of the Kullback-Leibler condition (2) for the prior. Both appear to be inspired by contemporary results regarding Hellinger rates of convergence for sieve MLE's, as well as on Barron-Schervish-Wasserman [4], which concerns posterior consistency based on controlled bracketing entropy for a sieve, up to subsets of negligible prior mass, following ideas that were first laid down in [3]. It is remarked already in [4] that their main theorem is easily re-formulated as a rate-of-convergence theorem, with reference to [63]. More recently, Walker, Lijoi and Prünster [72] have added to these considerations with a theorem for Hellinger rates of posterior concentration in models that are separable for the Hellinger metric, with a central condition that calls for summability of square-roots of prior masses of covers of the model by Hellinger balls, based on analogous consistency results in Walker [70]. More recent is [46], which shows that alternatives for the priors of [34, 63] exist.

THEOREM 4.4 *Assume that for all $n \geq 1$, the data $X^n \sim P_{\theta_0,n}$ for some $\theta_0 \in \Theta$. Fix priors $\Pi_n : \mathscr{G} \to [0,1]$ and assume that for given $B_n, V_n \in \mathscr{G}$ with $\Pi_n(B_n) > 0$ and $a_n, b_n \downarrow 0$ such that $a_n = o(b_n)$,*

(i) there are Bayesian tests $\phi_n : \mathscr{X}_n \to [0,1]$ such that,

$$\int_{B_n} P_{\theta,n} \phi_n \, d\Pi_n(\theta) + \int_{V_n} P_{\theta,n}(1 - \phi_n) \, d\Pi_n(\theta) = o(a_n), \tag{15}$$

(ii) The prior mass of B_n is lower-bounded, $\Pi_n(B_n) \geq b_n$,

(iii) The sequence $P_{\theta_0,n}$ satisfies $P_{\theta_0,n} \lhd b_n a_n^{-1} P_n^{\Pi_n|B_n}$.

Then $\Pi(V_n|X^n) \xrightarrow{P_{\theta_0,n}} 0$. □

EXAMPLE 4.5 To apply theorem 4.4, consider again the situation of a uniform distribution with an unknown location, as in examples 1.3 and 3.5. Taking V_n equal to $\{\theta : \theta - \theta_0 > \epsilon_n\}$ $\{\theta : \theta_0 - \theta > \epsilon_n\}$ respectively, with $\epsilon_n = M_n/n$ for some $M_n \to \infty$, suitable test sequences are constructed in example B.15, and in combination with example 3.5, lead to the conclusion that with a prior Π for θ that has a continuous and strictly positive Lebesgue density, the posterior is consistent at (any ϵ_n slower than) rate $1/n$.

EXAMPLE 4.6 Let us briefly review the conditions of [4, 34, 63] in light of theorem 4.4: let $\epsilon_n \downarrow 0$ denote the Hellinger rate of convergence we have in mind, let $M > 1$ be some constant and define,

$$V_n = \{P \in \mathscr{P} : H(P, P_0) \geq M\epsilon_n\},$$
$$B_n = \{P \in \mathscr{P} : -P_0 \log dP/dP_0 < \epsilon_n^2, \, P_0 \log^2 dP/dP_0 < \epsilon_n^2\}.$$

Theorems for posterior convergence at a rate propose a sieve of submodels satisfying entropy conditions like those of [7, 8, 51] and a negligibility condition for prior mass outside the sieve [3], based on the minimax Hellinger rate of convergence $\epsilon_n \downarrow 0$. Together, they guarantee the existence of Bayesian tests for Hellinger balls of radius ϵ_n versus complements of Hellinger balls of radius $M\epsilon_n$ of power $\exp(-DM^2 n\epsilon_n^2)$ for some $D > 0$ (see example B.8). Note that B_n is contained in the Hellinger ball of radius ϵ_n around P_0, so (15) holds. New in [34, 63] is the condition for the priors Π_n,

$$\Pi_n(B_n) \geq e^{-Cn\epsilon_n^2}, \tag{16}$$

for some $C > 0$. With the help of lemmas B.16 and 3.3-*(ii)*, we conclude that,

$$P_0^n \lhd e^{cn\epsilon_n^2} P_n^{\Pi|B_n}, \tag{17}$$

for any $c > 1$. If we choose M such that $DM^2 - C > 1$, theorem 4.4 proves that $\Pi(V_n|X^n) \xrightarrow{P_0} 0$, *i.e.* the posterior is Hellinger consistent at rate ϵ_n.

Certain (simple, parametric) models do not allow the definition of priors that satisfy (16), and alternative less restrictive choices for the sets B_n are possible under mild conditions on the model [46].

4.3   Consistent hypothesis testing with Bayes factors

The Neyman-Pearson paradigm notwithstanding, hypothesis testing and classification concern the same fundamental statistical question, to find a procedure to choose one subset from a given partition of the parameter space as the most likely to contain the parameter value of the distribution that has generated the data observed. Asymptotically one wonders whether choices following such a procedure focus on the correct subset with probability growing to one.

From a somewhat shifted perspective, we argue as follows: no statistician can be certain of the validity of specifics in his model choice and therefore always runs the risk of biasing his analysis from the outset. Non-parametric approaches alleviate his concern but imply greater uncertainty within the model, leaving the statistician with the desire to select the correct (sub)model on the basis of the data before embarking upon the statistical analysis proper (for a recent overview, see [69]). The issue also makes an appearance in asymptotic context, where over-parametrized models leave room for inconsistency of estimators, requiring regularization [9, 10, 12].

Model selection describes all statistical methods that attempt to determine from the data which model to use. (Take for example sparse variable selection, where one projects out the majority of covariates prior to actual estimation, and the model-selection question is which projection is optimal.) Methods for model selection range from simple rules-of-thumb, to cross-validation and penalization of the likelihood function. Here we propose to conduct the frequentist analysis with the help of a posterior: when faced with a (dichotomous) model choice, we let the so-called Bayes factor formulate our preference. For an analysis of hypothesis testing that compares Bayesian and frequentist views, see [5]. An objective Bayesian perspective on model selection is provided in [73].

DEFINITION 4.7 *For all $n \geq 1$, let the model be parametrized by maps $\theta \mapsto P_{\theta,n}$ on a parameter space $(\Theta, \mathscr{G})$ with priors $\Pi_n : \mathscr{G} \to [0,1]$. Consider disjoint, measurable $B, V \subset \Theta$. For given $n \geq 1$, we say that the* Bayes factor for testing B versus V,

$$F_n = \frac{\Pi(B|X^n)}{\Pi(V|X^n)} \frac{\Pi_n(V)}{\Pi_n(B)},$$

is consistent for testing B versus V, if for all $\theta \in V$, $F_n \xrightarrow{P_{\theta,n}} 0$ and for all $\theta \in B$, $F_n^{-1} \xrightarrow{P_{\theta,n}} 0$.

Let us first consider this from a purely Bayesian perspective: for fixed prior Π and *i.i.d.* data, theorem 2.4 says that the posterior gives rise to consistent Bayes factors for B versus V in a Bayesian (that is, Π-almost-sure) way, iff a Bayesian test sequence for B versus V exists. If the parameter space Θ is Polish and the maps $\theta \mapsto P_\theta(A)$ are Borel measurable for all $A \in \mathscr{B}$, proposition 2.5 says that any Borel set V is Bayesian testable versus $\Theta \setminus V$, so in Polish models for *i.i.d.* data, model selection with Bayes factors is Π-almost-surely consistent for all Borel measurable $V \subset \Theta$.

The frequentist requires strictly more, however, so we employ remote contiguity again to bridge the gap with the Bayesian formulation.

THEOREM 4.8 *For all $n \geq 1$, let the model be parametrized by maps $\theta \mapsto P_{\theta,n}$ on a parameter space with $(\Theta, \mathscr{G})$ with priors $\Pi_n : \mathscr{G} \to [0,1]$. Consider disjoint, measurable $B, V \subset \Theta$ with $\Pi_n(B), \Pi_n(V) > 0$ such that,*

(i) There exist Bayesian tests for B versus V of power $a_n \downarrow 0$,

$$\int_B P^n \phi_n \, d\Pi_n(P) + \int_V Q^n (1 - \phi_n) \, d\Pi_n(Q) = o(a_n),$$

(ii) For every $\theta \in B$, $P_{\theta,n} \lhd a_n^{-1} P_n^{\Pi_n|B}$, and for every $\theta \in V$, $P_{\theta,n} \lhd a_n^{-1} P_n^{\Pi_n|V}$.

Then the Bayes factor for B versus V is consistent.

Note that the second condition of theorem 4.8 can be replaced by a local condition: if, for every $\theta \in B$, there exists a sequence $B_n(\theta) \subset B$ such that $\Pi_n(B_n(\theta)) \geq b_n$ and $P_{\theta,n} \lhd a_n^{-1} b_n P_n^{\Pi_n|B_n}$, then $P_{\theta,n} \lhd a_n^{-1} P_n^{\Pi_n|B}$ (as a consequence of lemma 3.7 with $C_n = B$).

In example B.19, we use theorem 4.8 to prove the consistency of the Bayes factor for a goodness-of-fit test for the equilibrium distribution of an stationary ergodic Markov chain, based on large-length random-walk data, with prior and posterior defined on the space of Markov transition matrices.

4.4   Confidence sets from credible sets

The Bernstein-von Mises theorem [53] asserts that the posterior for a smooth, finite-dimensional parameter converges in total variation to a normal distribution centred on an efficient estimate with the inverse Fisher information as its covariance, if the prior has full support. The methodological implication is that Bayesian credible sets derived from such a posterior can be reinterpreted as asymptotically efficient confidence sets. This parametric fact begs for the exploration of possible non-parametric extensions but Freedman discourages us [32] with counterexamples (see also [17]) and concludes that: *"The sad lesson for inference is this. If frequentist coverage probabilities are wanted in an infinite-dimensional problem, then frequentist coverage probabilities must be computed."*

In recent years, much effort has gone into calculations that address the question whether non-parametric credible sets can play the role of confidence sets nonetheless. The focus lies on well-controlled examples in which both model and prior are Gaussian so that the posterior is conjugate and analyse posterior expectation and variance to determine whether credible metric balls have asymptotic frequentist coverage (for examples, see Szabó, van der Vaart and van Zanten [68] and references therein). Below, we change the question slightly and do not seek to justify the use of credible sets as confidence sets; from the present perspective it appears

more natural to ask in which particular fashion a credible set is to be transformed in order to guarantee the transform is a confidence set, at least in the large-sample limit.

In previous subsections, we have applied remote contiguity after the concentration inequality to control the $P_{\theta_0,n}$-expectation of the posterior probability for the alternative V through its $P_n^{\Pi|B_n}$-expectation. In the discussion of the coverage of credible sets that follows, remote contiguity is applied to control the $P_{\theta_0,n}$-probability that θ_0 falls outside the prospective confidence set through its $P_n^{\Pi|B_n}$-probability. The theorem below then follows from an application of Bayes's rule (A.22). Credible levels provide the sequence a_n.

DEFINITION 4.9 *Let* $(\Theta, \mathscr{G})$ *with prior* Π, *denote the sequence of posteriors by* $\Pi(\cdot|\cdot) : \mathscr{G} \times \mathscr{X}_n \to [0,1]$. *Let* $\mathscr{D}$ *denote a collection of measurable subsets of* Θ. *A sequence of credible sets* (D_n) *of credible levels* $1 - a_n$ *(where* $0 \leq a_n \leq 1$, $a_n \downarrow 0$*) is a sequence of set-valued maps* $D_n : \mathscr{X}_n \to \mathscr{D}$ *such that* $\Pi(\Theta \setminus D_n(x)|x) = o(a_n)$ *for* $P_n^{\Pi_n}$-*almost-all* $x \in \mathscr{X}_n$.

DEFINITION 4.10 *For* $0 \leq a \leq 1$, *a set-valued map* $x \mapsto C(x)$ *defined on* $\mathscr{X}$ *such that, for all* $\theta \in \Theta$, $P_\theta(\theta \notin C(X)) \leq a$, *is called a confidence set of level* $1 - a$. *If the levels* $1 - a_n$ *of a sequence of confidence sets* $C_n(X^n)$ *go to 1 as* $n \to \infty$, *the* $C_n(X^n)$ *are said to be asymptotically consistent.*

DEFINITION 4.11 *Let* D *be a (credible) set in* Θ *and let* $B = \{B(\theta) : \theta \in \Theta\}$ *denote a collection of model subsets such that* $\theta \in B(\theta)$ *for all* $\theta \in \Theta$. *A model subset* C' *is said to be (a confidence set) associated with* D *under* B, *if for all* $\theta \in \Theta \setminus C'$, $B(\theta) \cap D = \varnothing$. *The intersection* C *of all* C' *like above equals* $\{\theta \in \Theta : B(\theta) \cap D \neq \varnothing\}$ *and is called the minimal (confidence) set associated with* D *under* B *(see Fig 1).*

Example B.25 makes this construction explicit in uniform spaces and specializes to metric context.

THEOREM 4.12 *Let* $\theta_0 \in \Theta$ *and* $0 \leq a_n \leq 1$, $b_n > 0$ *such that* $a_n = o(b_n)$ *be given. Choose priors* Π_n *and let* D_n *denote level-*$(1 - a_n)$ *credible sets. Furthermore, for all* $\theta \in \Theta$, *let* $B_n = \{B_n(\theta) \in \mathscr{G} : \theta \in \Theta\}$ *denote a sequence such that,*

(i) $\Pi_n(B_n(\theta_0)) \geq b_n$,

(ii) $P_{\theta_0,n} \lhd b_n a_n^{-1} \, P_n^{\Pi_n|B_n(\theta_0)}$.

Then any confidence sets C_n *associated with the credible sets* D_n *under* B_n *are asymptotically consistent, i.e. for all* $\theta_0 \in \Theta$,

$$P_{\theta_0,n}\big(\theta_0 \in C_n(X^n)\big) \to 1. \tag{18}$$

This refutes Freedman's lesson, showing that the asymptotic identification of credible sets and confidence sets in smooth parametric models (the main inferential implication of the Bernstein-von Mises theorem) generalises to the above form of asymptotic congruence in non-parametric
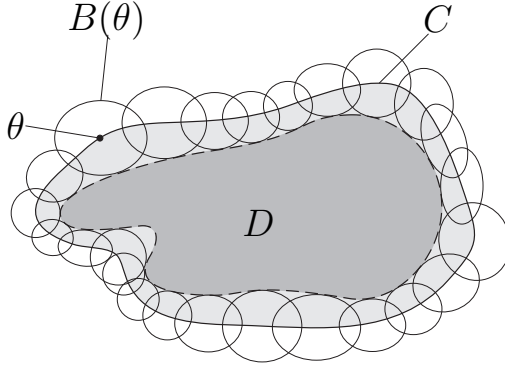
Figure 1: The relation between a credible set D and its associated (minimal) confidence set C under B in Venn diagrams: the extra points θ in the associated confidence set C not included in the credible set D are characterized by non-empty intersection $B(\theta) \cap D \neq \varnothing$.

models. The fact that this statement holds in full generality implies very practical ways to obtain confidence sets from posteriors, calculated, simulated or approximated. A second remark concerns the confidence levels of associated confidence sets. In order for the assertion of theorem 4.12 to be specific regarding the confidence level (rather than just resulting in asymptotic coverage), we re-write the last condition of theorem 4.12 as follows,

(ii') $c_n^{-1} P_{\theta_0,n} \lhd b_n a_n^{-1} P_n^{\Pi_n | B_n(\theta_0)}$,

so that the last step in the proof of theorem 4.12 is more specific; particularly, assertion (18) becomes,

$$P_{\theta_0,n}\big(\theta \in D_n(X^n) \big) = o(c_n),$$

i.e. the confidence level of the sets $D_n(X^n)$ is $1 - K c_n$ asymptotically (for some constant $K > 0$ and large enough n).

The following corollary that specializes to the *i.i.d.* situation is immediate (see example B.26). Let $\mathscr{P}$ denote a model of single-observation distributions, endowed with the Hellinger or total-variational topology.

COROLLARY 4.13 *For $n \geq 1$ assume that $(X_1, X_2, \ldots, X_n) \in \mathscr{X}^n \sim P_0^n$ for some $P_0 \in \mathscr{P}$. Let Π_n denote Borel priors on $\mathscr{P}$, with constant $C > 0$ and rate sequence $\epsilon_n \downarrow 0$ such that (16) is satisfied. Denote by D_n credible sets of level $1 - \exp(-C' n \epsilon_n^2)$, for some $C' > C$. Then the confidence sets C_n associated with D_n under radius-ϵ_n Hellinger-enlargement are asymptotically consistent.*

Note that in the above corollary,

$$\mathrm{diam}_H(C_n(X^n)) = \mathrm{diam}_H(D_n(X^n)) + 2\epsilon_n,$$

P_0^n-almost surely. If, in addition to the conditions in the above corollary, tests satisfying (15) with $a_n = \exp(-C'n\epsilon_n^2)$ exist, the posterior is consistent at rate ϵ_n and sets $D_n(X^n)$ have diameters decreasing as ϵ_n, *c.f.* theorem 4.4. In the case ϵ_n is the minimax rate of convergence for the problem, the confidence sets $C_n(X^n)$ attain rate-optimality [55]. Rate-adaptivity [42, 13, 68] is not possible like this because a definite, non-data-dependent choice for the B_n is required.

5   Conclusions

We list and discuss the main conclusions of this paper below.

Frequentist validity of Bayesian limits
There exists a systematic way of taking Bayesian limits into frequentist ones, if priors satisfy an extra condition relating true data distributions to localized prior predictive distributions. This extra condition generalises Schwartz's Kullback-Leibler condition and amounts to a weakened form of contiguity, termed *remote contiguity.*

For example regarding consistency with *i.i.d.* data, Doob shows that a Bayesian form of posterior consistency holds without any real conditions on the model. To the frequentist, 'holes' of potential inconsistency remain, in null-sets of the prior. Remote contiguity 'fills the holes' and elevates the Bayesian form of consistency to the frequentist one. Similarly, prior-almost-surely consistent tests are promoted to frequentist consistent tests and Bayesian credible sets are converted to frequentist confidence sets.

The nature of Bayesian test sequences
The existence of a Bayesian test sequence is equivalent to consistent posterior convergence in the Bayesian, prior-almost-sure sense. In theorems above, a Bayesian test sequence thus represents the Bayesian limit for which we seek frequentist validity through remote contiguity. Bayesian test sequences are more abundant than the more familiar uniform test sequences. Aside from prior mass requirements arising from remote contiguity, *the prior should assign little weight where testing power is weak and much where testing power is strong*, ideally.

Example B.19 illustrates the influence of the prior when constructing a test sequence. Aside from the familiar lower bounds for prior mass that arise from remote contiguity, existence of Bayesian tests also poses upper bounds for prior mass.

Systematic analysis of complex models and datasets
Although many examples have been studied on a case-by-case basis in the literature,

the systematic analysis of limiting properties of posteriors in cases where the data is dependent, or where the model, the parameter space and/or the prior are sample-size dependent, requires generalisation of Schwartz's theorem and its variations, which the formalism presented here provides.

To elaborate, given the growing interest in the analysis of dependent datasets gathered from networks (*e.g.* by *webcrawlers* that random walk linked webpages), or from time-series/stochastic processes (*e.g.* financial data of the high-frequency type), or in the form of high-dimensional or even functional data (biological, financial, medical and meteorological fields provide many examples), the development of new Bayesian methods involving such aspects benefits from a simple, insightful, systematic perspective to guide the search for suitable priors in concrete examples.

To illustrate the last point, let us consider consistent community detection in stochastic block models [64, 6]. Bayesian methods have been developed for consistent selection of the number of communities [41], for community detection with a controlled error-rate with a growing number of communities [16] and for consistent community detection using empirical priors [67]. A moment's thought on the discrete nature of the community assignment vector suggests a sequence of uniform priors, for which remote contiguity (of $B_n = \{P_{0,n}\}$) is guaranteed (at any rate) and prior mass lower bounded by $b_n = K_n! K_n^{-n}$ (where K_n is the number of communities at 'sample size' n). It would be interesting to see under which conditions a Bayesian test sequence of power $a_n = o(b_n)$ can be devised that tests the true assignment vector versus all alternatives (in the sparse regime [18, 1, 58]). Rather than apply a Chernoff bound like in [16], one would probably have to start from the probabilistic [58] or information-theoretic [1] analyses of respective algorithmic solutions in the (very closely related) planted bi-section model. If a suitably powerful test can be shown to exist, theorem 4.4 proves frequentist consistency of the posterior.

> *Methodology for uncertainty quantification*
> Use of a prior that induces remote contiguity allows one to convert credible sets of a calculated, simulated or approximated posterior into asymptotically consistent confidence sets, in full generality. This extends the main inferential implication of the Bernstein-von Mises theorem to non-parametric models without smoothness conditions.

The latter conclusion forms the most important and practically useful aspect of this paper.

A Definitions and conventions

Because we take the perspective of a frequentist using Bayesian methods, we are obliged to demonstrate that Bayesian definitions continue to make sense under the assumptions that the data X is distributed according to a true, underlying P_0.

REMARK A.1 We assume given for every $n \geq 1$, a measurable (sample) space $(\mathscr{X}_n, \mathscr{B}_n)$ and random sample $X^n \in \mathscr{X}_n$, with a model $\mathscr{P}_n$ of probability distributions $P_n : \mathscr{B}_n \to [0,1]$. It is also assumed that there exists an n-independent parameter space Θ with a Hausdorff, completely regular topology $\mathscr{T}$ and associated Borel σ-algebra $\mathscr{G}$, and, for every $n \geq 1$, a bijective model parametrization $\Theta \to \mathscr{P}_n : \theta \mapsto P_{\theta,n}$ such that for every $n \geq 1$ and every $A \in \mathscr{B}_n$, the map $\Theta \to [0,1] : \theta \mapsto P_{\theta,n}(A)$ is measurable. Any prior Π on Θ is assumed to be a Borel probability measure $\Pi : \mathscr{G} \to [0,1]$ and can vary with the sample-size n. (Note: in $i.i.d.$ setting, the parameter space Θ is $\mathscr{P}_1$, θ is the single-observation distribution P and $\theta \mapsto P_{\theta,n}$ is $P \mapsto P^n$.) As frequentists, we assume that there exists a 'true, underlying distribution for the data; in this case, that means that for every $n \geq 1$, there exists a distribution $P_{0,n}$ from which the n-th sample X^n is drawn. $\qquad\square$

Often one assumes that the model is *well-specified*: that there exists a $\theta_0 \in \Theta$ such that $P_{0,n} = P_{\theta_0,n}$ for all $n \geq 1$. We think of Θ as a topological space because we want to discuss estimation as a procedure of sequential, stochastic approximation of and convergence to such a 'true parameter value θ_0. In theorem 2.4 and definition 4.1 we assume, in addition, that the observations X^n are *coupled*, *i.e.* there exists a probability space $(\Omega, \mathscr{F}, P_0)$ and random variables $X^n : \Omega \to \mathscr{X}_n$ such that $P_0((X^n)^{-1}(A)) = P_{0,n}(X^n \in A)$ for all $n \geq 1$ and $A \in \mathscr{B}_n$.

DEFINITION A.2 *Given $n, m \geq 1$ and a prior probability measure $\Pi_n : \mathscr{G} \to [0,1]$, define the n-th prior predictive distribution on $\mathscr{X}_m$ as follows:*

$$P_m^{\Pi_n}(A) = \int_\Theta P_{\theta,m}(A)\, d\Pi_n(\theta), \tag{A.19}$$

for all $A \in \mathscr{B}_m$. If the prior is replaced by the posterior, the above defines the n-th posterior predictive distribution on $\mathscr{X}_m$,

$$P_m^{\Pi_n|X^n}(A) = \int_\Theta P_{\theta,m}(A)\, d\Pi(\theta|X^n), \tag{A.20}$$

for all $A \in \mathscr{B}_m$. For any $B_n \in \mathscr{G}$ with $\Pi_n(B_n) > 0$, define also the n-th local prior predictive distribution on $\mathscr{X}_m$,

$$P_m^{\Pi_n|B_n}(A) = \frac{1}{\Pi_n(B_n)} \int_{B_n} P_{\theta,m}(A)\, d\Pi_n(\theta), \tag{A.21}$$

as the predictive distribution on $\mathscr{X}_m$ that results from the prior Π_n when conditioned on B_n. If m is not mentioned explicitly, it is assumed equal to n.

The prior predictive distribution $P_n^{\Pi_n}$ is the marginal distribution for X^n in the Bayesian perspective that considers parameter and sample jointly $(\theta, X^n) \in \Theta \times \mathscr{X}_n$ as the random quantity of interest.

DEFINITION A.3 *Given $n \geq 1$, a (version of) the posterior is any map $\Pi(\,\cdot\,|X^n = \,\cdot\,) : \mathscr{G} \times \mathscr{X}_n \to [0,1]$ such that,*

(i) for $B \in \mathscr{G}$, the map $\mathscr{X}_n \to [0,1] : x^n \mapsto \Pi(B|X^n = x^n)$ is $\mathscr{B}_n$-measurable,

(ii) for all $A \in \mathscr{B}_n$ and $V \in \mathscr{G}$,

$$\int_A \Pi(V|X^n) \, dP_n^{\Pi_n} = \int_V P_{\theta,n}(A) \, d\Pi_n(\theta). \tag{A.22}$$

Bayes's Rule is expressed through equality (A.22) and is sometimes referred to as a 'disintegration' (of the joint distribution of (θ, X^n)). If the posterior is a Markov kernel, it is a $P_n^{\Pi_n}$-almost-surely well-defined probability measure on $(\Theta, \mathscr{G})$. But it does not follow from the definition above that a version of the posterior actually exists as a regular conditional probability measure. Under mild extra conditions, regularity of the posterior can be guaranteed: for example, if sample space and parameter space are Polish, the posterior is regular; if the model $\mathscr{P}_n$ is dominated (denote the density of $P_{\theta,n}$ by $p_{\theta,n}$), the fraction of integrated likelihoods,

$$\Pi(V|X^n) = \int_V p_{\theta,n}(X^n) \, d\Pi_n(\theta) \, \Big/ \int_\Theta p_{\theta,n}(X^n) \, d\Pi_n(\theta), \tag{A.23}$$

for $V \in \mathscr{G}$, $n \geq 1$ defines a regular version of the posterior distribution. (Note also that there is no room in definition (A.22) for X^n-dependence of the prior, so 'empirical Bayes' methods must be based on data Y^n independent of X^n, *i.e.* sample-splitting.)

REMARK A.4 As a consequence of the frequentist assumption that $X^n \sim P_{0,n}$ for all $n \geq 1$, the $P_n^{\Pi_n}$-almost-sure definition (A.22) of the posterior $\Pi(V|X^n)$ does not make sense automatically [29, 46]: null-sets of $P_n^{\Pi_n}$ on which the definition of $\Pi(\cdot|X^n)$ is ill-determined, may not be null-sets of $P_{0,n}$. To prevent this, we impose the domination condition,

$$P_{0,n} \ll P_n^{\Pi_n}, \tag{A.24}$$

for every $n \geq 1$. $\qquad\square$

To understand the reason for (A.24) in a perhaps more familiar way, consider a dominated model and assume that for certain n, (A.24) is *not* satisfied. Then, using (A.19), we find,

$$P_{0,n}\Big(\int p_{\theta,n}(X^n) \, d\Pi_n(\theta) = 0\Big) > 0,$$

so the denominator in (A.23) evaluates to zero with non-zero $P_{0,n}$-probability.

To get an idea of sufficient conditions for (A.24), it is noted in [46] that in the case of *i.i.d.* data where $P_{0,n} = P_0^n$ for some marginal distribution P_0, $P_0^n \ll P_n^\Pi$ for all $n \geq 1$, if P_0 lies in the Hellinger- or Kullback-Leibler-support of the prior Π. For the generalisation to the present setting we are more precise and weaken the topology appropriately.

DEFINITION A.5 *For all $n \geq 1$, let F_n denote the class of all bounded, $\mathscr{B}_n$-measurable $f : \mathscr{X}_n \to \mathbb{R}$. The topology $\mathscr{T}_n$ is the initial topology on $\mathscr{P}_n$ for the functions $\{P \mapsto Pf : f \in F_n\}$.* $\square$

Finite intersections of sets $U_{f,\epsilon} = \{(P, Q) \in \mathscr{P}_n^2 : |(P - Q)f| < \epsilon\}$ ($f \in \mathscr{F}_n$, $0 \leq f \leq 1$ and $\epsilon > 0$), form a fundamental system of entourages for a uniformity $\mathscr{U}_n$ on $\mathscr{P}_n$. A fundamental

system of neighbourhoods for the associated topology $\mathscr{T}_n$ on $\mathscr{P}$ is formed by finite intersections of sets of the form,

$$W_{P,f,\epsilon} = \{Q \in \mathscr{P}_n : |(P-Q)f| < \epsilon\},$$

with $P \in \mathscr{P}_n$, $f \in \mathscr{F}_n$, $0 \le f \le 1$ and $\epsilon > 0$.

If we model single-observation distributions $P \in \mathscr{P}$ for an *i.i.d.* sample, the topology $\mathscr{T}_n$ on $\mathscr{P}_n = \mathscr{P}^n$ induces a topology on $\mathscr{P}$ (which we also denote by $\mathscr{T}_n$) for each $n \ge 1$. The union $\mathscr{T}_\infty = \cup_n \mathscr{T}_n$ is an inverse-limit topology that allows formulation of conditions for the existence of consistent estimates that are not only sufficient, but also necessary [47], offering a precise perspective on what is estimable *and what is not* in *i.i.d.* context. The associated strong topology is that generated by total variation (or, equivalently, the Hellinger metric).

For more on these topologies, the reader is referred to Strasser (1985) [65] and to Le Cam (1986) [51]. We note explicitly the following fact, which is a direct consequence of Hoeffding's inequality.

PROPOSITION A.6 *(Uniform $\mathscr{T}_n$-tests)*
Consider a model $\mathscr{P}$ of single-observation distributions P for i.i.d. samples $(X_1, X_2, \ldots, X_n) \sim P^n$, $(n \ge 1)$. Let $m \ge 1$, $\epsilon > 0$, $P_0 \in \mathscr{P}$ and a measurable $f : \mathscr{X}^m \to [0,1]$ be given. Define $B = \big\{ P \in \mathscr{P} : |(P^m - P_0^m)f| < \epsilon \big\}$, and $V = \big\{ P \in \mathscr{P} : |(P^m - P_0^m)f| \ge 2\epsilon \big\}$. There exist a uniform test sequence (ϕ_n) such that,

$$\sup_{P \in B} P^n \phi_n \le e^{-nD}, \quad \sup_{Q \in V} Q^n(1 - \phi_n) \le e^{-nD},$$

for some $D > 0$.

PROOF The proof is an application of Hoeffding's inequality for the sum $\sum_{i=1}^n f(X_i)$ and is left to the reader. $\square$

The topologies $\mathscr{T}_n$ also play a role for condition (A.24).

PROPOSITION A.7 *Let (Π_n) be Borel priors on the Hausdorff uniform spaces $(\mathscr{P}_n, \mathscr{T}_n)$. For any $n \ge 1$, if $P_{0,n}$ lies in the $\mathscr{T}_n$-support of Π_n, then $P_{0,n} \ll P_n^{\Pi_n}$.*

PROOF Let $n \ge 1$ be given. For any $A \in \mathscr{B}_n$ and any $U' \subset \Theta$ such that $\Pi_n(U') > 0$,

$$P_{0,n}(A) \le \int P_{\theta,n}(A)\, d\Pi_n(\theta|U') + \sup_{\theta \in U'} |P_{\theta,n}(A) - P_{0,n}(A)|.$$

Let $A \in \mathscr{B}_n$ be a null-set of $P_n^{\Pi_n}$; since $\Pi_n(U') > 0$, $\int P_{\theta,n}(A)\, d\Pi_n(\theta|U') = 0$. For some $\epsilon > 0$, take U' equal to the $\mathscr{T}_n$-basis element $\{\theta \in \Theta : |P_{\theta,n}(A) - P_{\theta_0,n}(A)| < \epsilon\}$ to conclude that $P_{\theta_0,n}(A) < \epsilon$ for all $\epsilon > 0$. $\square$

In many situations, priors are Borel for the Hellinger topology, so it is useful to observe that the Hellinger support of Π_n in $\mathscr{P}_n$ is always contained in the $\mathscr{T}_n$-support.

Notation and conventions

l.h.s. and *r.h.s.* refer to left- and right-hand sides respectively. For given probability measures P, Q on a measurable space $(\Omega, \mathscr{F})$, we define the Radon-Nikodym derivative $dP/dQ : \Omega \to [0, \infty)$, P-almost-surely, referring *only* to the Q-dominated component of P, following [51]. We also *define* $(dP/dQ)^{-1} : \Omega \to (0, \infty] : \omega \mapsto 1/(dP/dQ(\omega))$, Q-almost-surely. Given a σ-finite measure μ that dominates both P and Q (*e.g.* $\mu = P + Q$), denote $dP/d\mu = q$ and $dQ/d\mu = p$. Then the measurable map $p/q \, 1\{q > 0\} : \Omega \to [0, \infty)$ is a μ-almost-everywhere version of dP/dQ, and $q/p \, 1\{q > 0\} : \Omega \to [0, \infty]$ of $(dP/dQ)^{-1}$. Define total-variational and Hellinger distances by $\|P - Q\| = \sup_A |P(A) - Q(A)|$ and $H(P, Q)^2 = 1/2 \int (p^{1/2} - q^{1/2})^2 \, d\mu$, respectively. Given random variables $Z_n \sim P_n$, weak convergence to a random variable Z is denoted by $Z_n \xrightarrow{P_n\text{-w.}} Z$, convergence in probability by $Z_n \xrightarrow{P_n} Z$ and almost-sure convergence (with coupling P^∞) by $Z_n \xrightarrow{P^\infty\text{-a.s.}} Z$. The integral of a real-valued, integrable random variable X with respect to a probability measure P is denoted PX, while integrals over the model with respect to priors and posteriors are always written out in Leibniz's notation. For any subset B of a topological space, $\bar{B}$ denotes the closure, $\mathring{B}$ the interior and ∂B the boundary. Given $\epsilon > 0$ and a metric space (Θ, d), the covering number $N(\epsilon, \Theta, d) \in \mathbb{N} \cup \{\infty\}$ is the minimal cardinal of a cover of Θ by d-balls of radius ϵ. Given real-valued random variables $X_1, \ldots, X_n$, the first order statistic is $X_{(1)} = \min_{1 \leq i \leq n} X_i$. The Hellinger diameter of a model subset C is denoted $\operatorname{diam}_H(C)$ and the Euclidean norm of a vector $\theta \in \mathbb{R}^n$ is denoted $\|\theta\|_{2,n}$.

B   Applications and examples

In this section of the appendix examples and applications are collected.

B.1   Inconsistent posteriors

Calculations that demonstrate instances of posterior inconsistency are many (for (a far-from-exhaustive list of) examples, see [20, 21, 17, 22, 23, 31, 32]). In this subsection, we discuss early examples of posterior inconsistency that illustrate the potential for problems clearly and without distracting technicalities.

EXAMPLE B.1 *(Freedman (1963) [29])*
Consider a sample $X_1, X_2, \ldots$ of random positive integers. Denote the space of all probability distributions on $\mathbb{N}$ by Λ and assume that the sample is *i.i.d.-*P_0, for some $P_0 \in \Lambda$. For any $P \in \Lambda$, write $p(i) = P(\{X = i\})$ for all $i \geq 1$. The total-variational and weak topologies on Λ are equivalent (defined, $P \to Q$ if $p(i) \to q(i)$ for all $i \geq 1$). Let $Q \in \Lambda \setminus \{P_0\}$ be given. To arrive at a prior with P_0 in its support, leading to a posterior that concentrates on Q, we consider sequences (P_m) and (Q_n) such that $Q_m \to Q$ and $P_m \to P_0$ as $m \to \infty$. The prior Π

places masses $\alpha_m > 0$ at P_m and $\beta_m > 0$ at Q_m ($m \geq 1$), so that P_0 lies in the support of Π. A careful construction of the distributions Q_m that involves P_0, guarantees that the posterior satisfies,

$$\frac{\Pi(\{Q_m\}|X^n)}{\Pi(\{Q_{m+1}\}|X^n)} \xrightarrow{P_0\text{-a.s.}} 0,$$

that is, posterior mass is shifted further out into the tail as n grows to infinity, forcing all posterior mass that resides in $\{Q_m : m \geq 1\}$ into arbitrarily small neighbourhoods of Q. In a second step, the distributions P_m and prior weights α_m are chosen such that the likelihood at P_m grows large for high values of m and small for lower values as n increases, so that the posterior mass in $\{P_m : m \geq 1\}$ also accumulates in the tail. However, the prior weights α_m may be chosen to decrease very fast with m, in such a way that,

$$\frac{\Pi(\{P_m : m \geq 1\}|X^n)}{\Pi(\{Q_m : m \geq 1\}|X^n)} \xrightarrow{P_0\text{-a.s.}} 0,$$

thus forcing all posterior mass into $\{Q_m : m \geq 1\}$ as n grows. Combination of the previous two displays leads to the conclusion that for every neighbourhood U_Q of Q,

$$\Pi(U_Q|X^n) \xrightarrow{P_0\text{-a.s.}} 1,$$

so the posterior is inconsistent. Other choices of the weights α_m that place more prior mass in the tail *do* lead to consistent posterior distributions. $\square$

Some objected to Freedman's counterexample, because knowledge of P_0 is required to construct the prior that causes inconsistency. So it was possible to argue that Freedman's counterexample amounted to nothing more than a demonstration that unfortunate circumstances could be created, probably not a fact of great concern in any generic sense. To strengthen Freedman's point one would need to construct a prior of full support without explicit knowledge of P_0.

EXAMPLE B.2 *(Freedman (1965) [30])*
In the setting of example B.1, denote the space of all distributions on Λ by $\pi(\Lambda)$. Note that since Λ is Polish, so is $\pi(\Lambda)$ and so is the product $\Lambda \times \pi(\Lambda)$.

THEOREM B.3 *(Freedman (1965) [30])*
Let $X_1, X_2, \ldots$ form an sample of i.i.d.-P_0 random integers, let Λ denote the space of all distributions on $\mathbb{N}$ and let $\pi(\Lambda)$ denote the space of all Borel probability measures on Λ, both in Prohorov's weak topology. The set of pairs $(P_0, \Pi) \in \Lambda \times \pi(\Lambda)$ such that for all open $U \subset \Lambda$,

$$\limsup_{n \to \infty} P_0^n \Pi(U|X^n) = 1,$$

is residual.

And so, the set of pairs $(P_0, \Pi) \in \Lambda \times \pi(\Lambda)$ for which the limiting behaviour of the posterior is acceptable to the frequentist, is *meagre* in $\Lambda \times \pi(\Lambda)$. The proof relies on the following construction: for $k \geq 1$, define Λ_k to be the subset of all probability distributions P on $\mathbb{N}$ such

that $P(X = k) = 0$. Also define Λ_0 as the union of all Λ_k, $(k \geq 1)$. Pick $Q \in \Lambda \setminus \Lambda_0$. We assume that $P_0 \in \Lambda \setminus \Lambda_0$ and $P_0 \neq Q$. Place a prior Π_0 on Λ_0 and choose $\Pi = \frac{1}{2}\Pi_0 + \frac{1}{2}\delta_Q$. Because Λ_0 is dense in Λ, priors of this type have full support in Λ. But P_0 has full support in $\mathbb{N}$ so for every $k \in \mathbb{N}$, $P_0^\infty(\exists_{m \geq 1} : X_m = k) = 1$: note that if we observe $X_m = k$, the likelihood equals zero on Λ_k so that $\Pi(\Lambda_k | X^n) = 0$ for all $n \geq m$, P_0^∞-almost-surely. Freedman shows this eliminates all of Λ_0 asymptotically, if Π_0 is chosen in a suitable way, forcing all posterior mass onto the point $\{Q\}$. (See also, Le Cam (1986) [51], section 17.7). $\qquad\square$

The question remains how Freedman's inconsistent posteriors relate to the work presented here. Since test sequences of exponential power exist to separate complements of weak neighbourhoods, *c.f.* proposition A.6, Freedman's inconsistencies must violate the requirement of remote contiguity in theorem 4.2.

EXAMPLE B.4 As noted already, Λ is a Polish space; in particular Λ is metric and second countable, so the subspace $\Lambda \setminus \Lambda_0$ contains a countable dense subset D. For $Q \in D$, let V be the set of all prior probability measures on Λ with finite support, of which one point is Q and the remaining points lie in Λ_0. The proof of the theorem in [30] that asserts that the set of consistent pairs (P_0, Π) is of the first category in $\Lambda \times \pi(\Lambda)$ departs from the observation that if P_0 lies in $\Lambda \setminus \Lambda_0$ and we use a prior from V, then,

$$\Pi(\{Q\}|X^n) \xrightarrow{P_0\text{-a.s.}} 1,$$

(in fact, as is shown below, with P_0^∞-probability one there exists an $N \geq 1$ such that $\Pi(\{Q\}|X^n) = 1$ for all $n \geq N$). The proof continues to assert that V lies dense in $\pi(\Lambda)$, and, through sequences of continuous extensions involving D, that posterior inconsistency for elements of V implies posterior inconsistency for all Π in $\pi(\Lambda)$ with the possible exception of a set of the first category.

From the present perspective it is interesting to view the inconsistency of elements of V in light of the conditions of theorem 4.2. Define, for some bounded $f : \mathbb{N} \to \mathbb{R}$ and $\epsilon > 0$, two subsets of Λ,

$$B = \{P : |Pf - P_0f| < \tfrac{1}{2}\epsilon\}, \quad V = \{P : |Pf - P_0f| \geq \epsilon\}.$$

Proposition A.6 asserts the existence of a uniform test sequence for B versus V of exponential power. With regard to remote contiguity, for an element Π of V with support of order $M + 1$, write,

$$\Pi = \beta\delta_Q + \sum_{m=1}^{M} \alpha_m \delta_{P_m},$$

where $\beta + \sum_m \alpha_m = 1$ and $P_m \in \Lambda_0$ $(1 \leq m \leq M)$. Without loss of generality, assume that ϵ and f are such that Q does not lie in B. Consider,

$$\frac{dP_n^{\Pi|B}}{dP_0^n}(X^n) = \frac{1}{\Pi(B)} \int_B \frac{dP^n}{dP_0^n}(X^n) \, d\Pi(P) \leq \frac{1}{\Pi(B)} \sum_{m=1}^{M} \alpha_m \frac{dP_m^n}{dP_0^n}(X^n).$$

For every $1 \leq m \leq M$, there exists a $k(m)$ such that $P_m(X = k(m)) = 0$, and the probability of the event E_n that none of the $X_1, \ldots, X_n$ equal $k(m)$ is $(1 - P_0(X = k(m)))^n$. Note that E_n is also the event that $dP_m^n / dP_0^n(X^n) > 0$.

Hence for every $1 \leq m \leq M$ and all X in an event of P_0^∞-probability one, there exists an $N_m \geq 1$ such that $dP_m^n / dP_0^n(X^n) = 0$ for all $n \geq N_m$. Consequently, for all X in an event of P_0^∞-probability one, there exists an $N \geq 1$ such that $dP_n^{\Pi|B} / dP_0^n(X^n) = 0$ for all $n \geq N$. Therefore, condition *(ii)* of lemma 3.3 is not satisfied for any sequence $a_n \downarrow 0$. A direct proof that (9) does not hold for any a_n is also possible: given the prior $\Pi \in V$, define,

$$\phi_n(X^n) = \prod_{m=1}^{M} 1_{\{\exists_{1 \leq i \leq n} : X_i = k(m)\}}.$$

Then the expectation of ϕ_n with respect to the local prior predictive distribution equals zero, so $P_n^{\Pi|B} \phi_n = o(a_n)$ for any $a_n \downarrow 0$. However, $P_0^n \phi_n(X^n) \to 1$, so the prior Π does *not* give rise to a sequence of prior predictive distributions $(P_n^{\Pi|B})$ with respect to which (P_0^n) is remotely contiguous, for any $a_n \downarrow 0$. $\qquad\square$

B.2 Consistency, Bayesian tests and the Hellinger metric

Let us first consider characterization of posterior consistency in terms of the family of real-valued functions on the parameter space that are bounded and continuous.

PROPOSITION B.5 *Assume that Θ is a Hausdorff, completely regular space. The posterior is consistent at $\theta_0 \in \Theta$, if and only if,*

$$\int f(\theta) \, d\Pi(\theta | X^n) \xrightarrow{P_{\theta_0, n}} f(\theta_0), \tag{B.25}$$

for every bounded, continuous $f : \Theta \to \mathbb{R}$.

PROOF Assume (13). Let $f : \Theta \to \mathbb{R}$ be bounded and continuous (with $M > 0$ such that $|f| \leq M$). Let $\eta > 0$ be given and let $U \subset \Theta$ be a neighbourhood of θ_0 such that $|f(\theta) - f(\theta_0)| < \eta$ for all $\theta \in U$. Integrate f with respect to the ($P_{\theta_0, n}$-almost-surely well-defined) posterior and to δ_{θ_0}:

$$\left| \int f(\theta) \, d\Pi(\theta | X^n) - f(\theta_0) \right|$$
$$\leq \int_{\Theta \setminus U} |f(\theta) - f(\theta_0)| \, d\Pi(\theta | X^n) + \int_U |f(\theta) - f(\theta_0)| \, d\Pi(\theta | X^n)$$
$$\leq 2M \, \Pi(\Theta \setminus U | X^n) \quad + \sup_{\theta \in U} |f(\theta) - f(\theta_0)| \, \Pi(U | X^n) \leq \eta + o_{P_{\theta_0, n}}(1),$$

as $n \to \infty$, so that (B.25) holds. Conversely, assume (B.25). Let U be an open neighbourhood of θ_0. Because Θ is completely regular, there exists a continuous $f : \Theta \to [0, 1]$ such that $f = 1$ at $\{\theta_0\}$ and $f = 0$ on $\Theta \setminus U$. Then,

$$\Pi(U | X^n) \geq \int f(\theta) \, d\Pi(\theta | X^n) \xrightarrow{P_{\theta_0, n}} \int f(\theta) \, d\delta_{\theta_0}(P) = 1.$$

Consequently, (13) holds. □

Proposition B.5 is used to prove consistency of frequentist point-estimators derived from the posterior.

EXAMPLE B.6 Consider a model $\mathscr{P}$ of single-observation distributions P on $(\mathscr{X}, \mathscr{B})$ for *i.i.d.* data $(X_1, X_2, \ldots, X_n) \sim P^n$, $(n \geq 1)$. Assume that the true distribution of the data is $P_0 \in \mathscr{P}$ and that the model topology is Prohorov's weak topology or stronger. Then for any bounded, continuous $g : \mathscr{X} \to \mathbb{R}$, the map,

$$f : \mathscr{P} \to \mathbb{R} : P \mapsto \big|(P - P_0)g(X)\big|,$$

is continuous. Assuming that the posterior is weakly consistent at P_0,

$$\big|P_1^{\Pi_n | X^n} g - P_0 g\big| \leq \int \big|(P - P_0)g\big| \, d\Pi(P|X^n) \xrightarrow{P_{\theta_0}} 0, \tag{B.26}$$

so posterior predictive distributions are consistent point estimators in Prohorov's weak topology. Replacing the maps g by bounded, measurable maps $\mathscr{X} \to \mathbb{R}$ and assuming posterior consistency in $\mathscr{T}_1$, one proves consistency of posterior predictive distributions in $\mathscr{T}_1$ in exactly the same way. Taking the supremum over measurable $g : \mathscr{X} \to [0, 1]$ in (B.26) and assuming that the posterior is consistent in the total variational topology, posterior predictive distributions are consistent in total variation as frequentist point estimators. □

The vast majority of non-parametric applications of Bayesian methods in the literature is based on the intimate relation that exists between testing and the Hellinger metric (see [51], section 16.4). Proofs concerning posterior consistency or posterior convergence at a rate rely on the existence of tests for small parameter subsets B_n surrounding a point $\theta_0 \in \Theta$, versus the complements V_n of neighbourhoods of the point θ_0. The building block in such constructions is the following application of the minimax theorem.

PROPOSITION B.7 *(Minimax Hellinger tests)*
Consider a model $\mathscr{P}$ of single-observation distributions P for i.i.d. data. Let $B, V \subset \mathscr{P}$ be convex with $H(B, V) > 0$. There exists a test sequence (ϕ_n) such that,

$$\sup_{P \in B} P^n \phi_n \leq e^{-nH^2(B,V)}, \ \sup_{Q \in V} Q^n (1 - \phi_n) \leq e^{-nH^2(B,V)}.$$

PROOF This is an application of the minimax theorem. See Le Cam (1986) [51], section 16.4 for details. □

Questions concerning consistency require the existence of tests in which at least one of the two hypotheses is a non-convex set, typically the complement of a neighbourhood. Imposing the model $\mathscr{P}$ to be of bounded entropy with respect to the Hellinger metric allows construction of such tests, based on the uniform tests of proposition B.7. Below, we apply well-known constructions for the uniform tests in Schwartz's theorem from the frequentist literature [49,

7, 8, 34] to the construction of Bayesian tests. Due to relations that exist between metrics for model parameters and the Hellinger metric in many examples and applications, the material covered here is widely applicable in (non-parametric) models for *i.i.d.* data.

EXAMPLE B.8 Consider a model $\mathscr{P}$ of distributions P for *i.i.d.* data $X^n \sim P^n$, $(n \geq 1)$ and, in addition, suppose that $\mathscr{P}$ is totally bounded with respect to the Hellinger distance. Let $P_0 \in \mathscr{P}$ and $\epsilon > 0$ be given, denote $V(\epsilon) = \{P \in \mathscr{P} : H(P_0, P) \geq 4\epsilon\}$, $B_H(\epsilon) = \{P \in \mathscr{P} : H(P_0, P) < \epsilon\}$. There exists an $N(\epsilon) \geq 1$ and a cover of $V(\epsilon)$ by H-balls $V_1, \ldots, V_{N(\epsilon)}$ of radius ϵ and for any point Q in any V_i and any $P \in B_H(\epsilon)$, $H(Q, P) > 2\epsilon$. According to proposition 2.6 with $\alpha = 1/2$ and (7), for each $1 \leq i \leq N(\epsilon)$ there exists a Bayesian test sequence $(\phi_{i,n})$ for $B_H(\epsilon)$ versus V_i of power (upper bounded by) $\exp(-2n\epsilon^2)$. Then, for any subset $B' \subset B_H(\epsilon)$,

$$
\begin{aligned}
P_n^{\Pi|B'} \Pi(V|X^n) &\leq \sum_{i=1}^{N(\epsilon)} P_n^{\Pi|B'} \Pi(V_i|X^n) \\
&\leq \frac{1}{\Pi(B')} \sum_{i=1}^{N(\epsilon)} \Big(\int_{B'} P^n \phi_n \, d\Pi(P) + \int_{V_i} P^n (1 - \phi_n) \, d\Pi(P) \Big) \qquad \text{(B.27)} \\
&\leq \sum_{i=1}^{N(\epsilon)} \sqrt{\frac{\Pi(V_i)}{\Pi(B')}} \exp(-2n\epsilon^2),
\end{aligned}
$$

which is smaller than or equal to $e^{-n\epsilon^2}$ for large enough n. If $\epsilon = \epsilon_n$ with $\epsilon_n \downarrow 0$ and $n\epsilon_n^2 \to \infty$, and the model's Hellinger entropy is upper-bounded by $\log N(\epsilon_n, \mathscr{P}, H) \leq K n \epsilon_n^2$ for some $K > 0$, the construction extends to tests that separate $V_n = \{P \in \mathscr{P} : H(P_0, P) \geq 4\epsilon_n\}$ from $B_n = \{P \in \mathscr{P} : H(P_0, P) < \epsilon_n\}$ asymptotically, with power $\exp(-nL\epsilon_n^2)$ for some $L > 0$. (See also the so-called *Le Cam dimension* of a model [49] and Birgé's rate-oriented work [7, 8].) It is worth pointing out at this stage that posterior inconsistency due to the phenomenon of 'data tracking' [4, 71], whereby weak posterior consistency holds but Hellinger consistency fails, can only be due to failure of the testing condition in the Hellinger case.

Note that the argument also extends to models that are Hellinger separable: in that case (B.27) remains valid, but with $N(\epsilon) = \infty$. The mass fractions $\Pi(V_i)/\Pi(B')$ become important (we point to strong connections with Walker's theorem [70, 72]). Here we see the balance between prior mass and testing power for Bayesian tests, as intended by the remark that closes the subsection on the existence of Bayesian test sequences in section 2. $\qquad \square$

To balance entropy and prior mass differently in Hellinger separable models, Barron (1988) [3] and Barron *et al.* (1999) [4] formulate an alternative condition that is based on the Radon property that any prior on a Polish space has.

EXAMPLE B.9 Consider a model $\mathscr{P}$ of distributions P for *i.i.d.* data $X^n \sim P^n$, $(n \geq 1)$, with priors (Π_n). Assume that the model $\mathscr{P}$ is Polish in the Hellinger topology. Let $P_0 \in \mathscr{P}$ and $\epsilon > 0$ be given; for a fixed $M > 1$, define $V = \{P \in \mathscr{P} : H(P_0, P) \geq M\epsilon\}$, $B_H = \{P \in$

$\mathscr{P} : H(P_0, P) < \epsilon\}$. For any sequence $\delta_m \downarrow 0$, there exist compacta $K_m \subset \mathscr{P}$ such that $\Pi(K_m) \geq 1 - \delta_m$ for all $m \geq 1$. For each $m \geq 1$, K_m is Hellinger totally bounded so there exists a Bayesian test sequence $\phi_{m,n}$ for $B_H(\epsilon) \cap K_m$ versus $V(\epsilon) \cap K_m$. Since,

$$\int_{B_H} P^n \phi_n \, d\Pi(P) + \int_V Q^n (1 - \phi_n) \, d\Pi(Q)$$
$$\leq \int_{B_H \cap K_m} P^n \phi_{m,n} \, d\Pi(P) \quad + \int_{V \cap K_m} Q^n (1 - \phi_{m,n}) \, d\Pi(Q) + \delta_m,$$

and all three terms go to zero, a diagonalization argument confirms the existence of a Bayesian test for B_H versus V. To control the power of this test and to generalise to the case where $\epsilon = \epsilon_n$ is n-dependent, more is required: as we increase m with n, the prior mass $\delta_{m(n)}$ outside of $K_n = K_{m(n)}$ must decrease fast enough, while the order of the cover must be bounded: if $\Pi_n(K_n) \geq 1 - \exp(-L_1 n \epsilon_n^2)$ and the Hellinger entropy of K_n satisfies $\log N(\epsilon_n, K_n, H) \leq L_2 n \epsilon_n^2$ for some $L_1, L_2 > 0$, there exist $M > 1$, $L > 0$, and a sequence of tests (ϕ_n) such that,

$$\int_{B_H(\epsilon_n)} P^n \phi_n \, d\Pi(P) + \int_{V(\epsilon_n)} Q^n (1 - \phi_n) \, d\Pi(Q) \leq e^{-Ln\epsilon_n^2},$$

for large enough n. (For related constructions, see Barron (1988) [3], Barron *et al.* (1999) [4] and Ghosal, Ghosh and van der Vaart (2000) [34].)

To apply corollary 4.3 consider the following steps.

EXAMPLE B.10 As an example of the tests required under condition *(i)* of corollary 4.3, consider $\mathscr{P}$ in the Hellinger topology, assuming totally-boundedness. Let U be the Hellinger-ball of radius 4ϵ around P_{θ_0} of example B.8 and let V be its complement. The Hellinger ball $B_H(\epsilon)$ in equation (B.27) contains the set $K(\epsilon)$. Alternatively we may consider the model in any of the weak topologies $\mathscr{T}_n$: let $\epsilon > 0$ be given and let U denote a weak neighbourhood of the form $\{P \in \mathscr{P} : |(P^n - P_0^n)f| \geq 2\epsilon\}$, for some bounded measurable $f : \mathscr{X}_n \to [0,1]$, as in proposition A.6. The set B of proposition A.6 contains a set $K(\delta)$, for some $\delta > 0$. Both these applications were noted by Schwartz in [62].

B.3 Some examples of remotely contiguous sequences

The following two examples illustrate the difference between contiguity and remote contiguity in the context of parametric and non-parametric regression.

EXAMPLE B.11 Let $\mathscr{F}$ denote a class of functions $\mathbb{R} \to \mathbb{R}$. We consider samples $X^n = ((X_1, Y_1), \ldots, (X_n, Y_n))$, $(n \geq 1)$ of points in $\mathbb{R}^2$, assumed to be related through $Y_i = f_0(X_i) + e_i$ for some unknown $f_0 \in \mathscr{F}$, where the errors are *i.i.d.* standard normal $e_1, \ldots, e_n \sim N(0,1)^n$ and independent of the *i.i.d.* covariates $X_1, \ldots, X_n \sim P^n$, for some (ancillary) distribution P on $\mathbb{R}$. It is assumed that $\mathscr{F} \subset L^2(P)$ and we use the L^2-norm $\|f\|_{P,2}^2 = \int f^2 \, dP$ to define a metric d on $\mathscr{F}$, $d(f, g) = \|f - g\|_{P,2}$. Given a parameter $f \in \mathscr{F}$, denote the sample distributions as $P_{f,n}$.

We distinguish two cases: *(a)* the case of linear regression, where $\mathscr{F} = \{f_\theta : \mathbb{R} \to \mathbb{R} : \theta \in \Theta\}$, where $\theta = (a, b) \in \Theta = \mathbb{R}^2$ and $f_\theta(x) = ax + b$; and *(b)* the case of non-parametric regression, where we do not restrict $\mathscr{F}$ beforehand.

Let Π be a Borel prior Π on $\mathscr{F}$ and place remote contiguity in context by assuming, for the moment, that for some $\rho > 0$, there exist $0 < r < \rho$ and $\tau > 0$, as well as Bayesian tests ϕ_n for $B = \{f \in \mathscr{F} : \|f - f_0\|_{P,2} < r\}$ versus $V = \{f \in \mathscr{F} : \|f - f_0\|_{P,2} \geq \rho\}$ under Π of power $a_n = \exp(-\frac{1}{2}n\tau^2)$. If this is the case, we may assume that $r < \frac{1}{2}\tau$ without loss of generality. Suppose also that Π has a support in $L^2(P)$ that contains all of $\mathscr{F}$.

Let us concentrate on case *(b)* first: a bit of manipulation casts the a_n-rescaled likelihood ratio for $f \in \mathscr{F}$ in the following form,

$$a_n^{-1} \frac{dP_{f,n}}{dP_{f_0,n}}(X^n) = e^{-\frac{1}{2}\sum_{i=1}^n \left(e_i(f-f_0)(X_i) + (f-f_0)^2(X_i) - \tau^2\right)}, \tag{B.28}$$

under $X^n \sim P_{f_0,n}$. The exponent is controlled by the law of large numbers,

$$\frac{1}{n}\sum_{i=1}^n \left(e_i(f - f_0)(X_i) + (f - f_0)^2(X_i) - \tau^2\right) \xrightarrow{P_{f_0}^\infty\text{-a.s.}} \|f - f_0\|_{P,2}^2 - \tau^2.$$

Hence, for every $\epsilon > 0$ there exists an $N(f, \epsilon) \geq 1$ such that the exponent in (B.28) satisfies the upper bound,

$$\sum_{i=1}^n \left(e_i(f - f_0)(X_i) + (f - f_0)^2(X_i) - \tau^2\right) \leq n(\|f - f_0\|_{P,2}^2 - \tau^2 + \epsilon^2),$$

for all $n \geq N(f, \epsilon)$. Since $\Pi(B) > 0$, we may condition Π on B, choose $\epsilon = \frac{1}{2}\tau$ and use Fatou's inequality to find that,

$$\liminf_{n\to\infty} e^{\frac{1}{2}n\tau^2} \frac{dP_n^{\Pi|B}}{dP_{f_0,n}}(X^n) \geq \liminf_{n\to\infty} e^{\frac{1}{4}n\tau^2} = \infty,$$

$P_{f_0}^\infty$-almost-surely. Consequently, for any choice of δ,

$$P_{f_0,n}\left(\frac{dP_n^{\Pi|B}}{dP_{f_0,n}}(X^n) < \delta\, e^{-\frac{1}{2}n\tau^2}\right) \to 0,$$

and we conclude that $P_{f_0,n} \lhd e^{\frac{1}{2}n\tau^2} P_n^{\Pi|B}$. Based on theorem 4.2, we conclude that,

$$\Pi\left(\|f - f_0\|_{P,2} < \rho \mid X^n\right) \xrightarrow{P_{f_0,n}} 1,$$

i.e. posterior consistency for the regression function in $L^2(P)$-norm obtains. $\qquad\square$

EXAMPLE B.12 As for case *(a)*, one has the choice of using a prior like above, but also to proceed differently: expression (B.28) can be written in terms of a local parameter $h \in \mathbb{R}^k$ which, for given θ_0 and $n \geq 1$, is related to θ by $\theta = \theta_0 + n^{-1/2}h$. For $h \in \mathbb{R}^2$, we write $P_{h,n} = P_{\theta_0 + n^{-1/2}h,n}$, $P_{0,n} = P_{\theta_0,n}$ and rewrite the likelihood ratio (B.28) as follows,

$$\frac{dP_{h,n}}{dP_{0,n}}(X^n) = e^{\frac{1}{\sqrt{n}}\sum_{i=1}^n h\cdot\ell_{\theta_0}(X_i,Y_i) - \frac{1}{2}h\cdot I_{\theta_0}\cdot h + R_n}, \tag{B.29}$$

where $\ell_{\theta_0} : \mathbb{R}^2 \to \mathbb{R}^2 : (x, y) \mapsto (y - a_0 x - b_0)(x, 1)$ is the score function for θ, $I_{\theta_0} = P_{\theta_0,1} \ell_{\theta_0} \ell_{\theta_0}^T$ is the Fisher information matrix and $R_n \xrightarrow{P_{\theta_0,n}} 0$. Assume that I_{θ_0} is non-singular and note the central limit,

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \ell_{\theta_0}(X_i, Y_i) \xrightarrow{P_{\theta_0,n}\text{-w.}} N_2(0, I_{\theta_0}),$$

which expresses local asymptotic normality of the model [48] and implies that for any fixed $h \in \mathbb{R}^2$, $P_{h,n} \lhd P_{0,n}$.

LEMMA B.13 *Assume that the model satisfies LAN condition (B.29) with non-singular I_{θ_0} and that the prior Π for θ has a Lebesgue-density $\pi : \mathbb{R}^d \to \mathbb{R}$ that is continuous and strictly positive in all of Θ. For given $H > 0$, define the subsets $B_n = \{\theta \in \Theta : \theta = \theta_0 + n^{-1/2}h, \|h\| \le H\}$. Then,*

$$P_{0,n} \lhd c_n^{-1} P_n^{\Pi | B_n}, \tag{B.30}$$

for any $c_n \downarrow 0$.

PROOF According to lemma 3 in section 8.4 of Le Cam and Yang (1990) [53], $P_{\theta_0,n}$ is contiguous with respect to $P_n^{\Pi | B_n}$. That implies the assertion. $\qquad\square$

Note that for some $K > 0$, $\Pi(B_n) \ge b_n := K(H/\sqrt{n})^d$. Assume again the existence of Bayesian tests for $V = \{\theta \in \Theta : \|\theta - \theta_0\| > \rho\}$ (for some $\rho > 0$) versus B_n (or some B such that $B_n \subset B$), of power $a_n = \exp(-\frac{1}{2}n\tau^2)$ (for some $\tau > 0$). Then $a_n b_n^{-1} = o(1)$, and, assuming (B.30), theorem 4.4 implies that $\Pi(\|\theta - \theta_0\| > \rho | X^n) \xrightarrow{P_{\theta_0,n}} 0$, so consistency is straightforwardly demonstrated.

The case becomes somewhat more complicated if we are interested in optimality of parametric rates: following the above, a logarithmic correction arises from the lower bound $\Pi(B_n) \ge K(H/\sqrt{n})^d$ when combined in the application of theorem 4.4. To alleviate this, we adapt the construction somewhat: define $V_n = \{\theta \in \Theta : \|\theta - \theta_0\| \le M_n n^{-1/2}\}$ for some $M_n \to \infty$ and B_n like above. Under the condition that there exists a uniform test sequence for any *fixed* $V = \{\theta \in \Theta : \|\theta - \theta_0\| > \rho\}$ versus B_n (see, for example, [45]), uniform test sequences for V_n versus B_n of power $e^{-K'M_n^2}$ exist, for some $k' > 0$. Alternatively, assume that the Hellinger distance and the norm on Θ are related through inequalities of the form,

$$K_1 \|\theta - \theta'\| \le H(P_\theta, P_{\theta'}) \le K_2 \|\theta - \theta'\|,$$

for some constants $K_1, K_2 > 0$. Then cover V_n with rings,

$$V_{n,k} = \left\{ \theta \in V_n : \frac{(M_n + k - 1)}{\sqrt{n}} \le \|\theta - \theta_0\| \le \frac{(M_n + k)}{\sqrt{n}} \right\},$$

for $k \ge 1$ and cover each ring with balls $V_{n,k,l}$ of radius $n^{-1/2}$, where $1 \le l \le L_{n,k}$ and $L_{n,k}$ the minimal number of radius-$n^{-1/2}$ balls needed to cover $V_{n,k}$, related to the *Le Cam dimension*

33

[49]. With the B_n defined like above, and the inequality,

$$\int P_{\theta,n}\Pi(V_{n,k,l}|X^n)\,d\Pi_n(\theta|B_n)$$

$$\leq \sup_{\theta\in B_n} P_{\theta,n}\phi_{n,k,l} + \frac{\Pi_n(V_{n,k,l})}{\Pi_n(B_n)}\sup_{\theta\in V_{n,k,l}} P_{\theta,n}(1-\phi_{n,k,l}),$$

where the $\phi_{n,k,l}$ are the uniform minimax tests for B_n versus $V_{n,k,l}$ of lemma B.7, of power $\exp(-K'(M_n+k-1)^2)$ for some $K' > 0$. We define $\phi_{n,k} = \max\{\phi_{n,k,l} : 1 \leq l \leq L_{n,k}\}$ for $V_{n,k}$ versus B_n and note,

$$\int P_{\theta,n}\Pi(V_{n,k}|X^n)\,d\Pi_n(\theta|B_n) \leq \left(L_{n,k} + \frac{\Pi_n(V_{n,k})}{\Pi_n(B_n)}\right)e^{-K(M_n+k-1)^2},$$

where the numbers $L_{n,k}$ are upper bounded by a multiple of $(M_n+k)^d$ and the fraction of prior masses $\Pi_n(V_{n,k})/\Pi_n(B_n)$ can be controlled without logarithmic corrections when summing over k next. $\qquad\square$

But remote contiguity also applies in more irregular situations: example 1.3 does not admit KL priors, but satisfies the requirement of remote contiguity. (Choose η equal to the uniform density for simplicity.)

EXAMPLE B.14 Consider $X_1, X_2, \ldots$ that form an *i.i.d.* sample from the uniform distribution on $[\theta, \theta+1]$, for unknown $\theta \in \mathbb{R}$. The model is parametrized in terms of distributions P_θ with Lebesgue densities of the form $p_\theta(x) = 1_{[\theta,\theta+1]}(x)$, for $\theta \in \Theta = \mathbb{R}$. Pick a prior Π on Θ with a continuous and strictly positive Lebesgue density $\pi : \mathbb{R} \to \mathbb{R}$ and, for some rate $\delta_n \downarrow 0$, choose $B_n = (\theta_0, \theta_0 + \delta_n)$. Note that for any $\alpha > 0$, there exists an $N \geq 1$ such that for all $n \geq N$, $(1-\alpha)\pi(\theta_0)\delta_n \leq \Pi(B_n) \leq (1+\alpha)\pi(\theta_0)\delta_n$. Note that for any $\theta \in B_n$ and $X^n \sim P_{\theta_0}^n$, $dP_\theta^n/dP_{\theta_0}^n(X^n) = 1\{X_{(1)} > \theta\}$, and correspondingly,

$$\frac{dP_n^{\Pi|B_n}}{dP_{\theta_0}^n}(X^n) = \Pi_n(B_n)^{-1}\int_{\theta_0}^{\theta_0+\delta_n} 1\{X_{(1)} > \theta\}\,d\Pi(\theta)$$

$$\geq \frac{1-\alpha}{1+\alpha}\frac{\delta_n \wedge (X_{(1)} - \theta_0)}{\delta_n},$$

for large enough n. As a consequence, for every $\delta > 0$ and all $a_n \downarrow 0$,

$$P_{\theta_0}^n\left(\frac{dP_n^{\Pi|B_n}}{dP_{\theta_0}^n}(X^n) < \delta\, a_n\right) \leq P_{\theta_0}^n\left(\delta_n^{-1}(X_{(1)} - \theta_0) < (1+\alpha)\delta\, a_n\right),$$

for large enough $n \geq 1$. Since $n(X_{(1)} - \theta_0)$ has an exponential weak limit under $P_{\theta_0}^n$, we choose $\delta_n = n^{-1}$, so that the *r.h.s.* in the above display goes to zero. So $P_{\theta_0,n} \lhd a_n^{-1}P_n^{\Pi_n|B_n}$, for any $a_n \downarrow 0$. $\qquad\square$

To show consistency and derive the posterior rate of convergence in example 1.3, we use theorem 4.4.

EXAMPLE B.15 Continuing with example B.14, we define $V_n = \{\theta : \theta - \theta_0 > \epsilon_n\}$. It is noted that, for every $0 < c < 1$, the likelihood ratio test,

$$\phi_n(X^n) = 1\{dP_{\theta_0+\epsilon_n,n}/dP_{\theta_0,n}(X^n) > c\} = 1\{X_{(1)} > \theta_0 + \epsilon_n\},$$

satisfies $P_\theta^n(1 - \phi_n)(X^n) = 0$ for all $\theta \in V_n$, and if we choose $\delta_n = 1/2$ and $\epsilon_n = M_n/n$ for some $M_n \to \infty$, $P_\theta^n \phi_n \le e^{-M_n+1}$ for all $\theta \in B_n$, so that,

$$\int_{B_n} P_\theta^n \phi_n(\, d\Pi(\theta) + \int_{V_n} P_\theta^n(1 - \phi_n)\, d\Pi(\theta) \le \Pi(B_n)\, e^{-M_n+1},$$

Using lemma 2.2, we see that $P_n^{\Pi|B_n}\Pi(V_n|X^n) \le e^{-M_n+1}$. Based on the conclusion of example B.14 above, remote contiguity implies that $P_{\theta_0}^n\Pi(V_n|X^n) \to 0$. Treating the case $\theta < \theta_0 - \epsilon_n$ similarly, we conclude that the posterior is consistent at (any ϵ_n slower than) rate $1/n$.

To conclude, we demonstrate the relevance of priors satisfying the lower bound (16). Let us repeat lemma 8.1 in [34], to demonstrate that the sequence (P_0^n) is remotely contiguous with respect to the local prior predictive distributions based on the B_n of example 4.6.

LEMMA B.16 *For all $n \ge 1$, assume that $(X_1, X_2, \dots, X_n) \in \mathscr{X}^n \sim P_0^n$ for some $P_0 \in \mathscr{P}$ and let $\epsilon_n \downarrow 0$ be given. Let B_n be as in example 4.6. Then, for any priors Π_n such that $\Pi_n(B_n) > 0$,*

$$P_{\theta_0,n}\bigg(\int \frac{dP_\theta^n}{dP_{\theta_0}^n}(X^n)\, d\Pi_n(\theta|B_n) < e^{-cn\epsilon_n^2} \bigg) \to 0,$$

for any constant $c > 1$.

B.4 The sparse normal means problem

For an example of consistency in the false-detection-rate (FDR) sense, we turn to the most prototypical instance of sparsity, the so-called *sparse normal means problem*: in recent years various types of priors have been proposed for the Bayesian recovery of a nearly-black vector in the Gaussian sequence model. Most intuitive in this context is the class of *spike-and-slab priors* [57], which first select a sparse subset of non-zero components and then draws those from a product distribution. But other proposals have also been made, *e.g.* the horseshoe prior [14], a scale-mixture of normals. Below, we consider FDR-type consistency with spike-and-slab priors.

EXAMPLE B.17 Estimation of a nearly-black vector of locations in the Gaussian sequence model is based on n-point samples $X^n = (X_1, \dots, X_n)$ assumed distributed according to,

$$X_i = \theta_i + \varepsilon_i, \tag{B.31}$$

(for all $1 \le i \le n$), where $\varepsilon_1, \dots, \varepsilon_n$ form an *i.i.d.* sample of standard-normally distributed errors. The parameter θ is a sequence $(\theta_i : i \ge 1)$ in $\mathbb{R}$, with n-dimensional projection

$\theta^n = (\theta_1, \ldots, \theta_n)$, for every $n \geq 1$. The corresponding distributions for X^n are denoted $P_{\theta,n}$ for all $n \geq 1$.

Denoting by p_n the number of non-zero components of the vector $\theta^n = (\theta_1, \ldots, \theta_n) \in \mathbb{R}^n$, sparsity is imposed through the assumption that θ is *nearly black*, that is, $p_n \to \infty$, but $p_n = o(n)$ as $n \to \infty$. For any integer $0 \leq p \leq n$, denote the space of n-dimensional vectors θ^n with exactly p non-zero components by $\ell_{0,n}(p)$. For later reference, we introduce, for every subset S of $I_n := \{1, \ldots, n\}$, the space $R_n^S := \{\theta^n \in \mathbb{R}^n : \theta_i = 1\{i \in S\}\theta_i, 1 \leq i \leq n\}$.

Popular sub-problems concern selection of the non-zero components [10] and (subsequent) minimax-optimal estimation of the non-zero components [25] (especially with the LASSO in related regression problems, see, for example, [75]). Many authors have followed Bayesian approaches; for empirical priors, see [44], and for hierarchical priors, see [15] (and references therein).

As n grows, the minimax-rate at which the L_2-error for estimation of θ^n grows, is bounded in the following, sparsity-induced way [24],

$$\inf_{\hat{\theta}^n} \sup_{\theta^n \in \ell_{0,n}(p_n)} P_{\theta,n} \|\hat{\theta}^n - \theta^n\|_{2,n}^2 \leq 2p_n \log \frac{n}{p_n}(1 + o(1)),$$

as $n \to \infty$, where $\hat{\theta}^n$ runs over all estimators for θ^n.

A natural proposal for a prior Π for θ [57] (or rather, priors Π_n for all θ^n ($n \geq 1$)), is to draw a sparse θ^n hierarchically [15]: given $n \geq 1$, first draw $p \sim \pi_n$ (for some distribution π_n on $\{0, 1, \ldots, n\}$), then draw a subset S of order p from $\{1, \ldots, n\}$ uniformly at random, and draw θ^n by setting $\theta_i = 0$ if $i \notin S$ and $(\theta_i : i \in S) \sim G^p$, for some distribution G on (all of) $\mathbb{R}$. The components of θ^n can therefore be thought of as having been drawn from a mixture of a distribution degenerate at zero (the *spike*) and a full-support distribution G (the *slab*).

To show that methods presented in this paper also apply in complicated problems like this, we give a proof of posterior convergence in the FDR sense. We appeal freely to useful results that appeared elsewhere, in particular in [15]: we adopt some of Castillo and van der Vaart's more technical steps to reconstitute the FDR-consistency proof based on Bayesian testing and remote contiguity: to compare, the testing condition and prior-mass lower bound of theorem 4.4 are dealt with simultaneously, while the remote contiguity statement is treated separately. (We stress that only the way of organising the proof, not the result is new. In fact, we prove only part of what [15] achieves.)

Assume that the data follows (B.31) and denote by θ_0 the true vector of normal means. For each $n \geq 1$, let p_n (respectively p) denote number of non-zero components of θ_0^n (respectively θ^n). We do not assume that the true degree of sparsity p_n is fully known, but for simplicity and brevity we assume that there is a known sequence of upper bounds q_n, such that for some constant $A > 1$, $p_n \leq q_n \leq A p_n$, for all $n \geq 1$. (Indeed, theorem 2.1 in [15] very cleverly shows

that if G has a second moment and the prior density for the sparsity level has a tail that is slim enough, then the posterior concentrates on sets of the form, $\{\theta^n \in \mathbb{R}^n : p \leq Ap_n\}$ under P_0, for some $A > 1$.)

Set $r_n^2 = p_n \log(n/p_n)$ and define two subsets of $\mathbb{R}^n$,

$$V_n = \big\{\theta^n : p \leq Ap_n, \|\theta^n - \theta_0^n\|_{2,n} > Mr_n\big\},$$
$$B_n = \big\{\theta^n : \|\theta^n - \theta_0^n\|_{2,n} \leq d\, r_n, \big\},$$

assuming for future reference that $\Pi(B_n) > 0$. As for V_n, we split further: define, for all $j \geq 1$,

$$V_{n,j} = \big\{\theta^n \in V_n : jMr_n < \|\theta^n - \theta_0^n\|_{2,n} \leq (j+1)Mr_n\big\}.$$

Next, we subdivide $V_{n,j}$ into intersections with the spaces R_n^S for $S \subset I_n$: we write $V_{n,j} = \cup\{V_{n,S,j} : S \subset I_n\}$ with $V_{n,S,j} = V_{n,j} \cap R_n^S$. For every $n \geq 1, j \geq 1$ and $S \subset I_n$, we cover $V_{n,S,j}$ by $N_{n,S,j}$ L_2-balls $V_{n,j,S,i}$ of radius $\frac{1}{2}jMr_n$ and centre points $\theta_{j,S,i}$. Comparing the problem of covering $V_{n,j}$ with that of covering $V_{n,1}$, one realizes that $N_{n,S,j} \leq N_{n,S} := N_{n,S,1}$.

Fix $n \geq 1$. Due to lemma 2.2, for any test sequences $\phi_{n,j,S,i}$,

$$P_n^{\Pi|B_n} \Pi(V_n|X^n) \leq \sum_{j \geq 1} \sum_{S \subset I_n} \sum_{i=1}^{N_{n,S,j}} P_n^{\Pi|B_n} \Pi(V_{n,j,S,i}|X^n)$$

$$\leq \frac{1}{\Pi(B_n)} \sum_{j \geq 1} \sum_{S \subset I_n} \sum_{i=1}^{N_{n,S,j}}$$
$$\left(\int_{B_n} P_{\theta,n} \phi_{n,j,S,i}\, d\Pi(\theta) + \int_{V_{n,j,S,i}} P_{\theta,n}(1 - \phi_{n,j,S,i})\, d\Pi(\theta) \right)$$

$$\leq \sum_{p=0}^{Ap_n} \binom{n}{p} \sum_{j \geq 1} N_{n,S} \frac{a_n(j)}{b_n},$$

where $b_n := \Pi(B_n)$, $a_n(j) := \max_{S \subset I_n, 1 \leq i \leq N_{n,S,j}} a_n(j, S, i)$ and,

$$\frac{a_n(j, S, i)}{b_n} = \int P_{\theta,n} \phi_{n,j,S,i}\, d\Pi(\theta|B_n)$$
$$+ \frac{\Pi(V_{n,j,S,i})}{\Pi(B_n)} \int P_{\theta,n}(1 - \phi_{n,j,S,i})\, d\Pi(\theta|V_{n,j,S,i})$$
$$\leq \sup_{\theta^n \in B_n} P_{\theta,n} \phi_{n,j,S,i} + \frac{\Pi(V_{n,j,S,i})}{\Pi(B_n)} \sup_{\theta^n \in V_{n,j,S,i}} P_{\theta,n}(1 - \phi_{n,j,S,i}).$$

A standard argument (see lemma 5.1 in [15]) shows that there exists a test $\phi_{n,j,S,i}$ such that,

$$P_{0,n} \phi_{n,j,S,i} + \frac{\Pi(V_{n,j,S,i})}{\Pi(B_n)} \sup_{\theta^n \in V_{n,j,S,i}} P_{\theta,n}(1 - \phi_{n,j,S,i})$$

$$\leq 2\sqrt{\frac{\Pi(V_{n,j,S,i})}{\Pi(B_n)}} e^{-\frac{1}{128}j^2 M^2 p_n \log(n/p_n)}$$

Note that for every measurable $0 \leq \phi \leq 1$, Cauchy's inequality implies that, for all $\theta, \theta' \in \mathbb{R}^n$

$$P_{\theta,n}\phi \leq \left(P_{\theta',n}\phi^2\right)^{1/2}\left(P_{\theta',n}(dP_{\theta,n}/dP_{\theta',n})^2\right)^{1/2} \leq \left(P_{\theta',n}\phi\right)^{1/2}e^{\frac{1}{2}\|\theta-\theta'\|_{2,n}^2} \tag{B.32}$$

We use this to generalise the first term in the above display to the test uniform over B_n at the expense of an extra factor, that is,

$$\sup_{\theta^n \in B_n} P_{\theta,n}\phi_{n,j,S,i} + \frac{\Pi(V_{n,j,S,i})}{\Pi(B_n)} \sup_{\theta^n \in V_{n,j,S,i}} P_{\theta,n}(1 - \phi_{n,j,S,i})$$
$$\leq 2\sqrt{\frac{\Pi(V_{n,j,S,i})}{\Pi(B_n)}}e^{-\frac{1}{256}j^2M^2 r_n^2+\frac{1}{2}d^2 r_n^2}$$

In what appears to be one of the essential (and technically very demanding) points of [15], the proofs of the lemma 5.4 (only after the first line) and of proposition 5.1 show that there exists a constant $K > 0$ such that,

$$\sqrt{\frac{\Pi(V_{n,j,S,i})}{\Pi(B_n)}} \leq e^{Kr_n^2},$$

if G has a Lebesgue density $g : \mathbb{R} \to \mathbb{R}$ such that there exists a constant $c > 0$ such that $|\log g(\theta) - \log g(\theta')| \leq c(1 + |\theta - \theta'|)$ for all $\theta, \theta' \in \mathbb{R}$. This allows for demonstration that (see the final argument in the proof of proposition 5.1 in [15]) if we choose $M > 0$ large enough, there exists a constant $K' > 0$ such that for large enough n,

$$P_n^{\Pi|B_n}\Pi(V_n|X^n) \leq e^{-K'r_n^2}.$$

Remote contiguity follows from (B.32): fix some $n \geq 1$ and note that for any $\theta^n \in B_n$,

$$(P_{0,n}\phi)^2 \leq e^{d^2 r_n^2}P_{\theta,n}\phi.$$

Integrating with respect to $\Pi(\cdot|B_n)$ on both sides shows that,

$$P_{0,n}\phi \leq e^{\frac{d^2}{2} r_n}(P_n^{\Pi|B_n}\phi)^{1/2},$$

so that $P_{0,n} \lhd e^{d^2 r_n^2}P_n^{\Pi|B_n}$. So if we choose $d^2 < K'$, remote contiguity guarantees that $P_{0,n}\Pi(V_n|X^n) \to 0$.

B.5   Goodness-of-fit Bayes factors for random walks

Consider the asymptotic consistency of goodness-of-fit tests for the transition kernel of a Markov chain with posterior odds or Bayes factors. Bayesian analyses of Markov chains on a finite state space are found in [66] and references therein. Consistency results *c.f.* [70] for random walk data are found in [36]. Large-deviation results for posterior distributions are derived in [59, 27]. The examples below are based on ergodicity for remote contiguity and Hoeffding's inequality for uniformly ergodic Markov chains [56, 37] to construct suitable tests. We first prove the analogue of Schwartz's construction in the case of an ergodic random walk.

Let $(S, \mathscr{S})$ denote a measurable state space for a discrete-time, stationary Markov process P describing a random walk $X^n = \{X_i \in S : 0 \leq i \leq n\}$ of length $n \geq 1$ (conditional on a starting position X_0). The chain has a Markov transition kernel $P(\cdot|\cdot) : \mathscr{S} \times S \to [0, 1]$ that describes $X_i|X_{i-1}$ for all $i \geq 1$.

Led by Pearson's approach to goodness-of-fit testing, we choose a finite partition $\alpha = \{A_1, \ldots, A_N\}$ of S and 'bin the data' in the sense that we switch to a new process Z^n taking values in the finite state space $S_\alpha = \{e_j : 1 \leq j \leq N\}$ (where e_j denotes the j-th standard basis vector in $\mathbb{R}^n$), defined by $Z^n = \{Z_i \in S_\alpha : 0 \leq i \leq n\}$, with $Z_i = (1\{X_i \in A_1\}, \ldots, 1\{X_i \in A_N\})$. The process Z^n forms a stationary Markov chain on S_α with distribution $P_{\alpha,n}$. The model is parametrized in terms of the convex set Θ of $N \times N$ Markov transition matrices p_α on the finite state space S_α,

$$p_\alpha(k|l) = P_{\alpha,n}(Z_i = e_k|Z_{i-1} = e_l) = P(X_i \in A_k|X_{i-1} \in A_l), \tag{B.33}$$

for all $0 \leq i \leq n$ and $1 \leq k, l \leq N$. We assume that $P_{\alpha,n}$ is ergodic with equilibrium distribution that we denote by π_α, and $\pi_\alpha(k) := \pi_\alpha(Z = k)$. We are interested in Bayes factors for goodness-of-fit type questions, given a parameter space consisting of transition matrices.

EXAMPLE B.18 Assume that the true transition kernel P_0 gives rise to a matrix $p_0 \in \Theta$ that generates an ergodic Markov chain Z^n. Denote the true distribution of Z^n by $P_{0,n}$ and the equilibrium distribution by π_0 (with $\pi_0(k) := \pi_0(Z = k)$). For given $\epsilon > 0$, define,

$$B' = \Big\{p_\alpha \in \Theta : \sum_{k,l=1}^N -p_0(l|k)\pi_0(k) \log \frac{p_\alpha(l|k)}{p_0(l|k)} < \epsilon^2\Big\}.$$

Assume that $\Pi(B') > 0$. According to the ergodic theorem, for every $p_\alpha \in B'$,

$$\frac{1}{n} \sum_{i=1}^n \log \frac{p_\alpha(Z_i|Z_{i-1})}{p_0(Z_i|Z_{i-1})} \xrightarrow{P_{0,n}\text{-a.s.}} \sum_{k,l=1}^N p_0(l|k)\pi_0(k) \log \frac{p(l|k)}{p_0(l|k)},$$

(compare with the rate-function in the large-deviation results in [59, 27]) so that, for large enough n,

$$\frac{dP_{\alpha,n}}{dP_{0,n}}(Z^n) = \prod_{i=1}^n \frac{p_\alpha(Z_i|Z_{i-1})}{p_0(Z_i|Z_{i-1})} \geq e^{-\frac{n}{2}\epsilon^2},$$

$P_{0,n}$-almost-surely. Just like in Schwartz's proof [62], in proposition B.22 and in example B.11, the assumption $\Pi(B') > 0$ and Fatou's lemma imply remote contiguity because,

$$P_{0,n}\Big(\int \frac{dP_{\alpha,n}}{dP_{0,n}}(Z^n) \, d\Pi(p_\alpha|B') < e^{-\frac{n}{2}\epsilon^2}\Big) \to 0.$$

So lemma 3.3 says that $P_{0,n} \lhd \exp(\frac{n}{2}\epsilon^2) P_n^{\Pi|B'}$.

However, exponential remote contiguity will turn out not to be enough for goodness-of-fit tests below, unless we impose stringent model conditions. Instead, we shall resort to local asymptotic normality for a sharper result.

EXAMPLE B.19 We formulate goodness-of-fit hypotheses in terms of the joint distribution for two consecutive steps in the random walk. Like Pearson, we fix some such distribution P_0 and consider hypotheses based on differences of 'bin probabilities' $p_\alpha(k,l) = p_\alpha(k|l)\pi_\alpha(l)$,

$$H_0 : \max_{1 \leq k,l \leq N} \left| p_\alpha(k,l) - p_0(k,l) \right| < \epsilon,$$
$$H_1 : \max_{1 \leq k,l \leq N} \left| p_\alpha(k,l) - p_0(k,l) \right| \geq \epsilon,$$
(B.34)

for some fixed $\epsilon > 0$. The sets B and V are defined as the sets of transition matrices $p_\alpha \in \Theta$ that satisfy hypotheses H_0 and H_1 respectively. We assume that the prior is chosen such that $\Pi(B) > 0$ and $\Pi(V) > 0$.

Endowed with some matrix norm, Θ is compact and a Borel prior on Θ can be defined in various ways. For example, we may assign the vector $(p_\alpha(\cdot|1), \ldots, p_\alpha(\cdot|N))$ a product of Dirichlet distributions. Conjugacy applies and the posterior for p_α is again a product of Dirichlet distributions [66]. For an alternative family of priors, consider the set $\mathscr{E}$ of N^N $N \times N$-matrices E that have standard basis vectors e_k in $\mathbb{R}^N$ as columns. Each $E \in \mathscr{E}$ is a deterministic Markov transition matrix on S_α and $\mathscr{E}$ is the extremal set of the polyhedral set Θ. According to Choquet's theorem, every transition matrix p_α can then be written in the form,

$$p_\alpha = \sum_{E \in \mathscr{E}} \lambda_E \, E,$$
(B.35)

for a (non-unique) combination of $\lambda_{\mathscr{E}} := \{\lambda_E : E \in \mathscr{E}\}$ such that $\lambda_E \geq 0$, $\sum_{\mathscr{E}} \lambda_E = 1$. If $\lambda_E > 0$ for all $E \in \mathscr{E}$, the resulting Markov chain is ergodic and we denote the corresponding distributions for Z^n by $P_{\alpha,n}$. Any Borel prior Π' (e.g. a Dirichlet distribution) on the simplex S_{N^N} in $\mathbb{R}^{N^N}$ is a prior for $\lambda_{\mathscr{E}}$ and induces a Borel prior Π on Θ. Note that all non-ergodic transition matrices lie in the boundary $\partial\Theta$, so if we choose Π' such that $\Pi(\mathring{\Theta}) = 1$, ergodicity may be assumed in all prior-almost-sure arguments. This is true for any Π' that is absolutely continuous with respect to the ($N^N - 1$-dimensional) Lebesgue measure on S_{N^N} (for example when we choose Π' equal to a Dirichlet distribution). Note that if the associated density is continuous and strictly positive, $\Pi(B) > 0$ and $\Pi(V) > 0$.

We intend to use theorem 4.8 with B and V defined by H_0 and H_1, so we first demonstrate that a Bayesian test sequence for B versus V exists, based on a version of Hoeffding's inequality valid for random walks [37]. First, define, for given $0 < \lambda_n \leq N^{-N}$ such that $\lambda_n \downarrow 0$,

$$S'_n := \left\{ \lambda_{\mathscr{E}} \in S^{N^N} : \lambda_E \geq \lambda_n / N^{N-1}, \text{for all } E \in \mathscr{E} \right\},$$

and denote the image of S'_n under (B.35) by S_n. Note that if $\Pi(\partial\Theta) = 0$, then $\pi_{S,n} := \Pi(\Theta \setminus S_n) \to 0$.

Now fix $n \geq 1$ for the moment. Recalling the nature of the matrices E, we see that for every $1 \leq k, l \leq N$, $p_\alpha(k|l)$ as in equation (B.35) is greater than or equal to λ_n. Consequently, the corresponding Markov chain satisfies condition (A.1) of Glynn and Ormoneit [37] (closely

related to the notion of uniform ergodicity [56]): starting in any point X_0 under a transition from S_n, the probability that X_1 lies in $A \subset S_\alpha$ is greater than or equal to $\lambda_n \phi(A)$, where ϕ is the uniform probability measure on S_α. This mixing condition enables a version of Hoeffding's inequality (see theorem 2 in [37]): for any $\lambda_\mathscr{E} \in S_n'$ and $1 \leq k, l \leq N$, the transition matrix of equation (B.35) is such that, with $\hat{p}_n(k, l) = n^{-1} \sum_i 1\{Z_i = k, Z_{i-1} = l\}$,

$$P_{\alpha,n}\big(\hat{p}_n(k, l) - p_\alpha(k, l) \geq \delta\big) \leq \exp\Big(-\frac{\lambda_n^2(n\delta - 2\lambda_n^{-1})^2}{2n}\Big). \tag{B.36}$$

Now define for a given sequence $\delta_n > 0$ with $\delta_n \downarrow 0$ and all $n \geq 1, 1 \leq k, l \leq N$,

$$B_n = \{p_\alpha \in \Theta : \max_{k,l}\big|p_\alpha(k, l) - p_0(k, l)\big| < \epsilon - \delta_n\},$$

$$V_{k,l} = \{p_\alpha \in \Theta : \big|p_\alpha(k, l) - p_0(k, l)\big| \geq \epsilon\},$$

$$V_{+,k,l,n} = \{p_\alpha \in \Theta : p_\alpha(k, l) - p_0(k, l) \geq \epsilon + \delta_n\},$$

$$V_{-,k,l,n} = \{p_\alpha \in \Theta : p_\alpha(k, l) - p_0(k, l) \leq -\epsilon - \delta_n\}.$$

Note that if Π' is absolutely continuous with respect to the Lebesgue measure on S^{N^N}, then $\pi_{B,n} := \Pi(B \setminus B_n) \to 0$ and $\pi_{n,k,l} := \Pi(V_{k,l} \setminus (V_{+,k,l,n} \cup V_{-,k,l,n})) \to 0$.

If we define the test $\phi_{+,k,l,n}(Z^n) = 1\{\hat{p}_n(k, l) - p_0(k, l) \geq \epsilon\}$, then for any $p_\alpha \in B_n \cap S_n$,

$$P_{\alpha,n}\phi_{+,k,l,n}(Z^n) \leq P_{\alpha,n}\big(\hat{p}_n(k, l) - p_\alpha(k, l) \geq \delta_n\big)$$
$$\leq \exp\Big(-\frac{\lambda_n^2(n\delta_n - 2\lambda_n^{-1})^2}{2n}\Big).$$

If on the other hand, p_α lies in the intersection of $V_{+,n,k,l}$ with S_n, we find,

$$P_{\alpha,n}(1 - \phi_{+,n,k,l}(Z^n)) = P_{\alpha,n}\big(\hat{p}_n(k, l) - p_\alpha(k, l) < -\delta_n\big)$$
$$\leq \exp\Big(-\frac{\lambda_n^2(n\delta_n - 2\lambda_n^{-1})^2}{2n}\Big).$$

Choosing the sequences δ_n and λ_n such that $n\delta_n^2\lambda_n^2 \to \infty$, we also have $\lambda_n^{-1} = o(n\delta_n)$, so the exponent on the right is smaller than or equal to $-\frac{1}{8}n\lambda_n^2\delta_n^2$.

So if we define $\phi_n(Z^n) = \max_{k,l}\{\phi_{-,k,l,n}(Z^n), \phi_{+,k,l,n}(Z^n)\}$,

$$\int_B P_{\alpha,n}\phi_n \, d\Pi(p_\alpha) + \int_V Q_{\alpha,n}(1 - \phi_n) \, d\Pi(q_\alpha)$$

$$\leq \int_{B \cap S_n} P_{\alpha,n}\phi_n \, d\Pi(p_\alpha) + \int_{V \cap S_n} Q_{\alpha,n}(1 - \phi_n) \, d\Pi(q_\alpha) + \Pi(\Theta \setminus S_n)$$

$$\leq \int_B \sum_{k,l=1}^N P_{\alpha,n}(\phi_{-,k,l,n} + \phi_{+,k,l,n}) \, d\Pi(p_\alpha)$$

$$+ \sum_{k,l=1}^N \Big(\int_{V_{-,k,l}} Q_{\alpha,n}(1 - \phi_{-,k,l,n}) \, d\Pi(q_\alpha)$$

$$+ \int_{V_{+,k,l}} Q_{\alpha,n}(1 - \phi_{+,k,l,n}) \, d\Pi(q_\alpha) \Big)$$

$$+ \sum_{k,l=1}^N \Pi\big(V_{n,k,l} \setminus (V_{+,n,k,l} \cup V_{+,n,k,l})\big) + \Pi(\Theta \setminus S_n) + \Pi(B \setminus B_n)$$

$$\leq 2N^2 e^{-\frac{1}{8}n\lambda_n^2\delta_n^2} + \pi_{B,n} + \pi_{S,n} + \sum_{k,l=1}^N \pi_{n,k,l}.$$

So if we choose a prior Π' on S^{N^N} that is absolutely continuous with respect to Lebesgue measure, then (ϕ_n) defines a Bayesian test sequence for B versus V.

Because we have not imposed control over the rates at which the terms on the *r.h.s.* go to zero, remote contiguity at exponential rates is not good enough. Even if we would restrict supports of a sequence of priors such that $\pi_{B,N} = \pi_{S,n} = \pi_{n,k,l} = 0$, the first term on the *r.h.s.* is subexponential. To obtain a rate sharp enough, we note that the chain Z^n is positive recurrent, which guarantees that the dependence $p_\alpha \to dP_{\alpha,n}/dP_{0,n}$ is locally asymptotically normal [43, 38]. According to lemma B.13, this implies that local prior predictive distributions based on $n^{-1/2}$-neighbourhoods of p_0 in Θ are c_n-remotely contiguous to $P_{0,n}$ for *any* rate c_n, if the prior has full support. If we require that the prior density π' with respect to Lebesgue measure on S^{N^N} is continuous and strictly positive, then we see that there exists a constant $\pi > 0$ such that $\pi'(\lambda) \geq \pi$ for all $\lambda \in S^{N^N}$, so that for every $n^{-1/2}$-neighbourhood B_n of p_0, there exists a $K > 0$ such that $\Pi(B_n) \geq b_n := K n^{-N^N/2}$. Although local asymptotic normality guarantees remote contiguity at arbitrary rate, we still have to make sure that $c_n \to 0$ in lemma B.13, *i.e.* that $a_n = o(b_n)$. Then the remark directly after theorem 4.8 shows that condition *(ii)* of said theorem is satisfied.

The above leads to the following conclusion concerning goodness-of-fit testing *c.f.* (B.34).

PROPOSITION B.20 *Let X^n be a stationary, discrete time Markov chain on a measurable state space $(S, \mathscr{S})$. Choose a finite, measurable partition α of S such that the Markov chain Z^n is ergodic. Choose a prior Π' on S^{N^N} absolutely continuous with respect to Lebesgue measure with a continuous density that is everywhere strictly positive. Assume that,*

  (i) $n\lambda_n^2\delta_n^2/\log(n) \to \infty$,

(ii) $\Pi(B \setminus B_n), \Pi(\Theta \setminus S_n) = o(n^{-(N^N/2)})$,

(iii) $\max_{k,l} \Pi(V_{k,l} \setminus (V_{+,k,l,n} \cup V_{-,k,l,n})) = o(n^{-(N^N/2)})$.

Then for any choice of $\epsilon > 0$, the Bayes factors F_n are consistent for H_0 versus H_1.

To guarantee ergodicity of Z^n one may use an empirical device, *i.e.* we may use an independent, finite-length realization of the random walk X^n to find a partition α such that for all $1 \leq k, l \leq N$, we observe some m-step transition from l to k. An interesting generalisation concerns a hypothesized Markov transition kernel P_0 for the process X^n and partitions α_n (with projections p_{0,α_n} as in (B.33)), chosen such that α_{n+1} refines α_n for all $n \geq 1$. Bayes factors then test a sequence of pairs of hypotheses (B.34) centred on the p_{0,α_n}. The arguments leading to proposition B.20 do not require modification and the rate of growth N_n comes into the conditions of proposition B.20. □

Example B.19 demonstrates the enhancement of the role of the prior as intended by the remark that closes the subsection on the existence of Bayesian test sequences in section 2: where testing power is relatively weak, prior mass should be scarce to compensate and where testing power is strong, prior mass should be plentiful. A random walk for which mixing does not occur quickly enough does not give rise to (B.36) and alternatives for which separation decreases too fast lose testing power, so the difference sets of proposition B.20 are the hard-to-test parts of the parameter space and conditions *(ii)–(iii)* formulate how scarce prior mass in these parts has to be.

B.6   Finite sample spaces and the tailfree case

EXAMPLE B.21 Consider the situation where we observe an *i.i.d.* sample of random variables $X_1, X_2, \ldots$ taking values in a space $\mathscr{X}_N$ of *finite order* N. Writing $\mathscr{X}_N$ as the set of integers $\{1, \ldots, N\}$, we note that the space M of all probability measures P on $(\mathscr{X}_N, 2^{\mathscr{X}_N})$ with the total-variational metric $(P, Q) \mapsto \|P - Q\|$ is in isometric correspondence with the simplex,

$$S_N = \big\{ p = (p(1), \ldots, p(N)) : \min_k p(k) \geq 0, \Sigma_i \, p(i) = 1 \big\},$$

with the metric $(p, q) \mapsto \|p - q\| = \Sigma_k \, |p(k) - q(k)|$ it inherits from $\mathbb{R}^N$ with the L_1-norm, when $k \mapsto p(k)$ is the density of $P \in M$ with respect to the counting measure. We also define $M' = \{P \in M : P(\{k\}) > 0, 1 \leq k \leq N\} \subset M$ (and $R_N = \{p \in S_N : p(k) > 0, 1 \leq k \leq N\} \subset S_N$).

PROPOSITION B.22 *If the data is an i.i.d. sample of $\mathscr{X}_N$-valued random variables, then for any $n \geq 1$, any Borel prior $\Pi : \mathscr{G} \to [0,1]$ of full support on M, any $P_0 \in M$ and any ball B around P_0, there exists an $\epsilon' > 0$ such that,*

$$P_0^n \lhd e^{\frac{1}{2}n\epsilon^2} P_n^{\Pi|B}, \tag{B.37}$$

for all $0 < \epsilon < \epsilon'$.

PROOF By the inequality $\|P - Q\| \leq -P \log(dQ/dP)$, the ball B around P_0 contains all sets of the form $K(\epsilon) = \{P \in M' : -P_0 \log(dP/dP_0) < \epsilon\}$, for some $\epsilon' > 0$ and all $0 < \epsilon < \epsilon'$. Fix such an ϵ. Because the mapping $P \mapsto -P_0 \log(dP/dP_0)$ is continuous on M', there exists an open neighbourhood U of P_0 in M such that $U \cap M' \subset K(\epsilon)$. Since both M' and U are open and Π has full support, $\Pi(K(\epsilon)) \geq \Pi(U \cap M') > 0$. With the help of example 3.2, we see that for every $P \in K(\epsilon)$,

$$e^{\frac{1}{2}n\epsilon^2} \frac{dP^n}{dP_0^n}(X^n) \geq 1,$$

for large enough n, P_0-almost-surely. Fatou's lemma again confirms condition *(ii)* of lemma 3.3 is satisfied. Conclude that assertion (B.37) holds. $\qquad\square$

EXAMPLE B.23 We continue with the situation where we observe an *i.i.d.* sample of random variables $X_1, X_2, \ldots$ taking values in a space $\mathscr{X}_N$ of finite order N. For given $\delta > 0$, consider the hypotheses,

$$B = \{P \in M : \|P - P_0\| < \delta\}, \; V = \{Q \in M : \|Q - P_0\| > 2\delta\}.$$

Noting that M is compact (or with the help of the simplex representation S_N) one sees that entropy numbers of M are bounded, so the construction of example B.8 shows that uniform tests of exponential power e^{-nD} (for some $D > 0$) exist for B versus V. Application of proposition B.22 shows that the choice for an $0 < \epsilon < \epsilon'$ small enough, guarantees that $\Pi(V|X^n)$ goes to zero in P_0^n-probability. Conclude that the posterior resulting from a prior Π of full support on M is consistent in total variation.

EXAMPLE B.24 With general reference to Ferguson (1973) [28], one way to construct non-parametric priors concerns a refining sequence of finite, Borel measurable partitions of a Polish sample space, say $\mathscr{X} = \mathbb{R}$: to define a 'random distribution' P on $\mathscr{X}$, we specify for each such partition $\alpha = \{A_1, \ldots, A_N\}$, a Borel prior Π_α on S_N, identifying $(p_1, \ldots, p_N)$ with the 'random variables' $(P(A_1), \ldots, P(A_N))$. Kolmogorov existence of the stochastic process describing all $P(A)$ in a coupled way subjects these Π_α to consistency requirements expressing that if A_1, A_2 partition A, then $P(A_1) + P(A_2)$ must have the same distribution as $P(A)$. If the partitions refine appropriately, the resulting process describes a probability measure Π on the space of Borel probability measures on $\mathscr{X}$, *i.e.* a 'random distribution' on $\mathscr{X}$. Well-known examples of priors that can be constructed in this way are the Dirichlet process prior (for which a so-called base-measure μ supplies appropriate parameters for Dirichlet distributions Π_α, see [28]) and Polya Tree prior (for detailed explanations, see, for example, [35]).

A special class of priors constructed in this way are the so-called *tailfree* priors. The process prior associated with a family of Π_α like above is said to be *tailfree*, if for all α, β such that $\beta = \{B_1, \ldots, B_M\}$ refines $\alpha = \{A_1, \ldots, A_N\}$, the following holds: for all $1 \leq k \leq N$, $(P(B_{l_1}|A_k), \ldots, P(B_{l_{L(k)}}|A_k))$ (where the sets $B_{l_1}, \ldots, B_{l_{L(k)}} \in \beta$ partition A_k) is independent

of $(P(A_1), \ldots, P(A_N))$. Although somewhat technical, explicit control of the choice for the Π_α render the property quite feasible in examples.

Fix a finite, measurable partition $\alpha = \{A_1, \ldots, A_N\}$. For every $n \geq 1$, denote by $\sigma_{\alpha,n}$ the σ-algebra $\sigma(\alpha^n) \subset \mathscr{B}^n$, generated by products of the form $A_{i_1} \times \cdots \times A_{i_n} \subset \mathscr{X}^n$, with $1 \leq i_1, \ldots, i_n \leq N$. Identify $\mathscr{X}_N$ with the collection $\{e_1, \ldots, e_N\} \subset \mathbb{R}^N$ and define the projection $\varphi_\alpha : \mathscr{X} \mapsto \mathscr{X}_N$ by,

$$\varphi_\alpha(x) = \big(1\{x \in A_1\}, \ldots, 1\{x \in A_N\}\big).$$

We view $\mathscr{X}_N$ (respectively $\mathscr{X}_N^n$) as a probability space, with σ-algebra σ_N equal to the power set (respectively $\sigma_{N,n}$, the power set of $\mathscr{X}_N^n$) and probability measures denoted $P_\alpha : \sigma_N \to [0,1]$ that we identify with elements of S_N. Denoting the space of all Borel probability measures on $\mathscr{X}$ by $M^1(\mathscr{X})$, we also define $\varphi_{*\alpha} : M^1(\mathscr{X}) \to S_N$,

$$\varphi_{*\alpha}(P) = \big(P(A_1), \ldots, P(A_N)\big),$$

which maps P to its restriction to $\sigma_{\alpha,1}$, a probability measure on $\mathscr{X}_N$. Under the projection ϕ_α, any Borel-measurable random variable X taking values in $\mathscr{X}$ distributed $P \in M^1(\mathscr{X})$ is mapped to a random variable $Z_\alpha = \varphi_\alpha(X)$ that takes values in $\mathscr{X}_N$ (distributed $P_\alpha = \varphi_{*\alpha}(P)$). We also define $Z_\alpha^n = (\varphi_\alpha(X_1), \ldots, \varphi_\alpha(X_n))$, for all $n \geq 1$.

Let Π_α denote a Borel prior on S_N. The posterior on S_N is then a Borel measure denoted $\Pi_\alpha(\cdot | Z_\alpha^n)$, which satisfies, for all $A \in \sigma_{N,n}$ and any Borel set V in S_N,

$$\int_A \Pi_\alpha(V | Z_\alpha^n) \, dP_n^{\Pi_\alpha} = \int_V P_\alpha^n(A) \, d\Pi_\alpha(P_\alpha),$$

by definition of the posterior. In the model for the original $i.i.d.$ sample X^n, Bayes's rule takes the form, for all $A' \in \mathscr{B}_n$ and all Borel sets V' in $M^1(\mathscr{X})$,

$$\int_{A'} \Pi(V' | X^n) \, dP_n^\Pi = \int_{V'} P^n(A') \, d\Pi(P),$$

defining the posterior for P. Now specify that V' is the pre-image $\varphi_{*\alpha}^{-1}(V)$ of a Borel measurable V in S_N: as a consequence of tailfreeness, the data-dependence of the posterior for such a V', $X^n \mapsto \Pi(V' | X^n)$, is measurable with respect to $\sigma_{\alpha,n}$ (see Freedman (1965) [30] or Ghosh (2003) [35]). So there exists a function $g_n : \mathscr{X}_N^n \to [0,1]$ such that,

$$\Pi(V' | X^n = x^n) = g_n(\varphi_\alpha(x_1), \ldots, \varphi_\alpha(x_n)),$$

for P_n^Π-almost-all $x^n \in \mathscr{X}^n$. Then, for given $A' \in \sigma_{\alpha,n}$ (with corresponding $A \in \sigma_{N,n}$),

$$\int_{A'} \Pi(V' | X^n) \, dP_n^\Pi = \int P^n(1_{A'}(X^n) \, \Pi(V' | X^n)) \, d\Pi(P)$$
$$= \int P_\alpha^n\big(1_A(Z_\alpha^n) \, g_n(Z_\alpha^n)\big) \, d\Pi_\alpha(P_\alpha) = \int_A g_n(Z_\alpha^n) \, dP_n^{\Pi_\alpha},$$

while also,

$$\int_{V'} P^n(A') \, d\Pi(P) = \int_V P_\alpha^n(A) \, d\Pi_\alpha(P_\alpha).$$

This shows that $Z_\alpha^n \mapsto g_n(Z_\alpha^n)$ is a version of the posterior $\Pi_\alpha(\,\cdot\,|Z_\alpha^n)$. In other words, we can write $\Pi(V'|X^n) = \Pi_\alpha(V|\phi_\alpha(X^n)) = \Pi_\alpha(V|Z_\alpha^n)$, P_n^Π-almost-surely.

Denote the true distribution of a single observation from X^n by P_0. For any V' of the form $\varphi_{*\alpha}^{-1}(V)$ for some α and a neighbourhood V of $P_{0,\alpha} = \varphi_{*\alpha}(P_0)$ in S_N, the question whether $\Pi(V'|X^n)$ converges to one in P_0-probability reduces to the question whether $\Pi(V|Z_\alpha^n)$ converges to one in $P_{0,\alpha}$-probability. Remote contiguity then only has to hold as in example B.21.

Another way of saying this is to note directly that, because $X^n \mapsto \Pi(V'|X^n)$ is $\sigma_{\alpha,n}$-measurable, remote contiguity (as in definition 3.1) is to be imposed *only* for $\phi_n : \mathscr{X}^n \to [0,1]$ that are measurable with respect to $\sigma_{\alpha,n}$ (rather than $\mathscr{B}^n$) for every $n \geq 1$. That conclusion again reduces the remote contiguity requirement necessary for the consistency of the posterior for the parameter $(P(A_1), \dots P(A_N))$ to that of a finite sample space, as in example B.21. Full support of the prior Π_α then guarantees remote contiguity for exponential rates as required in condition *(ii)* of theorem 4.2. In the case of the Dirichlet process prior, full support of the base measure μ implies full support for all Π_α, if we restrict attention to partitions $\alpha = (A_1, \dots, A_N)$ such that $\mu(A_i) > 0$ for all $1 \leq i \leq N$. (Particularly, we require $P_0 \ll \mu$ for consistent estimation.)

Uniform tests of exponential power for weak neighbourhoods complete the proof that tailfree priors lead to weakly consistent posterior distributions: (norm) consistency of $\Pi_\alpha(\,\cdot\,|Z_\alpha^n)$ for all α guarantees (weak $\mathscr{T}_1$-)consistency of $\Pi(\,\cdot\,|X^n)$, in this proof based on remote contiguity and theorem 4.2.

B.7 Credible/confidence sets in metric spaces

When enlarging credible sets to confidence sets using a collection of subsets B as in definition 4.11, measurability of confidence sets is guaranteed if $B(\theta)$ is *open* in Θ for all $\theta \in \Theta$.

EXAMPLE B.25 Let $\mathscr{G}$ be the Borel σ-algebra for a uniform topology on Θ, like the weak and metric topologies of appendix A. Let W denote a symmetric entourage and, for every $\theta \in \Theta$, define $B(\theta) = \{\theta' \in \Theta : (\theta, \theta') \in W\}$, a neighbourhood of θ. Let D denote any credible set. A confidence set associated with D under B is any set C' such that the complement of D contains the W-enlargement of the complement of C'. Equivalently (by the symmetry of W), the W-enlargement of D does not meet the complement of C'. Then the minimal confidence set C associated with D is the W-enlargement of D. If the $B(\theta)$ are all open neighbourhoods (*e.g.* whenever W is a symmetric entourage from a fundamental system for the uniformity on Θ), the minimal confidence set associated with D is open. The most common examples include the Hellinger or total-variational metric uniformities, but weak topologies (like Prohorov's or $\mathscr{T}_n$-topologies) and polar topologies are uniform too. □

EXAMPLE B.26 To illustrate example B.25 with a customary situation, consider a parameter space Θ with parametrization $\theta \mapsto P_\theta^n$, to define a model for *i.i.d.* data $X^n = (X_1, \ldots, X_n) \sim P_{\theta_0}^n$, for some $\theta_0 \in \Theta$. Let $\mathscr{D}$ be the class of all pre-images of Hellinger balls, *i.e.* sets $D(\theta, \epsilon) \subset \Theta$ of the form,

$$D(\theta, \epsilon) = \left\{ \theta' \in \Theta : H(P_\theta, P_{\theta'}) < \epsilon \right\},$$

for any $\theta \in \Theta$ and $\epsilon > 0$. After choice of a Kullback-Leibler prior Π for θ and calculation of the posteriors, choose D_n equal to the pre-image $D(\hat{\theta}_n, \hat{\epsilon}_n)$ of a (*e.g.* the one with the smallest radius, if that exists) Hellinger ball with credible level $1 - o(a_n)$, $a_n = \exp(-n\alpha^2)$ for some $\alpha > 0$. Assume, now, that for some $0 < \epsilon < \alpha$, the W of example B.25 is the Hellinger entourage $W = \{(\theta, \theta') : H(P_\theta, P_{\theta'}) < \epsilon\}$. Since Kullback-Leibler neighbourhoods are contained in Hellinger balls, the sets $D(\hat{\theta}_n, \hat{\epsilon}_n + \epsilon)$ (associated with D_n under the entourage W), is a sequence of asymptotic confidence sets, provided the prior satisfies (2). If we make ϵ vary with n, neighbourhoods of the form B_n in example 4.6 are contained in Hellinger balls of radius ϵ_n, and in that case,

$$C_n(X^n) = D(\hat{\theta}_n, \hat{\epsilon}_n + \epsilon_n),$$

is a sequence of asymptotic confidence sets, provided that the prior satisfies (16). $\qquad \square$

C   Proofs

In this section of the appendix, proofs from the main text are collected.

PROOF (theorem 1.1)
The argument (see, *e.g.*, Doob (1949) [26] or Ghosh and Ramamoorthi (2003) [35]) relies on martingale convergence and a demonstration of the existence of a measurable $f : \mathscr{X}^\infty \to \mathscr{P}$ such that $f(X_1, X_2, \ldots) = P$, P^∞-almost-surely for all $P \in \mathscr{P}$ (see also propositions 1 and 2 of section 17.7 in [51]). $\qquad \square$

PROOF (proposition 2.2)
Due to Bayes's Rule (A.22) and monotone convergence,

$$\int_B P_\theta(1 - \phi)\, \Pi(V|X)\, d\Pi(\theta)$$
$$\leq \int (1 - \phi)\, \Pi(V|X)\, dP^\Pi = \int_V P_\theta(1 - \phi)\, d\Pi(\theta).$$

Inequality (4) follows from the fact that $\Pi(V|X) \leq 1$. $\qquad \square$

PROOF (theorem 2.4)
Condition *(i)* implies *(ii)* by dominated convergence. Assume *(ii)* and note that by lemma 2.2,

$$\int P_{\theta,n}\Pi(V|X^n)\, d\Pi(\theta|B) \to 0.$$

Assuming that the observations X^n are coupled and can be thought of as projections of a random variable $X \in \mathscr{X}^\infty$ with distribution P_θ, martingale convergence in $L^1(\mathscr{X}^\infty \times \Theta)$ (relative to the probability measure Π^* defined by $\Pi^*(A \times B) = \int_B P_\theta(A)\, d\Pi(\theta)$ for measurable $A \subset \mathscr{X}^\infty$ and $B \subset \Theta$), shows there is a measurable $g : \mathscr{X}^\infty \to [0,1]$ such that,

$$\int P_\theta \big| \Pi(V|X^n) - g(X) \big|\, d\Pi(\theta|B) \to 0.$$

So $\int P_\theta g(X)\, d\Pi(\theta|B) = 0$, implying that $g = 0$, P_θ-almost-surely for Π-almost-all $\theta \in B$. Using martingale convergence again (now in $L^\infty(\mathscr{X}^\infty \times \Theta)$), conclude $\Pi(V|X^n) \to 0$, P_θ-almost-surely for Π-almost-all $\theta \in B$, from which *(iii)* follows. Choose $\phi(X^n) = \Pi(V|X^n)$ to conclude that *(i)* follows from *(iii)*. $\qquad \square$

PROOF (proposition 2.5)
Apply [51], section 17.1, proposition 1 with the indicator for V. See also [11]. $\qquad \square$

PROOF (lemma 3.3)
Assume *(i)*. Let $\phi_n : \mathscr{X}_n \to [0,1]$ be given and assume that $P_n\phi_n = o(a_n)$. By Markov's inequality, for every $\epsilon > 0$, $P_n(a_n^{-1}\phi_n > \epsilon) = o(1)$. By assumption, it now follows that $\phi_n \xrightarrow{Q_n} 0$. Because $0 \le \phi_n \le 1$ the latter conclusion is equivalent to $Q_n\phi_n = o(1)$.

Assume *(iv)*. Let $\epsilon > 0$ and $\phi_n : \mathscr{X}_n \to [0,1]$ be given. There exist $c > 0$ and $N \ge 1$ such that for all $n \ge N$,

$$Q_n\phi_n < c\,a_n^{-1}P_n\phi_n + \frac{\epsilon}{2}.$$

If we assume that $P_n\phi_n = o(a_n)$ then there is a $N' \ge N$ such that $c\,a_n^{-1}P_n\phi_n < \epsilon/2$ for all $n \ge N'$. Consequently, for every $\epsilon > 0$, there exists an $N' \ge 1$ such that $Q_n\phi_n < \epsilon$ for all $n \ge N'$.

To show that *(ii)* $\Rightarrow$ *(iv)*, let $\mu_n = P_n + Q_n$ and denote μ_n-densities for P_n, Q_n by $p_n, q_n : \mathscr{X}_n \to \mathbb{R}$. Then, for any $n \ge 1$, $c > 0$,

$$\Big\| Q_n - Q_n \wedge c\,a_n^{-1}P_n \Big\| = \sup_{A \in \mathscr{B}_n} \Big(\int_A q_n\, d\mu_n - \int_A q_n\, d\mu_n \wedge \int_A c\,a_n^{-1}p_n\, d\mu_n \Big)$$

$$\le \sup_{A \in \mathscr{B}_n} \int_A (q_n - q_n \wedge c\,a_n^{-1}p_n)\, d\mu_n \qquad\qquad \text{(C.38)}$$

$$= \int 1\{q_n > c\,a_n^{-1}p_n\}\,(q_n - c\,a_n^{-1}p_n)\, d\mu_n.$$

Note that the right-hand side of (C.38) is bounded above by $Q_n(dP_n/dQ_n < c^{-1}a_n)$.

To show that *(iii)* $\Rightarrow$ *(iv)*, it is noted that, for all $c > 0$ and $n \ge 1$,

$$0 \le \int c\,a_n^{-1}P_n(q_n > c\,a_n^{-1}p_n) \le Q_n(q_n > c\,a_n^{-1}p_n) \le 1,$$

so (C.38) goes to zero if $\liminf_{n\to\infty} c\,a_n^{-1}P_n(dQ_n/dP_n > c\,a_n^{-1}) = 1$.

To prove that *(v)* $\Leftrightarrow$ *(ii)*, note that Prohorov's theorem says that weak convergence of a subsequence within any subsequence of $a_n(dP_n/dQ_n)^{-1}$ under Q_n (see appendix A, *notation*

48

and conventions) is equivalent to the asymptotic tightness of $(a_n(dP_n/dQ_n)^{-1} : n \geq 1)$ under Q_n, *i.e.* for every $\epsilon > 0$ there exists an $M > 0$ such that $Q_n(a_n(dP_n/dQ_n)^{-1} > M) < \epsilon$ for all $n \geq 1$. This is equivalent to *(ii)*. $\qquad\square$

PROOF (proposition 3.6)

For every $\epsilon > 0$, there exists a constant $\delta > 0$ such that,

$$P_{\theta_0,n}\left(a_n \Big(\frac{dP_{\theta,n}}{dP_{\theta_0,n}} \Big)^{-1}(X^n) > \frac{1}{\delta} \right) < \epsilon,$$

for all $\theta \in B$, $n \geq 1$. For this choice of δ, condition *(ii)* of lemma 3.3 is satisfied for all $\theta \in B$ simultaneously, and *c.f.* the proof of said lemma, for given $\epsilon > 0$, there exists a $c > 0$ such that,

$$\| P_{\theta_0,n} - P_{\theta_0,n} \wedge c\, a_n^{-1} P_{\theta,n} \| < \epsilon, \tag{C.39}$$

for all $\theta \in B$, $n \geq 1$. Now note that for any $A \in \mathscr{B}_n$,

$$\begin{aligned} 0 \leq{}& P_{\theta_0,n}(A) - P_{\theta_0,n}(A) \wedge c\, a_n^{-1} P_n^{\Pi|B}(A) \\ \leq{}& \int \big(P_{\theta_0,n}(A) - P_{\theta_0,n}(A) \wedge c\, a_n^{-1} P_{\theta,n}(A) \big)\, d\Pi(\theta|B). \end{aligned}$$

Taking the supremum with respect to A, we find the following inequality in terms of total variational norms,

$$\big\| P_{\theta_0,n} - P_{\theta_0,n} \wedge c\, a_n^{-1} P_n^{\Pi|B} \big\| \leq \int \big\| P_{\theta_0,n} - P_{\theta_0,n} \wedge c\, a_n^{-1} P_{\theta,n} \big\|\, d\Pi(\theta|B).$$

Since the total-variational norm is bounded and $\Pi(\cdot|B)$ is a probability measure, Fatou's lemma says that,

$$\begin{aligned} \limsup_{n\to\infty} &\big\| P_{\theta_0,n} - P_{\theta_0,n} \wedge c\, a_n^{-1} P_n^{\Pi|B} \big\| \\ &\leq \int \limsup_{n\to\infty} \big\| P_{\theta_0,n} - P_{\theta_0,n} \wedge c\, a_n^{-1} P_{\theta,n} \big\|\, d\Pi(\theta|B), \end{aligned}$$

and the *r.h.s.* equals zero *c.f.* (C.39). According to condition *(iv)* of lemma 3.3 this implies the assertion. $\qquad\square$

PROOF (lemma 3.7)

Fix $n \geq 1$. Because $B_n \subset C_n$, for every $A \in \mathscr{B}_n$, we have,

$$\int_{B_n} P_{\theta,n}(A)\, d\Pi(\theta) \leq \int_{C_n} P_{\theta,n}(A)\, d\Pi(\theta),$$

so $P_n^{\Pi_n|B_n}(A) \leq \Pi_n(C_n)/\Pi_n(B_n)\, P_n^{\Pi_n|C_n}(A)$. So if for some sequence $\phi_n : \mathscr{X}_n \to [0,1]$, we have $P_n^{\Pi_n|C_n}\phi_n(X^n) = o(\Pi_n(B_n)/\Pi_n(C_n))$, then the $P_n^{\Pi_n|B_n}$-expectations of $\phi_n(X^n)$ are $o(1)$, proving the first claim. If $P_n^{\Pi_n|C_n}\phi_n(X^n) = o(a_n\Pi_n(B_n)/\Pi_n(C_n))$, then $P_n^{\Pi_n|B_n}\phi_n(X^n) = o(a_n)$ and, hence, $P_n\phi_n(X^n) = o(1)$. $\qquad\square$

PROOF (theorem 4.2)

Choose $B_n = B$, $V_n = V$ and use proposition 2.3 to see that $P_n^{\Pi|B}\Pi(V|X^n)$ is upper bounded

by $\Pi(B)^{-1}$ times the *l.h.s.* of (14) and, hence, is of order $o(a_n)$. Condition *(ii)* then implies that $P_{\theta_0,n}\Pi(V|X^n) = o(1)$, which is equivalent to $\Pi(V|X^n) \xrightarrow{P_{\theta_0,n}} 0$ since $0 \leq \Pi(V|X^n) \leq 1$, $P_{\theta_0,n}$-almost-surely, for all $n \geq 1$. $\qquad\square$

PROOF (corollary 4.3)

A prior Π satisfying condition *(ii)* guarantees that $P_0^n \ll P_n^\Pi$ for all $n \geq 1$, *c.f.* the remark preceding proposition A.7. Choose ϵ such that $\epsilon^2 < D$. Recall that for every $P \in B(\epsilon)$, the exponential lower bound (10) for likelihood ratios of dP^n/dP_0^n exists. Hence the limes inferior of $\exp(\frac{1}{2}n\epsilon^2)(dP^n/dP_0^n)(X^n)$ is greater than or equal to one with P_0^∞-probability one. Then, with the use of Fatou's lemma and the assumption that $\Pi(B(\epsilon)) > 0$,

$$\liminf_{n\to\infty} \frac{e^{nD}}{\Pi(B)} \int_B \frac{dP_\theta^n}{dP_{\theta_0}^n}(X^n)\, d\Pi(\theta) \geq 1,$$

with $P_{\theta_0}^\infty$-probability one, showing that sufficient condition *(ii)* of lemma 3.3 holds. Conclude that,

$$P_0^n \vartriangleleft e^{nD}\, P_n^{\Pi|B},$$

and use theorem 4.2 to see that $\Pi(U|X^n) \xrightarrow{P_{\theta_0,n}} 1$. $\qquad\square$

PROOF (theorem 4.4)

Proposition 2.3 says that $P_n^{\Pi_n|B_n}\Pi(V_n|X^n)$ is of order $o(b_n^{-1}a_n)$. Condition *(iii)* then implies that $P_{\theta_0,n}\Pi(V_n|X^n) = o(1)$, which is equivalent to $\Pi(V_n|X^n) \xrightarrow{P_{\theta_0,n}} 0$ since $0 \leq \Pi(V_n|X^n) \leq 1$, $P_{\theta_0,n}$-almost-surely for all $n \geq 1$. $\qquad\square$

PROOF (theorem 4.12)

Fix $n \geq 1$ and let D_n denote a credible set of level $1 - o(a_n)$, defined for all $x \in F_n \subset \mathscr{X}_n$ such that $P_n^{\Pi_n}(F_n) = 1$. For any $x \in F_n$, let $C_n(x)$ denote a confidence set associated with $D_n(x)$ under B. Due to definition 4.11, $\theta_0 \in \Theta \setminus C_n(x)$ implies that $B_n(\theta_0) \cap D_n(x) = \varnothing$. Hence the posterior mass of $B(\theta_0)$ satisfies $\Pi(B_n(\theta_0)|x) = o(a_n)$. Consequently, the function $x \mapsto 1\{\theta_0 \in \Theta \setminus C_n(x)\}\,\Pi(B(\theta_0)|x)$ is $o(a_n)$ for all $x \in F_n$. Integrating with respect to the n-th prior predictive distribution and dividing by the prior mass of $B_n(\theta_0)$, one obtains,

$$\frac{1}{\Pi_n(B_n(\theta_0))} \int 1\{\theta_0 \in \Theta \setminus C_n\}\,\Pi(B_n(\theta_0)|X^n)\, dP_n^{\Pi_n} \leq \frac{a_n}{b_n}.$$

Applying Bayes's rule in the form (A.22), we see that,

$$P_n^{\Pi_n|B_n(\theta_0)}\big(\theta_0 \in \Theta \setminus C_n(X^n)\big) = \int P_{\theta,n}\big(\theta_0 \in \Theta \setminus C_n(X^n)\big)\, d\Pi_n(\theta|B_n) \leq \frac{a_n}{b_n}.$$

By the definition of remote contiguity, this implies asymptotic coverage *c.f.* (18). $\qquad\square$

PROOF (corollary 4.13)

Define $a_n = \exp(-C'n\epsilon_n^2)$, $b_n = \exp(-Cn\epsilon_n^2)$, so that the D_n are credible sets of level $1 - o(a_n)$, the sets B_n of example 4.6 satisfy condition *(i)* of theorem 4.12 and $b_n a_n^{-1} = \exp(cn\epsilon_n^2)$ for some $c > 0$. By (17), we see that condition *(ii)* of theorem 4.12 is satisfied. The assertion now follows. $\qquad\square$

References

[1] E. Abbe, A. Bandeira and G. Hall, Exact recovery in the stochastic block model, arXiv:1405.3267 [cs.SI]

[2] A. Barron, Discussion on Diaconis and Freedman: the consistency of Bayes estimates. Ann. Statist. **14** (1986), 26–30.

[3] A. Barron, The exponential convergence of posterior probabilities with implications for Bayes estimators of density functions, Technical Report **7** (1988), Dept. Statistics, Univ. Illinois.

[4] A. Barron, M. Schervish and L. Wasserman, The consistency of distributions in nonparametric problems, Ann. Statist. **27** (1999), 536-561.

[5] M. Bayarri and J. Berger, The Interplay of Bayesian and Frequentist Analysis, Statist. Sci. **19** (2004), 58–80.

[6] P. Bickel and P. Chen, A nonparametric view of network models and Newman-Girvan and other modularities, Proc. Natl. Acad. Sci. USA **106** (2009), 21068-21073.

[7] L. Birgé, Approximation dans les espaces métriques et théorie de l'estimation, Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete **65** (1983), 181–238.

[8] L. Birgé, Sur un théorème de minimax et son application aux tests, Probability and Mathematical Statistics **3** (1984), 259–282.

[9] L. Birgé, and P. Massart, From model selection to adative estimation, Festschrift for Lucien Le Cam (eds. D. Pollard, E. Torgesen, G. Yang), Springer, New York (1997), 55–87.

[10] L. Birgé, and P. Massart, Gaussian model selection, J. Eur. Math. Soc. **3** (2001), 203–268.

[11] L. Breiman, L. Le Cam and L. Schwartz, Consistent Estimates and Zero-One Sets, Ann. Math. Statist. **35** (1964), 157–161.

[12] P. Bühlman and S. van de Geer, Statistics for High-Dimensional Data, Springer Verlag, New York (2011).

[13] T. Cai, M. Low and Y. Xia, Adaptive confidence interval for regression functions under shape constraints, Ann. Statist. **41** (2013), 722–750.

[14] C. Carvalho, N. Polson, and J. Scott, The horseshoe estimator for sparse signals, Biometrika **97** (2010), 465–480.

[15] I. Castillo and A. van der Vaart, Needles and straw in a haystack: posterior concentration for possibly sparse sequences, Ann. Statist. **40** (2012), 2069–2101.

[16] D. CHOI, P. WOLFE and E. AIROLDI, Stochastic block models with a growing number of classes, Biometrika **99** (2012), 273–284.

[17] D. COX, An analysis of Bayesian inference for non-parametric regression, Ann. Statist. **21** (1993), 903–924.

[18] A. DECELLE, F. KRZAKALA, C. MOORE and L. ZDEBOROVÁ, Inference and Phase Transitions in the Detection of Modules in Sparse Networks, Phys. Rev. Lett. **107** (2011), 065701

[19] P. DE BLASI, A. LIJOI, and I. PRÜNSTER, An asymptotic analysis of a class of discrete nonparametric priors, Statist. Sinica **23** (2013), 1299–1322.

[20] P. DIACONIS and D. FREEDMAN, On the Consistency of Bayes Estimates, Ann. Statist. **14** (1986), 1–26.

[21] P. DIACONIS and D. FREEDMAN, On Inconsistent Bayes Estimates of Location, Ann. Statist. **14** (1986), 68–87.

[22] P. DIACONIS and D. FREEDMAN, Nonparametric Binary Regression: A Bayesian Approach, Ann. Statist. **21** (1993), 2108–2137.

[23] P. DIACONIS and D. FREEDMAN, Consistency of Bayes estimates for nonparameteric regression: normal theory, Bernoulli **4** (1998), 411–444.

[24] D. DONOHO, I. JOHNSTONE, J. HOCH and A. STERN, Maximum entropy and the nearly black object, J. Roy. Statist. Soc. **B54** (1992), 41–81.

[25] D. DONOHO, and I. JOHNSTONE, Minimax risk over ℓ_p-balls for ℓ_q-error, Probab. Theory Related Fields, **99** (1994), 277-303.

[26] J. DOOB, Application of the theory of martingales, Colloque international Centre nat. Rech. Sci., Paris (1949), 22–28.

[27] P. EICHELSBACHER and A. GANESH, Bayesian inference for Markov chains, J. Appl. Probab. **39** (2002), 91–99.

[28] T. FERGUSON, A Bayesian Analysis of Some Nonparametric Problems, Ann. Statist. **1** (1973), 209–230.

[29] D. FREEDMAN, On the asymptotic behavior of Bayes estimates in the discrete case I, Ann. Math. Statist. **34** (1963), 1386–1403.

[30] D. FREEDMAN, On the asymptotic behavior of Bayes estimates in the discrete case II, Ann. Math. Statist. **36** (1965), 454–456.

[31] D. FREEDMAN, and P. DIACONIS, On Inconsistent Bayes Estimates in the Discrete Case, Ann. Statist. **11** (1983), 1109–1118.

[32] D. Freedman, On the Bernstein-von Mises theorem with infinite dimensional parameters, Ann. Statist. **27** (1999), 1119–1140.

[33] S. Ghosal, J. Ghosh, and R. Ramamoorthi, Consistency issues in Bayesian nonparametrics, In *Asymptotics, Nonparametrics and Time Series: A Tribute to Madan Lal Puri (Subir Ghosh, ed.)* Dekker, New York (1999), 639-667.

[34] S. Ghosal, J. Ghosh and A. van der Vaart, Convergence rates of posterior distributions, Ann. Statist. **28** (2000), 500–531.

[35] J. Ghosh and R. Ramamoorthi, Bayesian nonparametrics, Springer Verlag, New York (2003).

[36] S. Ghosal and Y.-Q. Tang Bayesian Consistency for Markov Processes, Sankhya **68** (2006), 227–239.

[37] P. Glynn and D. Ormoneit, Hoeffding's inequality for uniformly ergodic Markov chains, Statist. Probab. Lett. **56** (2002), 143–146.

[38] E. Gobet, LAN property for ergodic diffusions with discrete observations, Ann. I. H. Poincaré PR **38** (200), 711–737.

[39] P. Greenwood and A. Shiryaev, Contiguity and the statistical invariance principle, Gordon and Breach, New York (1985).

[40] J. Hajék and Z. Šidák, Theory of rank tests, Academic Press, New York (1967).

[41] K. Hayashi, T. Konishi and T. Kawamoto, A tractable fully Bayesian method for the stochastic block model, `arxiv:1602.02256 [cs.LG]`

[42] N. Hengartner, P. Stark, Finite-sample confidence envelopes for shape-restricted densities, Ann. Statist. **23** (1995), 525–550.

[43] R. Höpfner, J. Jacod and L. Ladelli, Local asymptotic normality and mixed normality for Markov statistical models, Probab. Th. Rel. Fields (1990) 86: 105.

[44] I. Johnstone and B. Silverman, Needles and straw in haystacks: empirical Bayes setimates of possibly sparse sequences, Ann. Statist. **32** (2004), 1594–1649.

[45] B. Kleijn and A. van der Vaart, The Bernstein-Von-Mises theorem under misspecification, Electron. J. Statist. **6** (2012), 354–381.

[46] B. Kleijn and Y.-Y. Zhao, Criteria for posterior consistency, arxiv:1308.1263 [MATH.ST].

[47] L. Le Cam and L. Schwartz, A necessary and sufficient condition for the existence of consistent estimates, Ann. Math. Statist. **31** (1960), 140–150.

[48] L. Le Cam, Locally asymptotically normal families of distributions, University of California Publications in Statistics **3** (1960), 37-98.

[49] L. Le Cam, Convergence of estimates under dimensionality restrictions, Ann. Statist. **1** (1973), 38–55.

[50] L. Le Cam, An inequality concerning Bayes estimates, University of California, Berkeley (197X), unpublished.

[51] L. Le Cam, Asymptotic methods in statistical decision theory, Springer, New York (1986).

[52] L. Le Cam and G. Yang, On the preservation of local asymptotic normality under information loss, Ann. Statist. **16** (1988), 483–520.

[53] L. Le Cam and G. Yang, Asymptotics in Statistics: some basic concepts, Springer, New York (1990).

[54] A. Lijoi, I. Prünster and S. Walker, Extending Doob's consistency theorem to nonparametric densities, Bernoulli **10** (2004), 651–663.

[55] M. Low, On nonparametric confidence intervals, Ann. Statist. **25** (1997), 2547–2554.

[56] S. Meyn and R. Tweedie, Markov Chains and Stochastic Stability, Cambridge University Press, New York (2009).

[57] T. Mitchell, and J. Beauchamp, Bayesian Variable Selection in Linear Regression, J. Amer. Statist. Assoc. **83** (1988), 1023–1032.

[58] E. Mossel, J. Neeman and A. Sly, Consistency thresholds for the planted bisection model, Electron. J. Probab. **21** (2016), 1–24.

[59] F. Papangelou, Large-deviations and the Bayesian estimation of higher-order Markov chains, J. Appl. Probab. **33** (1996), 18–27.

[60] G. Roussas, Contiguity of probability measures: some applications in statistics, Cambridge Tracts in Mathematics and Mathematical Physics **63** (1972), Cambridge University Press, London-New York.

[61] L. Schwartz, Consistency of Bayes procedures, PhD. thesis, Dept. of Statistics, University of California, Berkeley (1961).

[62] L. Schwartz, On Bayes procedures, Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete **4** (1965), 10–26.

[63] X. Shen and L. Wasserman, Rates of convergence of posterior distributions, Ann. Statist. **29** (2001), 687–714.

[64] K. NOWICKI and T. SNIJDERS, Estimation and Prediction for Stochastic Blockstructures, J. Amer. Statist. Assoc. **96** (2001), 1077–1087.

[65] H. STRASSER, Mathematical theory of statistics, de Gruyter, Berlin, 1985.

[66] C.STRELIOFF, J. CRUTCHFIELD and A. HUBLER, Inferring Markov chains: Bayesian estimation, model comparison, entropy rate, and out-of-class modeling, arXiv:math/0703715 [math.ST] (2007).

[67] S. SUWAN, D. LEE, R. TANG, D. SUSSMAN, M. TANG and C. PRIEBE Empirical Bayes estimation for the stochastic blockmodel, Electron. J. Statist. **10** (2016), 761–782.

[68] B. SZABÓ, A. VAN DER VAART and J. VAN ZANTEN, Frequentist coverage of adaptive nonparametric Bayesian credible sets, Ann. Statist. **43** (2015), 1391–1428.

[69] J. TAYLOR and R. TIBSHIRANI, Statistical learning and selective inference, Proc. Natl. Acad. Sc. **112** (2016), 7629–7634.

[70] S. WALKER, New approaches to Bayesian consistency, Ann. Statist. **32** (2004), 2028–2043.

[71] S. WALKER, A. LIJOI and I. PRÜNSTER, Data tracking and the understanding of Bayesian consistency, Biometrika **92** (2005), 765–778.

[72] S. WALKER, A. LIJOI and I. PRÜNSTER, On rates of convergence for posterior distributions in infinite-dimensional models, Ann. Statist. **35** (2007), 738–746.

[73] L. WASSERMAN, Bayesian Model Selection and Model Averaging, J. Math. Psychology **44** (2000), 92-107.

[74] G. YANG, A conversation with Lucien Le Cam, Statist. Sc. **14** (1999), 223–241.

[75] C.-H. ZHANG and J. HUANG, The sparsity and bias of the LASSO selection in high-dimensional linear regression, Ann. Statist. **36** (2008), 1567–1594.