On consistency issues in Bayesian nonparametric testing - a review

Alcune considerazioni riguardo ai test di ipotesi Bayesiani in ambito nonparametrico

Judith Rousseau

Abstract Although there have been a lot of developpements in the recent years on estimation in Bayesian nonparametric models, from a theoretical point view as well as from a methodological point of view, little has been done on Bayesian testing in nonparametric frameworks. In this talk I will be interested on asymptotic properties of Bayesian tests when at least one of the hypotheses is nonparametric. I will first give some results on goodness of fit types of tests where one is interested in testing a parametric model against a nonparametric alternative embedding the parametric model. Then I will discuss the more delicate problem where both hypotheses are nonparametric. Such cases involve in particular tests for monotonicity, two-sample tests and estimation of the number of components in nonparametric mixture models. It will be shown that the Bayes factor or equivalently the 0-1 loss function might not be appropriate in such cases and that modifications need to be considered.

Abstract *Nonostante i notevoli sviluppi nella stima di modelli nonparametrici degli ultimi anni, da un punto di vista teorico e metodologico i contributi su test di ipotesi Bayesiani in ambito nonparametrico sono stati limitati. In questa presentazione, si discutono le proprietà asintotiche dei test Bayesiani quando almeno una delle ipotesi è nonparametrica. Innanzitutto, si discutono risultati su indici di bontà statistica quando si sia interessati a testare un modello parametrico contro una alternativa nonparametrica che contiene il modello parametrico. Quindi, si discute il caso più delicato di ipotesi entrambe nonparametriche. Questi casi riguardano in particolare testi di monotonicità, test per due campioni e stima del numero di componenti in modelli mistura nonparametrici. Si mostra che il fattore di Bayes o, equivalentemente, una funzione di perdita 0-1 potrebbero non essere appropriati in tali casi e opportune modifiche si rendono necessarie.*

Key words: Bayesian nonparametrics, tests, Bayes factors, consistency ...

Judith Rousseau
Université Paris Dauphine and ENSAE-CREST, 3 avenue Pierre Larousse, 92245 Malakoff Cedex, FRANCE, e-mail: rousseau@ceremade.dauphine.fr

1 Introduction

Suppose we have two candidate models $\mathscr{M}_0$ and $\mathscr{M}_1$ for $y^n \in \mathscr{Y}^n$, a set of observations from an arbitrary distribution P^n which is absolutely continuous with respect to a commone measure μ^n on $\mathscr{Y}_n$, where n defines a known measure of information, such as the number of observations. We also assume two candidate models to have respective parameters and prior distributions, θ, $\pi_0(\theta)$, λ and $\pi_1(\lambda)$,

$$\mathscr{M}_0 = \{p^n_\theta(y^n), \theta \in \Theta,\ \pi_0(\theta)\},\ \ \mathscr{M}_1 = \{p^n_\lambda(y^n), \lambda \in \Lambda,\ \pi_1(\lambda)\}, \tag{1}$$

where $p^n_\theta(y^n)$ and $p^n_\lambda(y^n)$ denote the densities of y^n with respect to μ^n under $\mathscr{M}_0$ and $\mathscr{M}_1$ respectively.

The aim of this paper is to review some results on consistency of Bayesian testing procedures of $\mathscr{M}_0$ and $\mathscr{M}_1$, when at least one of the hypothesis corresponds to a nonparametric model. We focus mainly on the Bayes factor approach for testing, [7], which is defined as follows

$$B_{01} = \frac{p(y^n|\mathscr{M}_0)}{p(y^n|\mathscr{M}_1)} = \frac{m_0(y^n)}{m_1(y^n)} = \frac{\int p^n_\theta(y^n)\pi_0(\theta)d\theta}{\int p^n_\lambda(y^n)\pi_1(\lambda)d\lambda}. \tag{2}$$

Note that the large value of B_{01} indicates the strong evidence in support of model $\mathscr{M}_0$ ([7] and [8]). Accordingly, B_{01} is expected to converge to infinity as the sample size increases when $\mathscr{M}_0$ is the true model, and converge to 0 when the $\mathscr{M}_1$ is true. This notion is called *Bayes factor consistency* or consistency of the Bayes factor and is more precisely defined as :

$$\lim_{n\to\infty} B_{01} = \begin{cases} \infty, \text{ in } P^n_{\theta_0} \text{ probability, if } p^n_{\theta_0} \in \mathscr{M}_0 \\ 0, \text{ in } P^n_{\lambda_0} \text{ probability, if } p^n_{\lambda_0} \in \mathscr{M}_1. \end{cases} \tag{3}$$

In the i.i.d case, there some results have been obtained on Bayes factor consistency, either in specific models or providing general conditions, see [16], [3], [5], [13] and [11]. In the non i.i.d case some results have been obtained in the linear regression model, mostly in terms of variable selection, see [9], [1], [12], [15] and [2]. In [14] the authors generalize the works of [13] and [11] to the case of non i.i.d data, providing sufficient conditions for (3) to be valid when $\mathscr{M}_0$ corresponds to a parametric model and $\mathscr{M}_1$ is nonparametric and applying these conditions ot the case of the partial linear model.

In the case of embedded models, i.e. $\mathscr{M}_0$ is a submodel of $\mathscr{M}_1$ and if $\mathscr{M}_0$ is parametric, consistency under $\mathscr{M}_1$ is obtained by mainly considering posterior consistency conditions under $\mathscr{M}_1$. Consistency under $\mathscr{M}_0$ is slightly more subtil since roughly speaking $p^n_{\theta_0} \in \mathscr{M}_0 \subset \mathscr{M}_1$ so that both models are correct but the aim is to choose the smallest. Bayes factor consistency in such cases is obtained if the penalization induced by integration over the parameter $\lambda \in \Lambda$ is stronger than that induced by integration of $\theta \in \Theta$. Intuitively it is expected to happen, since Θ is fi-

nite dimensional while Λ is not. However there are cases where this is not the case. These notions are reviewed in Section 2.

The case where both $\mathscr{M}_0$ and $\mathscr{M}_1$ are nonparametric with $\mathscr{M}_0$ a submodel of $\mathscr{M}_1$ is more complicated. Consistency issues under $\mathscr{M}_1$ are similar to the former case, while they become significantly more difficult under $\mathscr{M}_0$. In section **??** we describe briefly two setups: the two sample case and the tests for monotony or more generally for some shape constraints.

2 Parametric versus non parametric

We now recall the sufficient conditions proposed in [13] and [11] and [14] and discuss cases were such conditions can fail and more importantly where consistency fails under $\mathscr{M}_0$. In this section we restrict our attntion to the case where $\mathscr{M}_0$ is parametric and is embedded in $\mathscr{M}_1$. We first recall the general conditions as provided in [14].

Let d_n be a semimetric on the set of models $\mathscr{M}_0$ and $\mathscr{M}_1$, define

$$h(\lambda_0) = \liminf_n \inf_{\theta\in\Theta} d_n(p^n_{\lambda_0}, p^n_\theta), \quad \lambda_0 \in \Lambda$$

Consistency under $\mathscr{M}_1$ corresponds to considering $\lambda_0 \in \Lambda$ such that

$$h(\lambda_0) > 0. \tag{4}$$

Define also

$$m_0(y^n) = \int_\Theta \frac{p^{(n)}_\theta(y^n)}{p^{(n)}_0(y^n)} d\pi_0(\theta), \quad m_1(y^n) = \int_\Lambda \frac{p^{(n)}_\lambda(y^n)}{p^{(n)}_0(y^n)} d\pi_1(\lambda)$$

where $p^{(n)}_0$ denotes the true distribution of the observations. Under $\mathscr{M}_1$ it is assumed that there exists $\lambda_0 \in \Lambda$ such that $p^{(n)}_0 = p^{(n)}_{\lambda_0}$ and under $\mathscr{M}_0$ it is assumed that there exists $\theta_0 \in \Theta$ such that $p^{(n)}_0 = p^{(n)}_{\theta_0}$. We first state the conditions ensuring consistency of the Bayes factor under $\mathscr{M}_1$.

Assumption A1. Let $\Lambda_0 \subset \Lambda$ satisfies : for all $\varepsilon > 0$,

$$\sup_{\lambda_0\in\Lambda_0} P^n_{\lambda_0}\left[m_1(y^n) < e^{-n\varepsilon^2}\right] = o(1)$$

Assumption A2.

(i) There exists $\varepsilon_0 > 0$ such that for all $\lambda_0 \in \Lambda_0$ and all $\varepsilon_0 > \varepsilon > 0$, there exists $\Theta_{n,\varepsilon}(\lambda_0) \subset \Theta$, such that

$$\pi_0(\Theta_{n,\varepsilon}(\lambda_0)^c) \le e^{-2n\varepsilon^2},$$

(ii)For all $\varepsilon > 0$ there exists $a > 0$ such that for all $\lambda_0 \in \Lambda_0$, there exists a sequence of tests $\phi_n(\lambda_0)$ satisfying

$$\sup_{\lambda_0 \in \Lambda_0} E^n_{\lambda_0}[\phi_n(\lambda_0)] = o(1), \quad \sup_{\lambda_0 \in \Lambda_0} \sup_{\theta \in \Theta_{n,\varepsilon}(\lambda_0)} E^n_\theta [1 - \phi_n(\lambda_0)] \leq e^{-an}.$$

Theorem 1. *Suppose that assumptions* **A1** *and* **A2** *hold, together with (4). Then for all* $\varepsilon > 0$ *there exists* $\delta > 0$ *such that*

$$\sup_{\lambda_0 \in \Lambda_0} P^n_{\lambda_0} \left[B_{01} e^{\delta n} > \varepsilon \right] = o(1)$$

That is, the Bayes factor is exponentially decreasing under $\mathscr{M}_1$ *(uniformly over* Λ_0*).*

As seen from assumptions **A1** and **A2**, obtaining consistency under $\mathscr{M}_1$ merely uses tools of posterior consistency. This is due to the fact that in such a case the true distribution is well separated from the wrong model, namely $\mathscr{M}_0$ so that it does not require much to see that from the likelihoods. This is not the case when $p_0^{(n)} \in \mathscr{M}_0$.

Let $KL(f,g)$ denote the Kullback-Leibler divergence between f and g and $V(f,g) = \int f(\log f/\log g)^2$ and consider the following assumptions:

Assumption B1. For all $\Theta_0 \subset \Theta$ compact and all $\theta_0 \in \Theta_0$, there exists $k_0 > 0$ such that

$$\inf_{\theta_0 \in \Theta_0} n^{k_0/2} \pi_0 \left[\{\theta : KL(p^n_{\theta_0}, p^n_\theta) \leq 1, \quad V(p^n_{\theta_0}, p^n_\theta) \leq 1)\} \right] \geq C$$

for some positive constants C.

Assumption B2. For all $\Theta_0 \subset \Theta$, for all $\theta_0 \in \Theta_0$, there exists $\varepsilon_n > 0$ going to 0, with $A_{\varepsilon_n}(\theta_0) = \{\lambda \in \Lambda : d_n(p^n_\lambda, p^n_{\theta_0}) < \varepsilon_n\}$ such that

B2-1.

$$\sup_{\theta_0 \in \Theta_0} P^n_{\theta_0} \left[\pi_1 \left[A^c_{\varepsilon_n}(\theta_0) | y^n \right] \right] = o(1)$$

and such that

B2-2.

$$\sup_{\theta_0 \in \Theta_0} n^{k_0/2} \pi_1 [A_{\varepsilon_n}(\theta_0)] = o(1).$$

where k_0 is the same positive constant as in Assumption **B1**.

Theorem 2. *Suppose that assumptions* **B1** *and* **B2** *hold. Let* $P^n_{\theta_0}$ *denote the joint distribution of* y^n*. Then for all compact subset* Θ_0 *of* Θ*, uniformly over* $\theta_0 \in \Theta_0$*,*

$$B_{01} \to \infty \quad \text{in } P^n_{\theta_0}\text{-probability}.$$

That is, the Bayes factor is increasing to infinity under $\mathscr{M}_0$*.*

Assumption **B2-1** only requires posterior consistency π_1, the extra condition, specific to studying the behaviour of the Bayes factor is **B2-2**. In [13], in the context of goodness of fit tests for i.i.d observations, a counter - example where consistency

does not hold under $\mathscr{M}_0$ is provided. The example corresponds to a prior π_1 such that θ_0 can be more easily approximated (i.e. a lesser number of constraints is required on the parameter) under the largest model than under the smaller model although it belongs to both. The example provided is however artificial and the set of $\theta_0 \in \Theta \subset \mathbb{R}^2$ having such pathological behaviour has null Lebesgue measure. Interestingly, in the case of partial linear model, [14] exhibits a more problematic counter- example where the subset of badly behaved θ_0 is much larger. To explain why things can go wrong recall that the partial linear model can be written as

$$y_i = \alpha + \beta x_i + f(z_i) + \varepsilon_i, \quad \varepsilon_i \sim \mathscr{N}(0, \sigma^2) \quad \text{independently}$$

where $x_i \in [-1,1]^p$, $z_i \in [-1,1]$, $\alpha \in \mathbb{R}, \beta \in \mathbb{R}^p$ and $f \in L^2([-1,1])$ and assume that the design is random satisfying $(x_i, z_i) \sim \mu$ independently, we write μ for both the joint and the marginal distributions of X, Z. Consider a hierarchical Gaussian prior on f in the form

$$f = f_\eta = \sum_{j=0}^{k} \eta_j \phi_j, \quad k \sim 1 + \mathscr{P}(\lambda); \quad \eta_j \sim \mathscr{N}(0, \tau),\, j \leq k \quad \text{i.i.d}$$

where $\mathscr{P}(\lambda)$ denotes a Poisson distribution with parameter $\lambda > 0$ and $(\phi_j)_{j\mathbb{N}}$ is the Fourier basis of $L_2([-1,1], \mu)$. For the sake of simplicity consider the case where $p = 1$, define $m(x) = E[Z_1 | X = x]$. It can be shown that if

$$\int_{-1}^{1} \phi_1(x) m(x) d\mu(x) = 0$$

then for all $\theta = (\alpha, \beta) \in \mathbb{R}^2$, the Bayes factor does not go to infinity under $p_\theta^{(n)}$. So in this case, the Bayes factor is never consistent under the null model, see [14].

Hence, when considering goodness of fit tests or more generally when testing a parametric model versus a nonparametric model, with the parametric model being a submodel of the nonparametric one, some apparently surprizingly unpleasant behaviour may happen if the prior is not carefully chosen. In the above case, the inconsistency issue can be fixed by forbidding the first r values of k for which $\int_{-1}^{1} \phi_j(x) m(x) d\mu(x) = 0$, for all $j \leq r$, the difficulty being that m is typically unknown.

When both hypothesis are nonparametric are unknown, then this becomes even more crucial. As an illustration, I describe in the following section two cases : the two sample test and the test for monotony.

3 Nonparametric versus nonparametric : a series of open problems

In the previous Section, we provide sufficient conditions to obtain consistency of the Bayes factor for testing model $\mathscr{M}_0$ versus model $\mathscr{M}_1$ with $\mathscr{M}_0$ a parametric sub-model of $\mathscr{M}_1$ and we showed that under the null hypothesis, consistency can be proved if prior mass of neighbourhoods of the true distribution under π_0 is significantly larger than that under π_1. In the case where both $\mathscr{M}_0$ and $\mathscr{M}_1$ are nonparametric with $\mathscr{M}_0 \subset \mathscr{M}_1$ (roughly speaking), such an approach can be too crude, specially if the $\mathscr{M}_0$ differs from $\mathscr{M}_1$ only by a finite number of parameters. This is for instance the case in the case of finite nonparametric mixture or HMM models, with unknown number of states. To our knowledge, if these models have been used in practice, see for instance [18] and studied in theory, see for instance [4] and [17], the problem of estimating the number of states remains an open problem so far. The problem of the two-sample test is simpler. However no asymptotic result exists on fully nonparametric Bayesian tests. [6] and [10] propose tests based on Polya trees or hierarchical versions of Polya trees but consistency is only proved under a finite dimensional version of the prior and is thus partial; [**?**] propose an interesting parametric test based on finite mixture models and prove some consistency under misspecification, but oes not cover the fully nonparametric version of the problem. Let us state the testing problem. Consider two samples $x^{(n_1)}$ and $y^{(n_2)}$ of independent and identically observations from distributions F_1 and F_2 respectively on $\mathbb{R}$, with densities f_1 and f_2 with respect to a given measure. Then the two samples test aims at choosing between the following two hypotheses:

$$\mathscr{M}_0 : F_1 = F_2, \quad \mathscr{M}_1 : F_1 \neq F_2.$$

Consider a probability distribution π on the set $\mathscr{F}$ of probability densities on $\mathbb{R}$. A bayes factor associated to π is then defined as

$$B_{0/1} = \frac{\int_{\mathscr{F}} f(x^{(n_1)}) f(y^{(n_2)}) d\pi(f)}{\int_{\mathscr{F}} f(x^{(n_1)}) d\pi(f) \int_{\mathscr{F}} f(y^{(n_2)}) d\pi(f)}.$$

To obtaining general conditions on π for the above Bayes factor to be consistent under the null hypothesis, we would need to bound from above the numerator and from below the denominator. In each case, the usual technics, based on the approach of [**?**], lead to controls of the form

$$\int_{\mathscr{F}} f(x^{(n_1)}) f(y^{(n_2)}) d\pi(f) \lesssim e^{-nc_1\varepsilon_n^2}, \quad n = n_1 + n_2$$

and

$$\int_{\mathscr{F}} f(x^{(n_1)}) d\pi(f) \int_{\mathscr{F}} f(y^{(n_2)}) d\pi(f) \gtrsim e^{-c_2(n_1\varepsilon_{n_1}^2 + n_2\varepsilon_{n_2}^2)},$$

with probability going to 1, where ε_n is some sequence converging to 0 and depending on properties of the true density f_0 and on the prior and $c_1, c_2 > 0$ are some

positive constant, whose control is not sharp and which satisfy $c_1 < c_2$. This leads to a bound on $B_{0/1}$ of the form

$$B_{0/1} \lesssim e^{n(c_2-c_1)n\varepsilon_n^2}$$

which does not go to 0. Hence a much tighter type of proof than the one used for instance to prove Theorem 2 in Section 2 is required.

Obtaining consistency under the H_1 is not so much of an issue, since then the marginal can be proved to be exponentially small, using the usual technics of proofs and under the usual types of assumptions.

Instead, in an ongoing work, we are proposing a modification of the Bayes factor, leading to a sequential approach for which consistency can be proved, both under the null and under the alternative.

Testing for monotonicity, in this respect is even more complicated, since the structure of the null space is more complex than in the two sample test.

Acknowledgements I am grateful to Michele Guindani for translating the title and abstract in Italian. This work was partially financed by Investissements d'Avenir (ANR-11-IDEX-0003/Labex Ecodec/ANR-11-LABX-0047) and by the grant ANR Bandhits ANR-2010-BLAN-0113-01.

References

1. Casella, G., Girón, F. J., Martínez, M. L. and Moreno, E., Consistency of Bayesian procedures for variable selection, Ann. Statist., **37**, 1207–1228 (2009)
2. Choi, T., Lee, J. and Roy, A. A note on the Bayes factor in a semiparametric regression model, J. Multivariate Anal. **100**, 1316–1327 (2009)
3. Dass, Sarat C. and Lee, Jaeyong, A note on the consistency of Bayes factors for testing point null versus non-parametric alternatives, J. Statist. Plann. Inference, **119**, 143–152 (2004)
4. Gassiat, E. and Rousseau, J. Non parametric finite translation hidden Markov models and extensions. http://arxiv.org/abs/1302.2345v1. (2013)
5. Ghosal, G. and Lember, J. and Van der Vaart, A., Nonparametric Bayesian model selection and averaging, Electron. J. Stat., **2**, 63–89 2008.
6. Holmes, C., Caron, F., Griffin, J., and Stephens, D. A.Two-sample Bayesian nonparametric hypothesis testing. arXiv:0910.5060 (2012).
7. Jeffreys, H.: Theory of probability,Third edition. Clarendon Press, Oxford (1961)
8. Kass, Robert E. and Raftery, A. Bayes factors, J. Amer. Statist. Assoc. **90**, 773–795 (1995)
9. Liang, F., Paulo, R., Molina, G., Clyde M. A. and Berger, J. O. Mixtures of g priors for Bayesian variable selection, J. Amer. Statist. Assoc. **103**, 410–423 (2008)
10. Ma, L. and Wong, W. H. Coupling optional Polya trees and the two sample problem. JASA, **106**, (2011)
11. McVinish, R., Rousseau, J. and Mengersen, K., Bayesian goodness of fit testing with mixtures of triangular distributions, Scand. J. Stat., **36**, 337–354 (2009).
12. Moreno, Elías and Girón, F. Javier and Casella, George, Consistency of objective Bayes factors as the model dimension grows, Ann. Statist., **38**, 1937–1952 (2010)
13. Rousseau, J. Approximating Interval hypothesis : p-values and Bayes factors. Bayesian statistics 8 (eds. Bernardo, J.M., Berger, J.O., Dawid, A.P. & Smith, A.F.M.), 417-452. Oxford University Press (USA). (2008)
14. Rousseau, J. and Choi, T. : Bayes factor consistency in non iid models (2013) Preprint.

15. Shang, Z. and Clayton, M. K., Consistency of Bayesian linear model selection with a growing number of parameters, J. Statist. Plann. Inference, **141**, 3463–3474 (2011)
16. Verdinelli, I. and Wasserman, L., Bayesian goodness-of-fit testing using infinite-dimensional exponential families, Ann. Statist., **26**, 1215–1241 (1998)
17. Vernet, E. Posterior consistency for nonparametric Hidden Markov Models with finite state space. arXiv:1311.3092.
18. Yau, C., Papaspiliopoulos, O., Roberts G. O. and Holmes, C. Bayesian non-parametric hidden Markov models with applications in genomics, Jour. Royal Statist. Soc. B, **73**, 1–21 (2011).