Bayesian Analysis (2018) **13**, Number 2, pp. 461–484

# Testing Un-Separated Hypotheses by Estimating a Distance

Jean-Bernard Salomond*

**Abstract.** In this paper we propose a Bayesian answer to testing problems when the hypotheses are not well separated. The idea of the method is to study the posterior distribution of a discrepancy measure between the parameter and the model we want to test for. This is shown to be equivalent to a modification of the testing loss. An advantage of this approach is that it can easily be adapted to complex hypotheses testing which are in general difficult to test for. Asymptotic properties of the test can be derived from the asymptotic behaviour of the posterior distribution of the discrepancy measure, and gives insight on possible calibrations. In addition one can derive separation rates for testing, which ensure the asymptotic frequentist optimality of our procedures.

**Keywords:** hypothesis testing, Bayesian inference, asymptotic properties of tests, nonparametric inference, goodness-of-fit, monotonicity.

## 1 Introduction

Bayesian hypothesis testing, although widely studied in the literature, is still subject to controversy (see Jeffreys, 1939; Bernardo, 1980; Berger and Sellke, 1987; Gelman, 2008, to name a few). In particular, a lot of efforts have been puted on reconciling Bayesian and frequentist testing procedures as in Berger and Sellke (1987), Berger et al. (1997) or Berger and Delampady (1987). In this paper, we focus on the specific case of two-hypotheses testing although we believe that the ideas developed here are more general; more precisely, we consider testing problems of the form:

$$H_0 : \theta \in \mathcal{M}_0, \text{ versus } H_1 : \theta \in \mathcal{M}_1, \tag{1}$$

where $\mathcal{M}_0$ and $\mathcal{M}_1$ are not well separated, i.e. $\bar{\mathcal{M}}_0 \cap \bar{\mathcal{M}}_1 \neq \emptyset$ where $\bar{\mathcal{F}}$ stands for the closure of $\mathcal{F}$. When considering prediction, it is now well known that standard Bayesian methods such as Bayesian Information Criterion (BIC) have a tendency to favour the simpler model, even when the more complex one gives better predictions, as shown in Erven et al. (2012). This phenomenon also occurs in a testing or model selection setting when hypotheses are nested, and induces a loss of power for the Bayesian test near the null. In our view, one reason for this lack of efficiency of standard Bayesian testing approaches, such as the Bayes Factor or the comparison of posterior probabilities, comes from the fact that parameters that are close to the boundary between both hypotheses can be approximated from both sides. Thus, depending on the prior distribution on both the null and the alternative, some inconsistency may occur. This phenomenon is shown on some examples in Section 2. This loss of power of Bayesian testing procedures,

*Université Paris-Est, Laboratoire d'Analyse et de Mathématiques Appliquées (UMR 8050), UPEM, UPEC, CNRS, F-94010, Créteil, France, jean-bernard.salomond@u-pec.fr

 https://doi.org/10.1214/17-BA1059

induced by the prior is troublesome as it is difficult to control for and strongly depends on the prior distribution. Finding good prior distributions for testing has been a subject of high interest in the recent years. In particular, Johnson and Rossell (2010) (or the actualised version of their ideas developed in Rossell and Telesca, 2017) consider a similar case of un-separated hypotheses. Their idea is to enforce separation through the prior distribution using *non-local* priors. As exposed in Rousseau and Robert (2010), this approach can be viewed as a modification of the loss used for testing. It appears on simple examples studied in Section 2 that imposing such a penalty can make it more difficult to detect parameters near the boundary between hypotheses (see Section 2.1 for instance). The problem of finding a good prior distribution for testing has also been tackled by Johnson (2013), where the author introduced uniformly most powerful Bayesian test. The author proposes to calibrate the method by maximizing the probability that the Bayes Factor exceeds a certain threshold under the alternative. However, the proposed method seems difficult to extend outside exponential models. In this paper we propose a novel approach to the problem of testing un-separated hypotheses, based on the evaluation of a *discrepancy* between the parameter $\theta$ and the hypothesis at hand. A great advantage of the approach is that it is in general easy to use in practice and it generalizes directly to nonparametric hypotheses testing. Let $D(\theta, \mathcal{M}_0)$ be a *discrepancy measure* between $\theta$ and $\mathcal{M}_0$. Following the frequentist approach to testing, our idea is to associate $\theta$ to $\mathcal{M}_0$ if $D(\theta, \mathcal{M}_0)$ is bellow a certain threshold $\tau$. This idea of choosing the model closer to the parameter for a certain metric is quite general and we believe that it could be applied in a wide variety of settings. In this paper, we might only focus on the simpler problem of two hypotheses testing.

Although not aiming at the same problem, this approach is similar to the idea of approximating precise hypotheses by point null hypotheses as studied in Berger and Delampady (1987), which can be re-interpreted as a use of non-local prior as argued in Johnson and Rossell (2010). This approximation of hypotheses where latter studied in Verdinelli and Wasserman (1998) and Rousseau (2007). More specifically, in the latter the author proposes a generalization of the $0-1$ loss function from which a Bayesian test is derived, and which induces a separation of the hypotheses. Following Rousseau (2007), we consider the following loss function

$$L(\theta, \delta) = \begin{cases} 0 & \text{if } \delta = \mathbb{I}_{D(\theta,\mathcal{M}_0)\leq\tau} \\ \gamma_0 & \text{if } \delta = 1 \text{ and } D(\theta, \mathcal{M}_0) \leq \tau \\ \gamma_1 & \text{if } \delta = 0 \text{ and } D(\theta, \mathcal{M}_0) > \tau \end{cases}, \tag{2}$$

where the parameters $\gamma_0$ and $\gamma_1$ have the same interpretation as for the standard weighted $0-1$ loss (see Robert, 2007) in terms of price of misclassification error. A default choice is to take $\gamma_0 = \gamma_1$. This modification of the loss function can also be viewed as a relaxation of the hypotheses

$$H_0^a : D(\theta, \mathcal{M}_0) \leq \tau, \;\; H_1^a : D(\theta, \mathcal{M}_0) > \tau. \tag{3}$$

For a fixed threshold $\tau$, the same idea was applied in Dunson and Peddada (2008) and Wang and Dunson (2011) for testing equality in distribution against stochastic ordering. From a decision theoretic point of view, this loss is relevant since it indicates that we do

not pay for misclassified parameters that lie in a region in which we cannot differentiate the null and the alternative. In addition, as argued in Berger and Delampady (1987), one is in general not so interested in knowing if $\theta$ belongs to $\mathcal{M}_0$ but rather if $\theta \in \mathcal{M}_0$ is reasonable approximation. From a more practical point of view, this approach gives a method for constructing Bayesian tests that separate well the hypotheses in a wide variety of contexts, including complex alternatives such as nonparametric models for example. Deriving the Bayesian answer to (3) can also lead to simpler procedures. The Bayesian estimate associated with a prior $\Pi$ on the parameter set $\mathcal{M} = \mathcal{M}_0 \cup \mathcal{M}_1$, the loss (2) and data $Y^n$, is given by

$$\delta_n^\pi(\tau) = \begin{cases} 0 & \text{if } \Pi\{D(\theta, \mathcal{M}_0) \leq \tau | Y^n\} \geq \frac{\gamma_0}{\gamma_0+\gamma_1} \\ 1 & \text{otherwise} \end{cases}, \tag{4}$$

where $\Pi(\cdot|Y^n)$ denote the posterior measure of the parameter $\theta$ given the observations $Y^n$. From this last equation, we see that the behaviour of a test based on our modified $0-1$ loss is driven by the behaviour of $D(\theta, \mathcal{M}_0)$. This will prove particularly useful when testing complex or nonparametric versus complex or nonparametric hypothesis which is known to be a difficult case to handle and has not received much attention in the Bayesian literature. In addition even for simpler model, the behaviour of $D(\theta, \mathcal{M}_0)$ may also be easy to study for a wide variety of priors, as shown in Section 3.1 for instance. From this formulation, we see that prior distributions that induce a good behaviour for $D(\theta, \mathcal{M}_0)$ in terms of concentration properties will also be good candidates for testing with this approach. Note that such priors may differ from the one that leads to good properties for estimating $\theta$, as shown for example in Section 3.2. Note also that to compute the Bayesian test with formulation (4), we only have to sample under the posterior. This thus gives leads to tackle two of the main difficulties in Bayesian testing when studying the Bayes Factor: choosing a appropriate prior and computing the marginal distribution.

Once the discrepancy measure is chosen, the remaining problem is calibrating the threshold $\tau$. In an informative context where one has prior knowledge on acceptable discrepancy from $\mathcal{M}_0$, $\tau$ can be calibrated subjectively. However, such a prior knowledge may not be available. We thus propose a calibration of $\tau$ based on asymptotic arguments. Heuristically, one would like to find a threshold $\tau$ that minimizes the testing error. Johnson (2013) proposed a similar idea for constructing uniformly most powerful Bayesian tests where he proposes to chose a prior for testing that maximizes the probability that the Bayes Factor exceed a certain threshold for all $\theta \in \mathcal{M}_1$. In our case, in general, minimizing the testing error might not be possible even for some simple models. We thus propose a calibration method based on the asymptotic control of the type I and type II errors. More precisely we chose $\tau = \tau_n$ to be the smallest sequence such that

$$\sup_{\theta \in \mathcal{M}_0} \mathrm{E}_\theta^n\{\delta_n^\pi(\tau_n)\} = o(1), \tag{5}$$

where $\mathrm{E}_\theta^n$ denote the expectation with respect to $Y^n \sim P_\theta$. Given the formulation of the test (4) finding such a calibration will only requires a control of the asymptotic

behaviour of $D(\theta, \mathcal{M}_0)$ under the posterior. We then study for which sequence of $\rho_n$ we have

$$\sup_{\theta \in \mathcal{M}_1, d(\theta, \mathcal{M}_0) > \rho_n} \mathrm{E}_\theta^n \{1 - \delta_n^\pi(\tau_n)\} = o(1). \tag{6}$$

The sequence $\rho_n$ is thus an upper bound on the separation rate of the test (see Lepski and Tsybakov, 2000). Separation rates indicates how close a parameter from $\mathcal{M}_1$ can be to $\mathcal{M}_0$ and sill be detected by the test. Although separation rates have been widely studied in the frequentist literature, to the author's best knowledge, the only related result in the Bayesian literature has been proposed in Rossell and Telesca (2017). Note that if the test separates both hypotheses at the best possible rate in the minimax sense, it indicates that the decision rule $\delta_n^\pi(\tau)$ although being a Bayesian answer to the relaxed testing problem (3), is also an asymptotically optimal frequentist answer for the original testing problem (1). This indicates that such a test can catch up with frequentist methods for detecting parameters close to the boundary between the hypotheses. This is to the best of our knowledge a new result for Bayesian test. A counterpart will be of course a loss in parsimony enforcement. In the remainder of the paper, we study on two examples the problems that can occurs at or close to the boundary between hypotheses. We then propose a general calibration for $\tau_n$ for some usual testing problems and show that our method achieve the minimax separation rates in these cases. On the last sections, we compare our approach to existing ones for a non-parametric test.

## 2 Boundary problems

In this section we illustrate on simple examples the problems faced by the *non-local prior approach to testing* proposed by Johnson and Rossell (2010) and further developed in Rossell and Telesca (2017) and the *standard priors* when the parameter is at, or near the boundary between the null and the alternative.

### 2.1 Point null hypotheses

Consider the following data generating process $X^n \sim \mathcal{N}(\theta, 1/\sqrt{n})$, and the test $H_0 : \theta = 0$ versus $H_1 : \theta \neq 0$. To compute the standard Bayes Factor for this problem, define $\mathcal{M}_0 = \{0\}$ and $\mathcal{M}_1 = \mathbb{R} \backslash \{0\}$, and let the prior distribution $\pi$ on $\theta$ be $\pi : \theta \sim \mathcal{N}(0, \sigma^2)$, and chose equal prior weights on both hypotheses. We can easily derive the usual Bayes-Factor for this problem and get

$$B_{0,1}(X^n) = \frac{\int_{\mathbb{R}} \pi(\theta) e^{-\frac{n}{2}(X^n - \theta)^2} d\theta}{e^{-\frac{n}{2}(X^n)^2}},$$

and compare it to 1. Comparing the Bayes Factor with the fixed threshold $c = 1$ is equivalent to comparing the posterior mass of $\mathcal{M}_0$ with $1/2$. For the non-local prior we use the method of moment proposed in Rossell and Telesca (2017) with parameter fixed as proposed in their paper, i.e.

$$\pi_1^M(\theta) = \frac{\theta^2}{\tau} \pi(\theta/\tau).$$

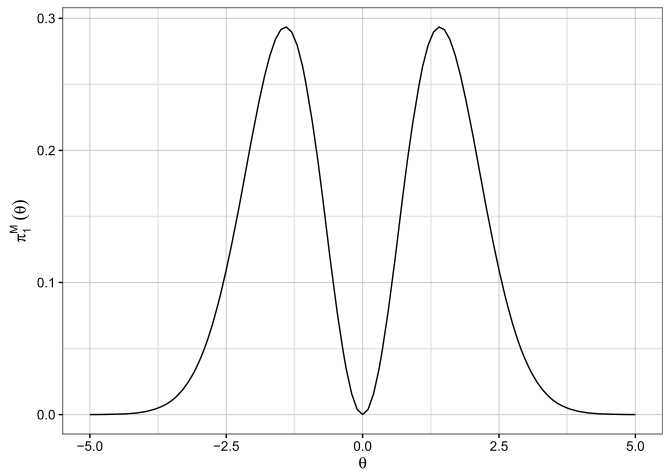


Figure 1: Prior on the alternative constructed with the Method of Moment from Rossell and Telesca (2017). The hyperparameter $\tau$ is fixed to 0.358.

The form of the proposed prior is displayed in Figure 1. We easily derive the Bayes-Factor associated with this prior

$$B_{0,1}^M(X^n) = \frac{\int_{\mathbb{R}} \pi_1^M(\theta) e^{-\frac{n}{2}(X^n-\theta)^2} d\theta}{e^{-\frac{n}{2}(X^n)^2}}.$$

Here again we shall compare it to 1. For the method proposed in this paper, we chose as a discrepancy measure $D(\theta, \mathcal{M}_0) = |\theta|$. We now have to calibrate $\tau_n$ such that the test satisfies (5)–(6). We shall see in the following Theorem 1 that in this case, choosing $\tau_n = u_n n^{-1/2}$ for any $u_n \to +\infty$ will ensure consistency. To calibrate $u_n$, note that we have

$$\pi(D(\theta, \mathcal{M}_0) > \tau_n | X^n) = 1 - \Phi\left(\frac{\tau_n - m_x}{\sigma_x}\right) + \Phi\left(\frac{-\tau_n - m_x}{\sigma_x}\right), \tag{7}$$

where $\sigma_x^2 = (n + \sigma^{-2})^{-1}$ and $m_x = nX^n\sigma_x^2$ are the posterior mean and the posterior variance respectively and $\Phi$ is the cumulative distribution function of a standard Gaussian. To get low type I error while not deteriorating the separation rate we choose $u_n = \max(\Phi^{-1}(0.05), \log(\log(n)))$. We run all three methods on simulated data generated for three different parameters $\theta_0$, namely $\sqrt{2\log(n)/n}$, $\sqrt{\log(n)/n}$ and 0. The first two parameters are getting closer and closer to the boundary between hypotheses as the number of observations grows while the third is in $\mathcal{M}_0$. The results are presented in Figure 2. We observe that even when the parameter is at a reasonable distance from $\mathcal{M}_0$, the non-local prior seems to penalize too much, and thus will contract on the simpler model, while the other approaches do detect the parameter as non-zero. When the parameter is at a distance $\sqrt{\log(n)/n}$ then the usual Bayes Factor does not clearly detect the parameter has non-null, while the proposed method asymptotically does. The price to pay is a slower decay of the type I error of the order of $\log(n)$ to be compared to a exponential decay for the Bayes Factor.

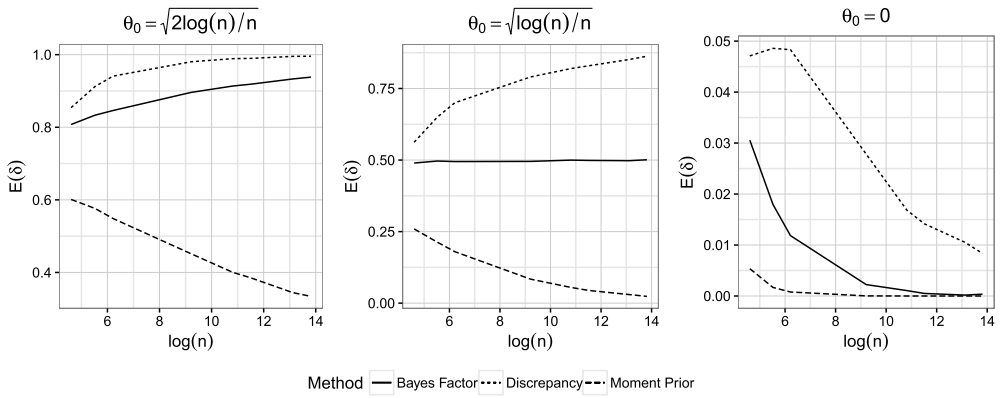


Figure 2: Proportion of test that classifies the parameter as non-null for $N = 5 \times 10^4$ replications of the test. The Bayes Factor obtained with Gaussian and non-local priors are compared to 1. For the discrepancy method $\tau_n = \max[1.96, \log(\log(n))]/\sqrt{n}$.

## 2.2 Un-separated hypotheses

We now consider a case where both the null and the alternative have similar sizes. Using the same setting as before we now test for $H_0 : \theta \leq 0$ versus $H_1 : \theta > 0$, and thus $\mathcal{M}_0 = (-\infty, 0]$ and $\mathcal{M}_1 = (0, +\infty)$, using the same prior $\pi$ as before. To compute the usual Bayes Factor, we thus have the following on $\mathcal{M}_0$ and $\mathcal{M}_1$ respectively:

$$\pi_0(\theta) = 2\pi(\theta)\mathbb{I}_{\theta\in\mathcal{M}_0}, \ \pi_1(\theta) = 2\pi(\theta)\mathbb{I}_{\theta\in\mathcal{M}_1}.$$

We can compute Bayes Factor

$$B_{0,1}(X^n) = \frac{\int_{\mathcal{M}_0} \pi(\theta|X^n)d\theta}{\int_{\mathcal{M}_1} \pi(\theta|X^n)d\theta} = \frac{\Phi\left(-m_x/\sigma_x\right)}{\Phi(m_x/\sigma_x)}.$$

From this formulation, we see that the Bayse Factor $B_{0,1}$ will not detect the parameters $\theta = 0$ (which is at the boundary between $\mathcal{M}_0$ and $\mathcal{M}_1$) as belonging to $\mathcal{M}_0$, leading to poor frequentist performances of such a test in this case. To compare this approach to the non-local prior method we construct a prior $\pi_1^M$ on the alternative using the method of moment described in Rossell and Telesca (2017) that will enforce a separation of the hypotheses. We consider the following modification of the prior $\pi_1^M(\theta) = \frac{\theta^2}{\tau}\pi_1(\theta/\tau)$. A plot of this prior is given in Figure 3. We can then compute the Bayes Factor

$$B_{0,1}^M(X^n) = \frac{\int_{\mathcal{M}_0} \pi_0(\theta)e^{-n(X^n-\theta)^2/2}d\theta}{\int_{\mathcal{M}_1} \pi_1^M(\theta)e^{-n(X^n-\theta)^2/2}d\theta}.$$

One can easily compute the marginals using simple Monte–Carlo integration. Here again we will compare the Bayes Factor $B_{0,1}^M$ with the fixed threshold 1. In order to compare these approaches with the one proposed in this paper, we first need to find a discrepancy measure $D$ and calibrate the threshold $\tau$. We choose $D(\theta, \mathcal{M}_0) = \min(\theta, 0)$. Using a

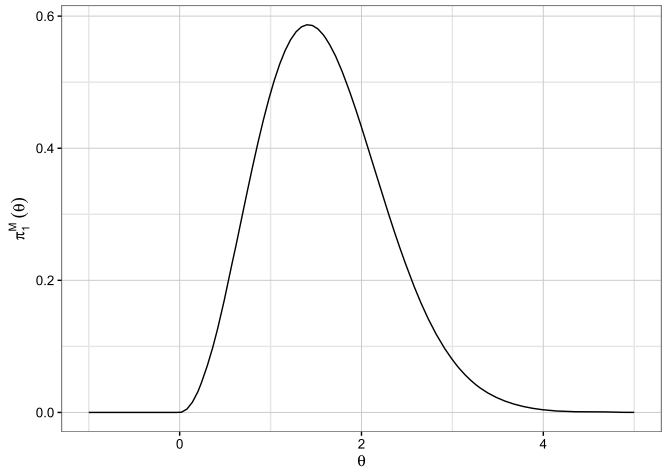


Figure 3: Prior on the alternative constructed with the Method of Moment from Rossell and Telesca (2017). The hyperparameter $\tau$ is fixed to 0.358.

simple standard Gaussian prior, we can easily calibrate the threshold $\tau_n$ using the same approach as before. We have that for all sequence $u_n$ that goes to infinity as slowly as needed, $\tau_n = Cu_n n^{-1/2}$ leads to a separation rate $\rho_n \leq 2\tau_n$. We now calibrate the constant $C$ and the sequence $u_n$ based on heuristics. Again, denoting $\hat{\sigma}_x^2 = (n+1/\sigma^2)^{-1}$ and $m_x = nX^n\hat{\sigma}_x^2$ the posterior mean and posterior variance respectively, we have

$$\Pi(D(\theta, \mathcal{M}_0) \geq \tau_n | X^n) = 1 - \Phi\left(\frac{\tau_n - \hat{m}_x}{\hat{\sigma}_x}\right).$$

We then choose again $u_n = \max(\log(\log(n)), \Phi^{-1}(0.05))$ which insure consistency while not deteriorating the separation rate too much.

Similarly to what we did in the previous section, we compare the results obtained with the three different methods on simulated data generated with a parameter $\theta_0 = \sqrt{2\log(n)/n}$, $\sqrt{\log(n)/n}$ and 0. The results are given in Figure 4. We observe that the Bayes Factor constructed using the non-local prior of Rossell and Telesca (2017) has difficulties to detect parameters in $\mathcal{M}_1$ but close to $\mathcal{M}_0$ as positive due to the penalization induced by the prior. On the other hand the usual Bayes Factor based on the simple conjugate Gaussian prior do not detect $\theta = 0$ in $\mathcal{M}_0$ while the other two methods have good asymptotic behaviour. We thus see that both the usual Bayes Factor and the approach based on non-local priors have difficulties to detect parameters at or near the boundary. More worryingly, the behaviour of these methods near the boundary strongly depends on the sets $\mathcal{M}_0$ and $\mathcal{M}_1$ and on the prior contraction on these sets. On the other hand the proposed method, although a little less efficient for finite sample sizes, does detect the parameter at or near the boundary. An easy fix in this particular setting to get better results for the Bayes Factor and non-local priors would be to elicit $\tilde{\mathcal{M}}_0 = \{0\}$. Nevertheless the same behaviours exposed in the previous section would remain. Furthermore, for more complex hypotheses, it could be difficult to single out

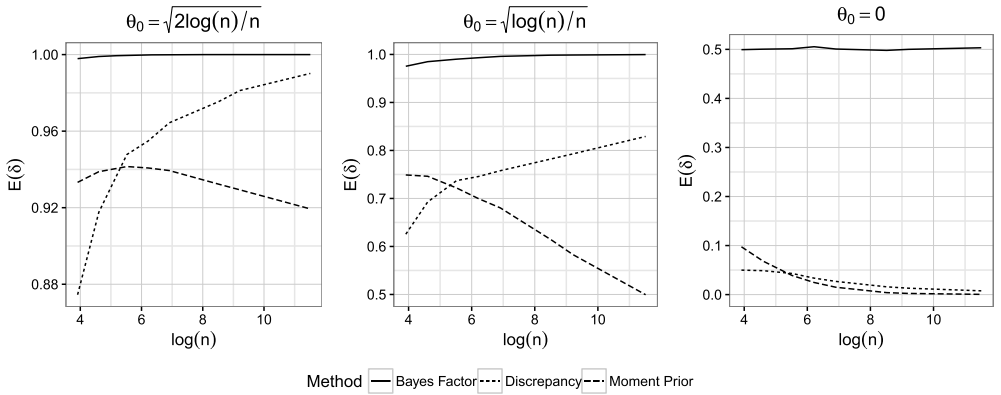


Figure 4: Proportion of test that classifies the parameter as negative for $N = 5 \times 10^4$ replications of the test. The Bayes Factor obtained with Gaussian and non-local priors are compared to 1. For the discrepancy method $\tau_n = \max[1.96, \log(\log(n))]/\sqrt{n}$.

the boundary as a separated hypothesis. We shall see in the next section that for these examples, the proposed method attains asymptotically the minimax separation rate.

# 3 Application to standard testing problems

## 3.1 Testing parametric hypotheses

Consider the following parametric model for some fixed $p > 0$, $Y^n \sim P_\theta^n$, for $\theta \in \Theta \subset \mathbb{R}^p$. For a fixed subset $\Theta_0 \subset \Theta$ we want to test $H_0 : \theta \in \mathcal{M}_0 = \Theta_0$, versus $H_1 : \theta \in \mathcal{M}_1 = \Theta \cap \Theta_0^c$. This problem has been widely studied in the Bayesian literature (see Robert, 2007, for instance).

In this simple case, the following theorem gives a calibration for the threshold $\tau_n$ in (3) such that the testing procedure satisfies condition (5) and (6), and gives an upper bound for the separation rate $\rho_n$.

**Theorem 1.** *Let $\Pi$ be a prior distribution on $\Theta$ and $d$ be a metric on the parameter space $\Theta$. Assume that for some positive sequence $\epsilon_n$ we have*

$$\sup_{\theta^* \in \Theta} \mathrm{E}_{\theta^*}^n \Pi[d(\theta, \theta^*) > \epsilon_n | Y^n] = o(1). \tag{8}$$

*Then choosing $D(\theta, \mathcal{M}_0) = \inf_{\theta^* \in \Theta_0} d(\theta, \theta^*)$ and $\tau_n = \epsilon_n$ in (3) the decision rule $\delta_n^\pi$ in (4) satisfies*

$$\sup_{\theta \in \Theta_0} \mathrm{E}_\theta^n[\delta_n^\pi(\tau_n)] = o(1), \qquad \sup_{\theta \in \Theta, d(\theta, \Theta_0) > 2\epsilon_n} \mathrm{E}_\theta^n[1 - \delta_n^\pi(\tau_n)] = o(1).$$

Condition (8) is the standard concentration property of the posterior which is known to hold for regular models with $\epsilon_n = n^{-1/2} u_n$ where $u_n$ is any positive sequence increas-

ing to infinity (see for instance Ghosal et al., 2000a; Ghosal and van der Vaart, 2007). In this case the separation rate $\rho_n$ for the proposed test is the minimax separation rate $n^{-1/2}$ up to some factor $u_n$. The proof of this Theorem is postponed to Section 6.1. From the proof of Theorem 1, we can also derive an upper bound for $\sup_{\theta\in\Theta_0} \mathrm{E}_\theta^n[\delta_n^\pi(\tau)]$ and $\sup_{\theta\in\Theta, d(\theta,\Theta_0)>2\epsilon_n} \mathrm{E}_\theta^n[1-\delta_n^\pi(\tau)]$ of the order of $\sup_{\theta^*\in\Theta} \mathrm{E}_{\theta^*}^n \Pi[d(\theta,\theta^*) > \epsilon_n | Y^n]$. Under some regularity assumptions on the models, we get that the type I and type II error can be uniformly bounded by $e^{-Cu_n^2}$ for some constant $C > 0$. Choosing $u_n$ of the order of $\sqrt{\log(n)}$ will thus give polynomial decay uniformly for both errors. As argued in Johnson and Rossell (2010), the Bayes Factor usually contracts at an exponentially fast rate, for a true alternative. However, this is to be balanced with the fact that here the proposed control is uniform over all $\theta \in \Theta$ such that $d(\theta, \Theta_0) > 2\epsilon_n$.

## 3.2 Detection of signal in white noise

We now apply our approach to the problem of detecting signal in the standard white noise model. This problem is closely related to the well studied goodness-of-fit testing problem, where one is interested in testing a parametric hypothesis versus a non-parametric one. Here again this problem has been extensively studied in the literature. Goodness of fit testing have been considered both from a frequentist and Bayesian point of view, see for instance Ingster and Suslina (2003), Dass and Lee (2004) or see Tokdar et al. (2010) for a review. The specific problem of detection of signal in white noise has also been treated in Ingster (1987); Lepski and Spokoiny (1999); Lepski and Pouet (2008).

Here we consider the equivalent infinite Gaussian sequence model

$$Y_i = f_i + \frac{\epsilon_i}{\sqrt{n}},\ \epsilon_i \overset{iid}{\sim} \mathcal{N}(0,1),\ i \geq 1, \tag{9}$$

where $f = (f_i) \in l^2 = \{g, \sum_i g_i^2 < \infty\}$. Similarly to Lepski and Spokoiny (1999) we test $f = 0$ against a Sobolev ellipsoid of fixed smoothness $s$, $W_2^s(L) = \{f \in l_2, \sum_{i=1}^\infty f_i^2 i^{2s} \leq L\}$. We thus have $\mathcal{M}_0 = \{f = 0\}$ and $\mathcal{M}_1 = \{f \in W_s^2(L), f \neq 0\}$. We consider a conjugate Gaussian prior on as in Section 3 of Castillo and Rousseau (2015). For $k_n = n^{2/(4s+1)}$ and all increasing sequence $s = (s_1, s_2, \ldots)$ such that $s_{k_n} \leq n^{4s/(4s+1)}$ and $\sum_{i=1}^{k_n} 1/(n+s_i) \leq \rho_n/4$ we define $\Pi$ by

$$(f_1, \ldots, f_{k_n}) \sim \bigotimes_{i=1}^{k_n} \mathcal{N}(0, s_i^{-1}),\ f_j = 0\ \forall j > k_n. \tag{10}$$

We choose the discrepancy measure $D(f, \mathcal{M}_0)$ to be the $l_2$ norm of $f$, $||f||_2 = (\sum_{i=0}^\infty f_i^2)^{1/2}$. The following Theorem gives a calibration for the threshold $\tau_n$ in (3) and an upper bound for the separation rate of our testing procedure.

**Theorem 2.** *Let $Y^n$ be sample from (9) and consider a prior on $f$ as defined in (10). Let $v_n$ be any sequence increasing to infinity and let $\rho_n = v_n n^{-2s/(4s+1)}$ and $\tau_n$ be such that $\tau_n^2 = C\rho_n/2 + k_n/n + \sum_{i=1}^{k_n} \frac{1}{n+s_i}$ for some positive constant $C$. Setting $d$ to be the $l_2$ norm, the decision rule $\delta_n^\pi$ as defined in (4) satisfies*

$$\mathrm{E}_0^n(\delta_n^\pi) = o(1), \qquad \sup_{f\in\mathcal{M}_1, ||f||_2>\rho_n} \mathrm{E}_f^n(1-\delta_n^\pi) = o(1). \tag{11}$$

Here again the separation rate $\rho_n$ of the test is the minimax separation rate as shown in Ingster (1987). An interesting aspect of this test is that it does not rely on the precise estimation of the true underlying function but rather on the semiparametric estimation of $D(f, \mathcal{M}_0)$ which allows us to obtain a separation rate polynomially faster than the estimation rate for Sobolev alternative. It is to be noted that the prior (10) is not optimal for the estimation problem but leads to the best possible separation rate for the testing problem. The proof of this theorem is postponed to Section 6.2.

# 4 Shape constraints testing

## 4.1 Statistical setting

We consider the nonparametric fixed design regression problem with Gaussian residuals for $n > 0$

$$Y_j = f(j/n) + \sigma\epsilon_i, \ j = 1, \dots, n, \tag{12}$$

where $\sigma > 0$ and $(\epsilon_1, \dots, \epsilon_n)$ is a sequence of independent standard Gaussian random variable. The approach presented in this paper are also valid for non-uniform design and random design under additional condition but considering these cases will only make the computations more complex and will thus not be treated here. For this problem, we consider a piecewise constant prior distribution on the regression function $f$ and a prior with density $\pi_\sigma$ with respect to the Lebesgue measure on $\sigma$. More precisely, for $I_i = [i - 1/k, i/k)$ the uniform partition of $[0, 1]$, we define functions $f_{\omega,k}$ as

$$f_{\omega,k}(\cdot) = \sum_{i=1}^{k} \omega_i \mathbb{I}_{I_i}.$$

We choose the following form for the prior on $f$

$$d\Pi(f) = \pi_k(k)\pi_\omega(\omega_1, \dots, \omega_k|k)d\lambda_k(\omega_1, \dots, \omega_k)d\nu(k), \tag{13}$$

where $\lambda_k$ is the Lebesgue measure on $\mathbb{R}^k$ and $\nu$ the counting measure on $\mathbb{N}$. Note that a similar prior has been studied in Holmes and Heard (2003) for modelling monotone functions. Here again, although this prior is not well suited for the estimation problem, it gives good theoretical and practical results for testing the shape constraints studied in this paper as shown bellow. For simplicity we consider a product form for $\pi_\omega$, $\pi_\omega(\omega_1, \dots, \omega_k|k) = \prod_{i=1}^{k} g(\omega_i)$ where $g$ is a density on $\mathbb{R}$. In addition we assume that the following conditions holds

**C1** the density $\pi_\sigma$ is bounded and continuous and $\pi_\sigma(\sigma) > 0$ for all $\sigma \in (0, \bar{\sigma})$,

**C2** the density $g$ is continuous positive on $\mathbb{R}$ and bounded from above.

**C3** $\pi_k$ is such that there exists positive constants $C_d$ and $C_u$ such that

$$e^{-C_d k L(k)} \leq \pi_k(k) \leq e^{-C_u k L(k)}, \tag{14}$$

where $L(k)$ is either $\log(k)$ or 1.

The condition **C1** and **C2** are mild and are satisfied for a large variety of distributions. In Section 5.1 we will take $g$ to be a Gaussian density and $\pi_\sigma$ to be a inverse gamma density. Simple algebra shows that for this choice of prior, both conditions are satisfied. Condition **C3** is a usual condition when considering mixture models with random number of components, see e.g. Rousseau (2010) and is satisfied by Poisson or Geometric distribution for instance.

Define the sets

$$\mathcal{F}_+ = \{f \in L_\infty([0,1]) : \forall x \in [0,1], f(x) > 0\}$$
$$\mathcal{F}_\searrow(K) = \{f : ||f||_\infty \leq K, \forall x \leq y \ f(x) \geq f(y)\}$$

of positive and monotone non-increasing functions respectively. For $\alpha > 0$ and $L > 0$, define $\mathcal{H}(\alpha, L) = \{f, ||f||_{H,\alpha} \leq L\}$ where $||\cdot||_{H,\alpha}$ is the Hölder norm. We consider both testing problems $H_0 : f \in \mathcal{M}_0 = \mathcal{F}_+$ versus $H_1 : f \in \mathcal{M}_1 = \mathcal{H}(\alpha, L) \cap \mathcal{F}_+^c$ and $H_0 : f \in \mathcal{M}_0 = \mathcal{F}_\searrow(K)$ versus $H_1 : f \in \mathcal{M}_1 = \mathcal{H}(\alpha, L) \cap \mathcal{F}_\searrow(K)^c$. These problem has been considered in the literature in Juditsky and Nemirovski (2002) and Baraud et al. (2005) for instance. Note that with a prior chosen as in (13) we have $\pi(\mathcal{F}_+) > 0$ and $\pi(\mathcal{F}_\searrow(K)) > 0$. Furthermore, if the true regression function $f_0$ is in $\mathcal{F}_+$ or $\mathcal{F}_\searrow(K)$ then the piecewise constant function with $k$ pieces of the form (13) which minimizes the Kullback Leibler divergence with $P_{f_0}$ will also be in $\mathcal{F}_+$, respectively $\mathcal{F}_\searrow(K)$, for all $k$.

We then study the posterior separation rate of the test with respect to the metric $d = d_\infty$ defined as

$$d_\infty(f, g) = \sup_{x \in [0,1]} |f(x) - g(x)|.$$

For each test we compute the separation rate of our procedure and compare it with the minimax separation rates, which is $n^{-\alpha/(2\alpha+1)}$ in both cases.

Our approach could also apply to other types of shape constraints such as convexity or unimodality using similar methods.

## 4.2 Testing for positivity

We first consider positivity constraints. There exist a few methods to test for positivity in a nonparametric setting, see for instance Baraud et al. (2005). We propose the following discrepancy measure for $D$ in (3)

$$D(f, \mathcal{F}_+) = -\inf_{x \in [0,1]} f(x). \tag{15}$$

We immediately have that $D(f, \mathcal{F}_+) \leq 0$ if and only if $f \in \mathcal{F}_+$. Here the discrepancy measure can be related to the supremum distance with the set of positive functions. For piecewise constant functions $f_{\omega,k}$, $D(f_{\omega,k}, \mathcal{F}_+)$ has the simple expression $D(f_{\omega,k}, \mathcal{F}_+) = -\min_{1 \leq i \leq k}(\omega_i)$. This turn out to be particularly useful for the calibration of the threshold $\tau_n$. Let $\mathcal{G}_k$ be the set of piecewise constant function with $k$ pieces. The idea of the calibration of $\tau_n$ is the following. In the model $\mathcal{G}_k$, the a posteriori uncertainty for estimating $\omega = (\omega_1, \ldots, \omega_k)$ is of order $(k/n)^{1/2}$. Hence any positive

function $f_{\omega,k}$ such that for all $i$, $\omega_i \geq O\{(k/n)^{1/2}\}$ might be detected as possibly positive. We thus choose a threshold $\tau_n^k$ for each model $\mathcal{G}_k$ of similar order. The results are presented in the following theorem.

**Theorem 3.** *Under the assumptions **C1** to **C3**, and if $\underline{\sigma} < \sigma \leq \bar{\sigma}$ for fixed $0 < \underline{\sigma} \leq \bar{\sigma}$, then for a fixed constant $M_0 > 0$, setting $\tau = \tau_n^k = M_0\{k\log(n)n^{-1}\}^{1/2}$ and $\delta_n^\pi$ the testing procedure defined in (4), for all $K > 0$ there exists some $M > 0$ such that uniformly for $\alpha \in [\alpha_0, 1]$, $\forall \alpha_0 > 0$*

$$\begin{aligned} \sup_{\underline{\sigma}<\sigma\leq\bar{\sigma}} \sup_{f\in\mathcal{F}_+} \mathrm{E}_{f,\sigma}^n(\delta_n^\pi) &= o(1) \\ \sup_{\underline{\sigma}<\sigma\leq\bar{\sigma}} \sup_{f,d_\infty\{f,\mathcal{F}_+\}>\rho, f\in\mathcal{H}(\alpha,L)} \mathrm{E}_{f,\sigma}^n(1-\delta_n^\pi) &= o(1) \end{aligned} \tag{16}$$

*for all $\rho > \rho_n(\alpha) = M\{n/\log(n)\}^{-\alpha/(2\alpha+1)}v_n$ where $v_n = 1$ when $L(k) = \log(k)$ and $v_n = \{\log(n)\}^{1/2}$ when $L(k) = 1$.*

## 4.3 Testing for monotonicity

We now consider monotonicity constraints. Tests for monotonicity have been well studied in the frequentist literature, see for instance Baraud et al. (2003, 2005); Ghosal et al. (2000b); Bowman et al. (1998). In a Bayesian setting, only Scott et al. (2015) proposed a test for monotonicity using non-local priors. Define the discrepancy measure between $f$ and $\mathcal{F}_+$ as

$$D(f, \mathcal{F}_+) = \sup_{0\leq x<y\leq 1} \{f(y) - f(x)\}. \tag{17}$$

Here again when considering piecewise constant functions $f_{\omega,k}$, (17) we get the simple formulation $D(f_{\omega,k}, \mathcal{F}_+) = \max_{1\leq i\leq j\leq k}(\omega_j - \omega_i)$ which allows for a simple calibration of $\tau_n$ in a similar way as in Section 4.2.

**Theorem 4.** *Under the assumptions **C1** to **C3**, for a fixed constant $M_0 > 0$, setting $\tau = \tau_n^k = M_0\{k\log(n)n^{-1}\}^{1/2}$ and $\delta_n^\pi$ the testing procedure defined in (4), for all $K > 0$ then there exists some $M > 0$ such that uniformly for $\alpha \in [\alpha_0, 1]$, $\forall \alpha_0 > 0$*

$$\begin{aligned} \sup_{\underline{\sigma}<\sigma\leq\bar{\sigma}} \sup_{f\in\mathcal{F}_\searrow(K)} \mathrm{E}_f^n(\delta_n^\pi) &= o(1) \\ \sup_{\underline{\sigma}<\sigma\leq\bar{\sigma}} \sup_{f,d_\infty\{f,\mathcal{F}_\searrow(K)\}>\rho, f\in\mathcal{H}(\alpha,L)} \mathrm{E}_f^n(1-\delta_n^\pi) &= o(1) \end{aligned} \tag{18}$$

*for all $\rho > \rho_n(\alpha) = M\{n/\log(n)\}^{-\alpha/(2\alpha+1)}v_n$ where $v_n = 1$ when $L(k) = \log(k)$ and $v_n = \{\log(n)\}^{1/2}$ when $L(k) = 1$.*

Neither the prior nor the threshold depend on the regularity $\alpha$ of the regression function under the alternative. Moreover for all $\alpha \in (0, 1]$, the separation rate $\rho_n(\alpha)$ is the minimax separation rate up to a $\log(n)$ term. Thus our test is almost minimax adaptive. The $\log(n)$ term seems to follow from our definition of the consistency where we do not fix a level for the Type I or Type II error contrariwise to the frequentist

procedures. The conditions on the prior are quite loose, and are satisfied in a wide variety of cases. The constant $M_0$ does not influence the asymptotic behaviour of our test but has a great influence in practice for finite $n$. A practical way of choosing $M_0$ is given in Section 5.1.

# 5 Simulation study for positivity and monotonicity testing

## 5.1 Prior specification and sampling strategy

Conditions on the prior in Theorem 4 are satisfied for a wide variety of distributions. However, when no further information is available, some specific choices can ease the computations and lead to good results in practice. We present in this section such a specific choice for the prior and a way to calibrate the hyperparameters. We also fix $\gamma_0 = \gamma_1 = 1/2$ in the definition of $\delta_n^\pi$.

A practical default choice is the usual conjugate prior, given $k$, i.e. a Gaussian prior on $\omega$ with variance proportional to $\sigma^2$ and an Inverse Gamma prior on $\sigma^2$. This will considerably accelerate the computations as sampling under the posterior is then straightforward. Condition (14) on $\pi_k$ is satisfied by the two classical distributions on the number of parameters in a mixture model, namely the Poisson distribution and the Geometric distribution. It seems that choosing a Geometric distribution is more appropriate as it is less spiked. We thus choose for $\lambda, a, b > 0$, $m \in \mathbb{R}$ and $\mu > 0$

$$\Pi = \begin{cases} k \sim \text{Geom}(\lambda), \\ \sigma^2 | k \sim IG(a, b), \\ \omega_i | k, \sigma \overset{iid}{\sim} \mathcal{N}(m, \sigma^2/\mu). \end{cases} \tag{19}$$

Standard algebra leads to a close form for the posterior distribution up to a normalizing constant. Let $n_i = \text{Card}\{j, j/n \in [(i-1)/k, i/k)\}$, we denote

$$\tilde{b}_k = b + \frac{1}{2} \sum_{i=1}^{k} \left\{ \sum_{j, j/n \in I_i} \left(Y_j - \overline{Y_i}\right)^2 + \frac{n_i \mu}{n_i + \mu} (\overline{Y_i} - m)^2 \right\},$$

where $\overline{Y_i}$ is the empirical mean of the $Y_l$ on the set $\{j, j/n \in [(i-1)/k, i/k)\}$, we have

$$\pi_k(k|Y^n) \propto \pi(k) \tilde{b}_k^{-(\alpha + n/2)} \mu^{k/2} \prod_{i=1}^{k} (n_i + \mu)^{-1/2}.$$

We can thus compute the posterior distribution of $k$ up to a constant. We will thus be able to sample from $\pi_k(k|Y^n)$ using a truncated approximation of the posterior. In the examples we choose to truncate at some $k_0 \leq n$. We then compute the posterior distribution of $\omega$ and $\sigma$ given $k$

$$\sigma^2 | k, Y^n \sim IG(a + n/2, \tilde{b}_k)$$

$$\omega_j | k, \sigma^2, Y^n \overset{ind.}{\sim} \mathcal{N}\left(\frac{m\mu + n_j \bar{Y}_j}{n_j + \mu}, \frac{\sigma^2}{n_j + \mu}\right).$$

Given $k$, sampling from the posterior is thus straightforward.

A crucial hyperparameter that needs to be calibrate is for $M_0$ the constant in $\tau$. A close inspection of the proofs (in particular the proof of Lemma 2) using the fact that we have a Gaussian posterior, gives us that taking

$$\tau_n = \sqrt{\frac{\log(k/n) k\sigma^2}{n + k\mu\sigma^2}},$$

would induce the desired results.

## 5.2 Simulated examples

In this section we run our testing procedure on simulated data to study the behaviour of our test for finite sample sizes. We first examine the behaviour of the proposed test for positivity on an example that illustrate that the separation rate of the test is indeed upper bounded by $(\log(n)/n)^{\alpha/(2\alpha+1)}$ up to some constant. We then compare our test for monotonicity to other methods proposed in the literature, and get comparable results for finite sample size.

### Testing for positivity

Consider the test for positivity proposed in Section 4.2. Similarly to the examples of Section 2, we will consider a sequence of function that are in $\mathcal{M}_1$ i.e. not positive, but are getting closer and closer to the boundary. More precisely we take

$$f_n(x) = 10\rho_n(|x - 0.1| - 0.1)\mathbb{I}_{|x-0.5|<0.1},$$

and thus $\rho_n = d_\infty(f, \mathcal{F}_+)$. Plots of $f_n$ for different values of $n$ are given in Figure 5. Since for all $n$ this function is piecewise linear, we thus have that $f \in \mathcal{H}(\alpha, L)$, with $\alpha = 1$. Given Theorem 3, we have that for some constant $M$ large enough, the test should be consistent for $f_n$ if $\rho_n > M(\log(n)/n)^{1/3}$.

We run our test on simulated data generated from the model (12) with $f = f_n$ for different values of $M$ and with $f = 0$ that lies at the boundary between hypotheses. The results are given in Figure 6. We observe that the test detects parameter at the boundary as positive, even for moderate values of $n$. In addition, for $M > 0.4$, the function $f_n$ are detected as non-positive, and the asymptotic regime is attained around $n = 2000$, while for $M < 0.4$ the functions $f_n$ are not detected as non-positive. This indicates that the test does separate the hypotheses at the rate at least $0.4(\log(n)/n)^{1/3}$, and we thus recover the results from Theorem 3.

### Testing for monotonicity

We now compare our approach to test for monotonicity with the ones proposed in the literature. We consider the following nine functions adapted from Scott et al. (2015)

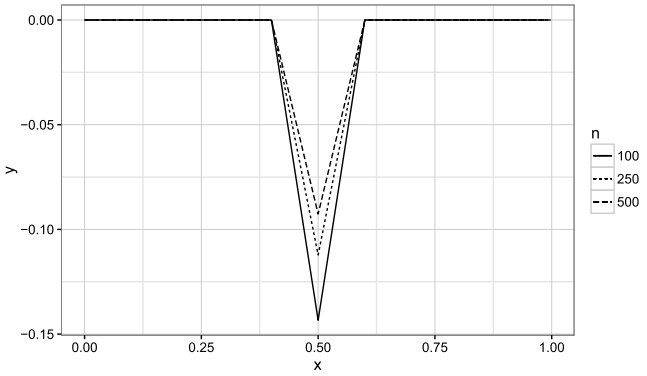


Figure 5: Plots of $f_n$ for different values of $n$ with $\rho_n = 0.4(\log(n)/n)^{1/3}$.

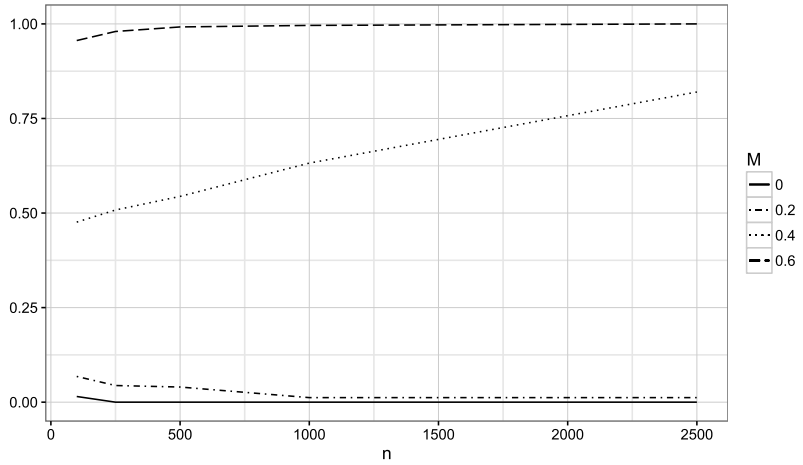


Figure 6: Proportion of functions classified as non-positive estimated on $K = 500$ independent replication of simulated data generated from model (12) with $f = 10M(\log(n)/n)^{1/3}(|x - 0.1| - 0.1)\mathbb{I}_{|x-0.5|<0.1}$ for different values of $M$.

and Baraud et al. (2003) and plot in Figure 7.

$$
\begin{aligned}
f_1(x) =& -4(x-0.5)^3\mathbb{I}_{x\leq 1/2} - \\
& 0.1(x-0.5) + 0.25e^{-250(x-0.25)^2} \\
f_2(x) =& 0.1x \\
f_3(x) =& 0.1e^{-50(x-0.5)^2} \\
f_4(x) =& -0.1\cos(6\pi x) \\
f_5(x) =& -0.2x + f_3(x) \\
f_6(x) =& -0.2x + f_4(x) \\
f_7(x) =& -(1+x) + 0.25e^{-50(x-0.5)^2} \\
f_8(x) =& -x - 1 + 0.45e^{-50(x-0.5)^2} \\
f_9(x) =& -0.5x^2 \\
f_{10}(x) =& 0 \\
f_{11}(x) =& -x - 1
\end{aligned}
\tag{20}
$$

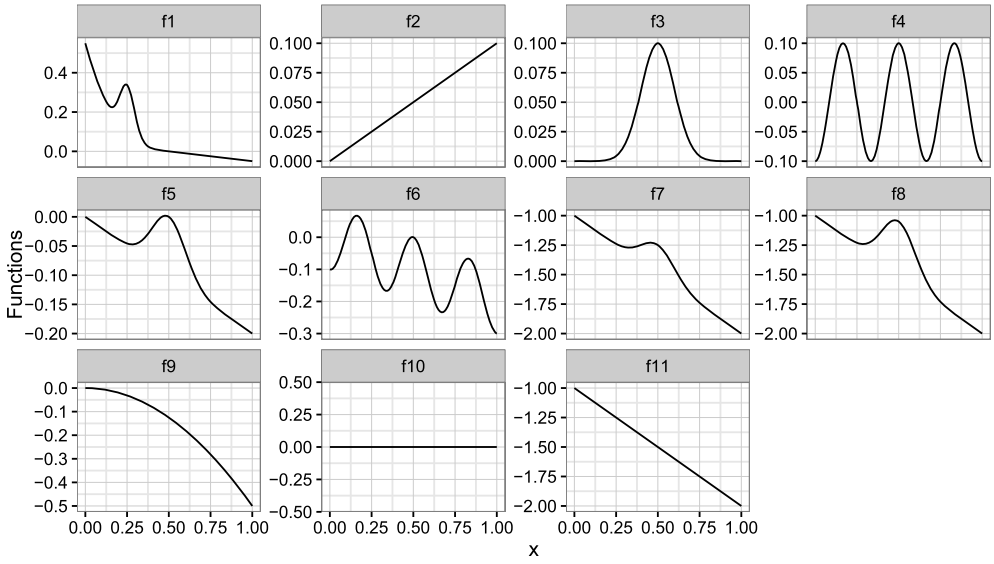


Figure 7: Regression functions used in the simulated examples.

The functions $f_1$ to $f_6$ are clearly not in $\mathcal{F}_{\searrow}(K)$ with $K = 2$. The function $f_7$ has a small bump around $x = 0.5$ which can be seen as a local departure from monotonicity. This function is thus expected to be difficult to detect for small datasets given our parametrization. The function $f_9$ is a completely flat function and belongs to $\mathcal{F}_{\searrow}(K)$.

For several values of $n$, we generate $N = 500$ replicates of the data $Y^n = \{y_i, i = 1, \ldots, n\}$ from model (12). For each dataset, we approximate $\pi\{D(f_{\omega,k}, \mathcal{F}_{\searrow}) > \tau_n^k | Y^n\}$ based on $K = 5 \times 10^4$ samples from the posterior and reject the null if

$$\hat{\pi}\left\{D(f_{\omega,k}, \mathcal{F}_{\searrow}) > \tau_n^k | Y^n\right\} = \frac{1}{K}\sum_{i=1}^{K} \mathbb{I}\left\{D(f_{\omega^i,k^i}, \mathcal{F}_{\searrow}) > \tau_n^{k^i}\right\} \geq 1/2.$$

The results are given in Table 1.

For all the considered functions, the computational time is reasonable even for large values of $n$. For instance, for $f_1$, we require less than 2 seconds to perform the test for $n = 2500$ using a simple R script available on demand. We compare our results with the ones obtained in Scott et al. (2015) for the Gaussian prior and the methods proposed by Baraud et al. (2003) and Akakpo et al. (2014). The results are given in Table 1. The proposed method is a little less efficient than the one based on non-local prior in average, but it seems to perform better for some functions (e.g. $f_3$). When $n$ grows, the percentage of correctly classified function goes to 1 as predicted by the theory.

| | $f_0$ | $\sigma^2$ | Barraud | Akakpo | Scott | Discrepancy |
|---|---|---|---|---|---|---|
| $H_1$ | $f_1$ | 0.01 | 6 | 9 | 100 | 80 |
| | $f_2$ | 0.01 | 64 | 33 | 74 | 75 |
| | $f_3$ | 0.01 | 53 | 43 | 35 | 77 |
| | $f_4$ | 0.01 | 92 | 92 | 91 | 100 |
| | $f_5$ | 0.01 | 24 | 25 | 85 | 17 |
| | $f_6$ | 0.01 | 77 | 75 | 99 | 99 |
| | $f_7$ | 0.01 | 1 | 4 | 91 | 39 |
| | $f_8$ | 0.01 | 71 | 82 | 99 | 99 |
| $H_0$ | $f_9$ | 0.01 | 100 | 100 | 93 | 93 |
| | $f_{10}$ | 0.01 | 97 | 94 | 95 | 89 |
| | $f_{11}$ | 0.01 | 100 | 99 | 97 | 94 |
| Average | | | 62.3 | 59.5 | 87.2 | 78.36 |

Table 1: Results of the simulation study. Each entry is the percentage of correctly classified functions estimated on $K = 500$ independent replication of the experiment. Baraud method for Baraud et al. (2003), Akakpo for Akakpo et al. (2014), Scott for Scott et al. (2015) with Gaussian prior and Discrepancy for the proposed method.

# 6 Proofs

## 6.1 Proof for the parametric test

We prove that with the proposed calibration for $\tau_n$ and $D(\theta, \Theta_0) = \inf_{t\in\Theta_0} d(\theta, t)$, the decision rule satisfies (5)–(6). For $\theta^* \in \Theta_0$, we directly have that

$$D(\theta, \Theta_0) \leq d(\theta, \theta^*),$$

which together with (8) gives (5). Now for $\theta^* \in \Theta \cap \Theta_0^c$ we have if $d(\theta^*, \Theta_0) > \rho_n$

$$D(\theta, \Theta_0) \geq d(\theta^*, \Theta_0) - d(\theta, \theta^*) \geq \rho_n - d(\theta, \theta^*).$$

We deduce (6) directly from condition (8) which ends the proof.

## 6.2 Proof for the detection of signal in white noise

We fist prove that with the proposed calibration of $\tau_n$ the decision rule (4) satisfies (5). In the sequel $c$ will denote a generic absolute constant that may change from one line to another. We want to bound $\Pi(||f||^2 > \tau_n^2 | Y^n)$ when $f_0 = 0$ for $\tau_n^2 = \rho_n/2 + k_n/n + \sum_{i=1}^{k_n} \frac{1}{n+s_i}$. For all $t \leq 2n$ we have, using the Chernoff bound we have with $P_0^n$-probability that goes to 1

$$\begin{aligned}\log\{\Pi(||f||^2 > \tau_n^2 | Y^n)\} &\leq -t\tau_n^2 + \sum_{i=1}^{k_n} t\frac{\epsilon_i^2}{n} + \sum_{i=1}^{k_n} \frac{1}{n+s_i} \\ &\leq -t\frac{\rho_n}{2}\end{aligned}$$

$$\leq 1 - c$$

for $c$ large enough, which give the result

We now state an auxiliary result that will be needed for the remainder of the proof. Define $H(s,\rho) = \{f \in W_2^s(L), ||f||_2 > \rho\}$

**Lemma 1.** *Let $k_n = n^{2/(4s+1)}$ and consider $Z_n = \sum_{i=1}^{k_n}(Y_i^2 - 1/n)$. We thus have if $f_0 \in H(s,\rho_n)$ with $\rho_n = v_n n^{-4s/(4s+1)}$ where $v_n \to \infty$ slowly with $n$ then for some $C > 0$*

$$P_0^n(Z_n \leq \rho_n) = o(1). \tag{21}$$

The proof of this Lemma can be found in the supplementary materials (Salomond, 2017).

We now end the proof by showing that $\delta_n^\pi$ satisfies (6). We want to bound $\Pi(||f||^2 \leq \tau_n|Y^n)$ when $f_0 \in H(s,\rho_n)$. For all $n/2 > t > 0$ and all increasing sequence $s_i$ such that $s_{k_n} \leq n^{4s/(4s+1)}$ we have for $Y^n$ such that $P_0^n(Z_n > \rho_n)$, using the Chernoff bound and the fact that $\sum_{i=1}^{k_n} 1/(n+s_i) \leq \rho_n/4$

$$\begin{aligned}
\log\{\Pi(||f||^2 \leq \tau_n^2|Y^n)\} &\leq tu_n - \sum_{i=1}^{k_n}\left\{\frac{tn^2Y_i^2}{(n+s_i)(n+s_i+2t)} - \frac{1}{2}\log\left(1+\frac{2t}{n+s_i}\right)\right\} \\
&\leq t\tau_n^2 - \sum_{i=1}^{k_n}\left\{tY_i^2(1-2\frac{s_i+t}{n}) - \frac{t}{n+s_i}\right\} \\
&\leq -\frac{t\rho_n}{4} + 2k_n\frac{s_{k_n}+t}{n^2} + \frac{k_nt^2}{n^2} \\
&\leq c,
\end{aligned}$$

for some $c$ for $v_n$ large enough by taking $t \asymp \rho_n^{-1}$ and $s_{k_n} \leq t^2$, where the second line comes from the fact that $\frac{1}{(n+s_i)(n+s_i+2t)} \geq \frac{1}{n^2}(1-2\frac{s_i+t}{n})$.

## 6.3 Proof for shape constraints

## 6.4 Auxiliary result

For all functions $f_0$ in $L_\infty([0,1])$ denote by $P_0$ the probability distribution of $Y^n$ generated with $f = f_0$ and $f_{\omega^0,k}$ the function of $\mathcal{G}_k$ the set of piecewise constant functions with $k$ pieces, that minimizes the Kullback Leibler divergence between $P_f$ and $P_0$. Standard computation gives

$$\omega_i^0 = n_i^{-1} \sum_{j, j/n \in [(i-1)/k, i/k)} f_0(j/n),\ n_i = \text{Card}\left\{j, j/n \in [(i-1)/k, i/k)\right\}. \tag{22}$$

The following lemma gives some concentration result for $f_{\omega,k}$ that will be useful for the study of $D(f,\mathcal{F}_+)$ or $D(f,\mathcal{F}_\searrow(K))$ respectively for both monotonicity and positivity constraints.

**Lemma 2.** *Let $M$ be a positive constant. Let $\Pi$ be as define in (13) such that it satisfies condition **C1**, **C2** and **C3**. Denote by $\omega_0$ the minimizer of the Kulback–Leibler divergence $KL(P_{f_{\omega,k}}, P_0)$. Then if there exists a constant $C$ such that $\Pi(\sigma_0/\sigma < C|Y^n) = o_{P_0^n}(1)$ for a constant $A > 0$ large enough, we have*

$$P_0^n \left\{ \Pi \left( \max_{j=1,\dots,k} |\omega_j - \omega_j^0| \geq A\xi_n^k | Y^n \right) \leq \frac{\gamma_1}{\gamma_0 + \gamma_1} \right\} \to 1, \tag{23}$$

*where $\xi_n^k = [\{k \log(n)\}/n]^{1/2}$ for all fixed positive $\gamma_0$ and $\gamma_1$.*

The proof of this lemma is given in the supplementary materials. We also state the following lemma that gives a control on the posterior distribution of $k$.

**Lemma 3.** *Let $k_n = n\epsilon_n^2/\log(n)$ if $L(k) = \log(k)$ and $k_n = n\epsilon_n^2$ if $L(k) = 1$ where $\epsilon_n$ is either $\epsilon_n(\mathcal{F})$ if $f_0 \in \mathcal{F}$ or $\epsilon_n(\alpha)$ if $f_0 \in \mathcal{H}(\alpha, L)$. For $C_1$ a positive constant that my depend on $K$ or $L$, let $\mathcal{K}_n = \{k \leq C_1 k_n\}$. If $\Pi$ is define as in (13) and satisfies **C1** or **C1'**, **C2** and **C3** we have*

$$\Pi\left(\mathcal{K}_n^c | Y^n\right) \leq o_{P_0^n}(1). \tag{24}$$

The proof is given in the supplementary materials.

### Proof for the test for positivity

We first prove that $\delta_n^\pi(\tau_n)$ satisfies (5) for $\tau_n = \{k \log(n)/n\}^{1/2}$. Let $f_0 \in \mathcal{F}_+$, then for all $k > 0$ we have $f_{\omega^0,k} \in \mathcal{F}_+$ which in turns gives

$$D(f_{\omega,k}, \mathcal{F}_+) = -\min(\omega) \leq \max_{j=1,\dots,k} |\omega_j - \omega_j^0|.$$

Note that if $\sigma_0 \leq \bar{\sigma}$ we get directly for $C$ large enough that $\Pi(\sigma_0/\sigma < C|Y^n) = o_{P_0^n}(1)$

Applying Lemma 2 gives us immediately (5) for $M_0$ large enough. We now show that $\delta_n^\pi(\tau_n)$ satisfies (6) with $\rho = \rho_n = M\{n/\log(n)\}^{-\alpha/(2\alpha+1)} v_n$ for $v_n$ as in Theorem 3. First note that for $f_0$ such that $f_0 \in \mathcal{H}(\alpha, L)$ $d_\infty(f, \mathcal{F}_+) > \rho_n$ we have for all $k$

$$-\min_{j=1,\dots,k}(\omega_j^0) \geq \rho_n - k^{-\alpha},$$

which leads to

$$\begin{aligned} -\min_{j=1,\dots,k}(\omega_j) &\geq -\min_{j=1,\dots,k}(\omega_j^0) - \max_{j=1,\dots,k} |\omega_j - \omega_j^0| \\ &\geq \rho_n - k^{-\alpha} - \max_{j=1,\dots,k} |\omega_j - \omega_j^0|. \end{aligned}$$

We thus deduce the following upper bound for $\Pi\{D(f, \mathcal{F}_+) \leq \tau_n | Y^n\}$:

$$\Pi\{D(f, \mathcal{F}_+) \leq \tau_n | Y^n\} \leq \Pi(\max_{j=1,\dots,k} |\omega_j - \omega_j^0| \geq \rho_n - k^{-\alpha} - \tau_n | Y^n).$$

We ends the proof by applying Lemma 2 together with Lemma 3.

**Proof for the test for monotonicity**

We first prove consistency under $H_0$. Let $f_0 \in \mathcal{F}$ then

$$D(f_{\omega,k}, \mathcal{F}_\searrow) \leq 2 \max_{i=1,\dots,k} |\omega_i - \omega_i^0|$$

and thus

$$P_0^n \left[ \Pi\{D(f_{\omega,k}, \mathcal{F}_\searrow) \geq \tau_n^k | Y_n\} < \frac{\gamma_1}{\gamma_0 + \gamma_1} \right] \to 1$$

as soon as $\tau_n^k \geq 2A\xi_n^k$, which gives the consistency under $H_0$ given Lemma 2.

We now prove consistency under $H_1$. Let $f_0 \notin \mathcal{F}$ and $f_0 \in \mathcal{H}(\alpha, L)$ we have

$$D(f_{\omega,k}, \mathcal{F}_\searrow) \geq D(f_{\omega^0,k}, \mathcal{F}_\searrow) - 2 \max_{i=1,\dots,k} |\omega_i - \omega_i^0|. \tag{25}$$

Assume that $\rho_n(\alpha) < d_\infty(f_0, \mathcal{F})$, we derive a lower bound for $D(f_{\omega^0,k}, \mathcal{F}_\searrow)$. Let $g^*$ be the monotone non-increasing piecewise constant function on the partition $\{[0, 1/k), \dots, [(k-1)/k, 1)\}$, with for $1 \leq i \leq k$, $g_i^* = \min_{j \leq i} \omega_j^0$. Given that $d_\infty(f_{\omega^0,k}, \mathcal{F}) = \inf_{g \in \mathcal{F}} d_\infty(f_{\omega^0,k}, g)$ we get

$$d_\infty(f_{\omega^0,k}, \mathcal{F}_\searrow) \leq d_\infty(f_{\omega^0,k}, g^*) \leq D(f_{\omega^0,k}, \mathcal{F}_\searrow).$$

And therefore, given that $d_\infty(f_0, \mathcal{F}_\searrow) \leq d_\infty(f_{\omega^0,k}, \mathcal{F}_\searrow) + d_\infty(f_{\omega^0,k}, f_0)$

$$\Pi\left\{D(f_{\omega,k}, \mathcal{F}_\searrow) < \tau_n^k | Y_n\right\} \leq \Pi\left\{ \max_{i=1,\dots,k} |\omega_i - \omega_i^0| \geq \frac{\rho_n(\alpha) - d_\infty(f_{\omega^0,k}, f_0) - C\tau_n^k}{4} | Y^n \right\}.$$

For $\mathcal{K}_n$ as in Lemma 3 and $k \in \mathcal{K}_n$ and $M$ large enough we have $\rho_n(\alpha)/4 > \tau_n^k$. On the set $\mathcal{K}_n$ we have for $M$, the constant in $\rho_n(\alpha)$ large enough $\rho_n(\alpha)/4 \geq d_\infty(f_{\omega^0,k}, f_0)$ which in turns gives

$$\Pi\left\{D(f_{\omega,k}, \mathcal{F}_\searrow) < \tau_n^k | Y_n\right\} \leq \Pi\left[ \left\{ \max_{i=1,\dots,k} |\omega_i - \omega_i^0| \geq \rho_n(\alpha)/8 \right\} \cap \{\mathcal{K}_n \cap B_n\} | Y^n \right] + o_{P_0^n}(1).$$

Given (23), we get that for all $f_0$ such that $d_\infty(f_0, \mathcal{F}_\searrow) > \rho_n(\alpha)$

$$P_0^n \left[ \Pi\{D(f_{\omega,k}, \mathcal{F}_\searrow) < \tau_n^k | Y_n\} < \frac{\gamma_0}{\gamma_0 + \gamma_1} \right] \to 1,$$

which ends the proof.

# 7 Discussion

In this paper we present an approach for testing un-separated hypotheses that relies on the estimation of a distance between the parameter and the null set. This approach can be viewed either as a modification of the testing loss function or as a relaxation

of the hypotheses at hand. The test obtained using this approach have been shown to be consistent and to achieve the minimax separation rates when testing parametric hypotheses and in some nonparametric settings.

The approach proposed here currently focus on two hypotheses testing, however, we believe that it could also be applied to more general settings, in particular for the sparse Gaussian sequence model. In this case Carvalho et al. (2010) proposed a model selection method for the Horseshoe prior. The idea of their approach is somehow similar to the one proposed here. Bogdan et al. (2011) and Datta and Ghosh (2013) have derived some upper bounds on the multiple testing risk from asymptotic properties of each individual test for the Horseshoe prior. We thus believe that the approach presented in this paper could also lead to interesting results in this setting. In particular one could adopt a minimax version of the risk studied in Bogdan et al. (2011) and adapt the approach studied here.

## Supplementary Material

Supplement for "Testing un-separated hypotheses by estimating a distance" (DOI: 10.1214/17-BA1059SUPP; .pdf).

## References

Akakpo, N., Balabdaoui, F., and Durot, C. (2014). "Testing monotonicity via local least concave majorants." *Bernoulli*, 20(2): 514–544. MR3178508. doi: https://doi.org/10.3150/12-BEJ496. 476, 477

Baraud, Y., Huet, S., and Laurent, B. (2003). "Adaptive tests of qualitative hypotheses." *ESAIM Probabilités Et Statistique*, 7: 147–159. MR1956076. doi: https://doi.org/10.1051/ps:2003006. 472, 475, 476, 477

Baraud, Y., Huet, S., and Laurent, B. (2005). "Testing convex hypotheses on the mean of a Gaussian vector. Application to testing qualitative hypotheses on a regression function." *The Annals of Statistics*, 33(1): 214–257. MR2157802. doi: https://doi.org/10.1214/009053604000000896. 471, 472

Berger, J. O., Boukai, B., and Wang, Y. (1997). "Unified frequentist and Bayesian testing of a precise hypothesis." *Statistical Science*, 12(3): 133–160. With comments by Dennis V. Lindley, Thomas A. Louis and David Hinkley and a rejoinder by the authors. MR1617518. doi: https://doi.org/10.1214/ss/1030037904. 461

Berger, J. O. and Delampady, M. (1987). "Testing precise hypotheses." *Statistical Science*, 2(3): 317–352. With comments and a rejoinder by the authors. MR0920141. 461, 462, 463

Berger, J. O. and Sellke, T. (1987). "Testing a point null hypothesis: irreconcilability of $P$-values and evidence." *Journal of the American Statistical Association*, 82(397): 112–139. With comments and a rejoinder by the authors. MR0883340. 461

Bernardo, J. (1980). "A Bayesian analysis of classical hypothesis testing." *Trabajos de Estadistica Y de Investigacion Operativa*, 31(1): 605–647. http://dx.doi.org/10.1007/BF02888370 461

Bogdan, M., Chakrabarti, A., Frommlet, F., and Ghosh, J. K. (2011). "Asymptotic Bayes-optimality under sparsity of some multiple testing procedures." *Annals of Statistics*, 39(3): 1551–1579. MR2850212. doi: https://doi.org/10.1214/10-AOS869. 481

Bowman, A., Jones, M., and Gijbels, I. (1998). "Testing monotonicity of regression." *Journal of computational and Graphical Statistics*, 7(4): 489–500. 472

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2010). "The horseshoe estimator for sparse signals." *Biometrika*, 97(2): 465–480. MR2650751. doi: https://doi.org/10.1093/biomet/asq017. 481

Castillo, I. and Rousseau, J. (2015). "A General Bernstein–von Mises Theorem in semiparametric models." *The Annals of Statistics*. To appear. MR3405597. doi: https://doi.org/10.1214/15-AOS1336. 469

Dass, S. C. and Lee, J. (2004). "A note on the consistency of Bayes factors for testing point null versus non-parametric alternatives." *Journal of statistical planning and inference*, 119(1): 143–152. MR2018454. doi: https://doi.org/10.1016/S0378-3758(02)00413-5. 469

Datta, J. and Ghosh, J. K. (2013). "Asymptotic properties of Bayes risk for the horseshoe prior." *Bayesian Analysis*, 8(1): 111–131. MR3036256. doi: https://doi.org/10.1214/13-BA805. 481

Dunson, D. B. and Peddada, S. D. (2008). "Bayesian nonparametric inference on stochastic ordering." *Biometrika*, 95(4): 859–874. http://biomet.oxfordjournals.org/content/95/4/859.abstract MR2461216. doi: https://doi.org/10.1093/biomet/asn043. 462

Erven, T. v., Grünwald, P., and de Rooij, S. (2012). "Catching up faster by switching sooner: a predictive approach to adaptive estimation with an application to the AIC-BIC dilemma." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(3): 361–417. MR2925369. doi: https://doi.org/10.1111/j.1467-9868.2011.01025.x. 461

Gelman, A. (2008). "Objections to Bayesian statistics." *Bayesian Analasys*, 3(3): 445–449. MR2434394. doi: https://doi.org/10.1214/08-BA318. 461

Ghosal, S., Ghosh, J. K., and Van Der Vaart, A. W. (2000a). "Convergence rates of posterior distributions." *The Annals of Statistics*, 28(2): 500–531. MR1790007. doi: https://doi.org/10.1214/aos/1016218228. 469

Ghosal, S., Sen, A., and van der Vaart, A. W. (2000b). "Testing monotonicity of regression." *The Annals of Statistics*, 28(4): 1054–1082. MR1810919. doi: https://doi.org/10.1214/aos/1015956707. 472

Ghosal, S. and van der Vaart, A. (2007). "Convergence rates of posterior distributions for non-i.i.d. observations." *The Annals of Statistics*, 35(1): 192–223. MR2332274. doi: https://doi.org/10.1214/009053606000001172. 469

Holmes, C. and Heard, N. (2003). "Generalized monotonic regression using random change points." *Statistics in Medicine*, 22(4): 623–638. 470

Ingster, Y. I. (1987). "Asymptotically minimax testing of nonparametric hypotheses." In *Probability theory and mathematical statistics, Vol. I (Vilnius, 1985)*, 553–574. VNU Sci. Press, Utrecht. MR0901514. 469, 470

Ingster, Y. I. and Suslina, I. A. (2003). *Nonparametric goodness-of-fit testing under Gaussian models*, volume 169 of *Lecture Notes in Statistics*. Springer-Verlag, New York. MR1991446. doi: https://doi.org/10.1007/978-0-387-21580-8. 469

Jeffreys, H. (1939). *Theory of Probability*. Oxford University Press, Oxford. MR0000924. 461

Johnson, V. E. (2013). "Uniformly most powerful Bayesian tests." *Ann. Statist.*, 41(4): 1716–1741. MR3127847. doi: https://doi.org/10.1214/13-AOS1123. 462, 463

Johnson, V. E. and Rossell, D. (2010). "On the use of non-local prior densities in Bayesian hypothesis tests." *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, 72(2): 143–170. MR2830762. doi: https://doi.org/10.1111/j.1467-9868.2009.00730.x. 462, 464, 469

Juditsky, A. and Nemirovski, A. (2002). "On Nonparametric Tests of Positivity/Monotonicity/Convexity." *The Annals of Statistics*, 30(2): pp. 498–527. http://www.jstor.org/stable/2699966 MR1902897. doi: https://doi.org/10.1214/aos/1021379863. 471

Lepski, O. and Tsybakov, A. B. (2000). "Asymptotically exact nonparametric hypothesis testing in sup-norm and at a fixed point." *Probability Theory and Related Fields*, 117(1): 17–48. MR1759508. doi: https://doi.org/10.1007/s004400050265. 464

Lepski, O. V. and Pouet, C. F. (2008). "Hypothesis testing under composite functions alternative." In *Topics in stochastic analysis and nonparametric estimation*, volume 145 of *IMA Vol. Math. Appl.*, 123–150. Springer, New York. MR2406269. doi: https://doi.org/10.1007/978-0-387-75111-5_7. 469

Lepski, O. V. and Spokoiny, V. G. (1999). "Minimax Nonparametric Hypothesis Testing: The Case of an Inhomogeneous Alternative." *Bernoulli*, 5(2): pp. 333–358. http://www.jstor.org/stable/3318439 MR1681702. doi: https://doi.org/10.2307/3318439. 469

Robert, C. P. (2007). *The Bayesian choice*. Springer Texts in Statistics. Springer, New York, second edition. From decision-theoretic foundations to computational implementation. MR2723361. 462, 468

Rossell, D. and Telesca, D. (2017). "Non-Local Priors for High-Dimensional Estimation." *Journal of the American Statistical Association*, 112(517): 1–33. MR3646569. doi: https://doi.org/10.1080/01621459.2015.1130634. 462, 464, 465, 466, 467

Rousseau, J. (2007). "Approximating interval hypothesis: $p$-values and Bayes factors." In *Bayesian statistics 8*, Oxford Sci. Publ., 417–452. Oxford: Oxford Univ. Press. MR2433202. 462

Rousseau, J. (2010). "Rates of convergence for the posterior distributions of mixtures of betas and adaptive nonparametric estimation of the density." *The Annals of Statistics*, 38(1): 146–180. MR2766853. doi: https://doi.org/10.1214/10-AOS811. 471

Rousseau, J. and Robert, C. (2010). "On moment priors for Bayesian model choice: a discussion." *Bayesian Statistics*, 9: 1–2. 462

Salomond, J.-B. (2017). "Supplement for "Testing un-separated hypotheses by estimating a distance"." *Bayesian Analysis*. doi: https://doi.org/10.1214/17-BA1059SUPP. 478

Scott, J. G., Shively, T. S., and Walker, S. G. (2015). "Nonparametric Bayesian testing for monotonicity." *Biometrika*, 102(3): 617–630. MR3394279. doi: https://doi.org/10.1093/biomet/asv023. 472, 474, 476, 477

Tokdar, S. T., Chakrabarti, A., and Ghosh, J. K. (2010). "Bayesian nonparametric goodness of fit tests." *Frontiers of Statistical Decision Making and Bayesian Analysis, M.-H. Chen, DK Dey, P. Mueller, D. Sun, and K. Ye, Eds*. 469

Verdinelli, I. and Wasserman, L. (1998). "Bayesian goodness-of-fit testing using infinite-dimensional exponential families." *The Annals of Statistics*, 26(4): 1215–1241. MR1647645. doi: https://doi.org/10.1214/aos/1024691240. 462

Wang, L. and Dunson, D. B. (2011). "Bayesian isotonic density regression." *Biometrika*, 98(3): 537–551. MR2836405. doi: https://doi.org/10.1093/biomet/asr025. 462

**Acknowledgments**

I am grateful to the Editor, Associate Editor and the two anonymous referees for careful reading and suggestions. I also thank Judith Rousseau and Peter Grünwald for helpful discussion.