On rates of convergence for posterior distributions under misspecification

By HENG LIAN

182 George St, Division of Applied Mathematics, Brown University,
Providence, RI, 02912 USA.
Heng_Lian@brown.edu

SUMMARY

We extend the approach of Walker (2003, 2004) to the case of misspecified models. A sufficient condition for establishing rates of convergence is given based on a key identity involving martingales, which does not require construction of tests. We also show roughly that the result obtained by using tests can also be obtained by our approach, which demonstrates the potential wider applicability of this method.

Some key words: α-covering; Bayesian nonparametrics; Prior misspecification.

1 INTRODUCTION

Bayesian inference distinguishes itself from the frequentist school by its explicit quantification of uncertainty of the parameter with prior specification. The classical approach uses subjective priors elicited from field experts and domain knowledge. The modern Bayesian school have instead shifted attention more towards the construction of priors using formal rules in the hope of dealing with arbitrariness of the prior.

Asymptotics for infinite-dimensional Bayesian statistics has been receiving a lot of attention recently. In these studies, the Bayesian inference is approached from a frequentist point of view, that is, we assume there is a true underlying probability distribution that generates the data. Naturally one desired property is that as more and more observations are made from the underlying generating mechanism, we will obtain accurate estimate of the true distribution. While traditional Bayesians do not believe in such an assumption, it is shown by Blackwell & Dubins (1962) that this property is the same as intersubjective agreement, which means two Bayesians will eventually come to roughly the same conclusion after seeing enough data.

The posterior distribution typically behaves well under regular parametric models. Doob showed that consistency is achieved under almost no assumptions on the model, except for a zero measure set under the prior, although in topological terms this set can be large. For infinite-dimensional models, however, the matter is more subtle. Strange behavior can be observed under some priors as documented in Diaconis &

Freeman (1986). Given the prior Π on the set $\mathcal{P}$ of probability distribution, the posterior is a random measure:

$$\Pi^n(B|X_1, \ldots, X_n) = \frac{\int_B \prod_{i=1}^n p(X_i) d\Pi(P)}{\int \prod_{i=1}^n p(X_i) d\Pi(P)}$$

For ease of notation, we will omit the conditioning and only write $\Pi^n(B)$ for the posterior distribution. We say that the posterior is consistent if

$$\Pi^n(P \in \mathcal{P} : d(P, P_0) > \epsilon) \to 0 \text{ in } P_0^n \text{ probability.}$$

where P_0 is the true distribution and d is some suitable distance function between probability measures.

To study rates of convergence, let ϵ_n be a sequence decreasing to zero, we say the rate is at least ϵ_n if for sufficiently large constant M

$$\Pi^n(P : d(P, P_0) \geq M\epsilon_n) \to 0 \text{ in } P_0^n \text{ probability.}$$

We can also have a slightly weaker definition of rates of convergence by replacing M with a sequence M_n and requiring that the above posterior mass converge to zero for any sequence M_n that diverges to infinity.

On the positive side, Schwartz (1965) shows consistency for specific distributions by constructing a sequence of tests of the true distribution against distributions some positive distance away. The tests can trivially be constructed for weak neighborhoods. The construction of similar tests for stronger topology (typically measured in Hellinger distance, for example) is not so straightforward and requires extra works. Barron et al. (1999) gives sufficient conditions that guarantee consistency of infinite-dimensional models by bounding the likelihood ratio under bracketing entropy constraint on sieves. Shen & Wasserman (2001) studied rates of convergence. A related approach by constructing a sequence of tests appeared in Ghosal et al. (2000).

The conditions imposed in the above are sufficient but not necessary. It is important to see to what extent these conditions can be relaxed. Another line of work parallel to the development above by Stephen Walker and his collaborators proves consistency and rates of convergence under slightly less stringent conditions. These results are established by constructing a certain supermartingale and consistency and rates of convergence is shown by focusing on the distance of certain predictive distributions to the true one. This approach does not require construction of sequence of tests or sieves. It is shown that this new approach can lead to somewhat weaker sufficient conditions or faster rates.

In Kleijn & van der Vaart (2006), the authors consider the situation where one cannot expect to achieve consistency since the prior is misspecified. In this case, it is not surprising that the posterior will converge to the distribution in the support of the prior that is closest to the true distribution measured in Kullback-Leibler divergence. Instead of using the usual entropy number or its local version, they used a new concept called covering number for testing under misspecification and studied rates by constructing a sequence of tests between the true distribution P_0 and another measure that is not necessarily a probability distribution. The new entropy number can be reduced to the usual entropy in the well-specified case. In this paper, we study the posterior distribution also under the misspecified situation, without constructing a sequence of tests.

The goal of this paper is two fold. First, we show that the approach in Walker(2003, 2004) can be extended to the situation of misspecified prior rather straightforwardly, by introducing an α-entropy condition that is slightly stronger than that of Kleijn & van der Vaart (2006). Second, we show that using a more refined analysis, a result similar to Theorem 2.2 in Kleijn & van der Vaart (2006) can be recovered. In particular, it shows that under the well-specified case, this approach indeed is more general than the approach of constructing a sequence of tests.

In §2, we introduce necessary notations and concepts and present the martingale construction due to Walker (2003). In §3, we prove the main result and show that this approach is somehow more general than the one presented in Kleijn & van der Vaart (2006). We end this paper with a discussion in §4.

2 PRELIMINARIES

Let $\{X_1, X_2, \ldots\}$ be independent samples generated from distribution P_0, with corresponding lower case letter p_0 denoting the density with respect to some dominating measure μ. We are given a collection of distributions $\mathcal{P}$, and a prior Π on it with $\Pi(\mathcal{P}) = 1$. For simplicity, we assume that there exists a unique distribution $P^* \in \mathcal{P}$ that achieves minimum value of Kullback-Leibler divergence to the true distribution, that is

$$E_0(\log \frac{p_0}{p^*}) \le E_0(\log \frac{p_0}{p}), \text{ for all } p \in \mathcal{P}$$

where E_0 denotes the expectation under the true distribution P_0.

Let $R_n(p) = \prod_{i=1}^n p(X_i)/p^*(X_i)$, then the posterior mass for a set B is

$$\Pi^n(B) = \frac{\int_B R_n(p)\Pi(P)}{\int R_n(p)\Pi(P)} \tag{1}$$

Following Kleijn & van der Vaart (2006), for $\epsilon > 0, 0 < \alpha < 1$ and some suitable semi-metric d on $\mathcal{P}$, we define the α-covering of the set $A = \{P \in \mathcal{P} : d(P, P^*) \ge \epsilon\}$ as a collection of convex sets $\{A_1, A_2, \ldots\}$ that covers A with the additional property that for any j,

$$\inf_{P \in A_j} -\log E_0(\frac{p}{p^*})^\alpha \ge \frac{\epsilon^2}{4} \tag{2}$$

and denote by $N_t(\epsilon, \alpha, A)$ the minimum integer N such that there exists $\{A_1, \ldots, A_N\}$ that forms such a cover, if N is finite.

This condition appears to be stronger than the concept of covering for testing under misspecification introduced by Kleijn & van der Vaart (2006), which only requires that

$$\inf_{P\in A_j}\sup_{0<\alpha<1} -\log E_0(\frac{p}{p^*})^\alpha \geq \frac{\epsilon^2}{4} \tag{3}$$

In all the examples they gave in their paper, though, we can find a certain value of α only depending on the specification of the model that satisfies our condition. As shown in Kleijn & van der Vaart (2006), when $\mathcal{P}$ is convex, we have $d^2(P, P^*) \leq -\log E_0(p/p^*)^{1/2}$ where d is a generalized Hellinger distance defined by $d^2(P_1, P_2) = \frac{1}{2}\int(p_1^{1/2} - p_2^{1/2})^2 p_0/p^*\, d\mu$, which reduces to the usual Hellinger distance in the well-specified case. In this situation, the 1/2-covering for testing can be replaced by the usual covering as shown in Kleijn & van der Vaart (2006). In general, allowing α to be different than 1/2 is required, since in the misspecified case, we cannot guarantee that $-\log E_0(p/p^*)^{1/2} > 0$, and we are obliged to choose some smaller α in order to find the covering.

The predictive density constrained to a general set A is defined as

$$p_{nA}(x) = \int_A p(x)\Pi_A^n(P)$$

, where $\Pi_A^n(P) = 1_{\{P\in A\}}\Pi^n(P)/\Pi^n(A)$ is the posterior measure conditioned on A. The key identity noted by Walker (2003) is the following:

$$\int_A R_{n+1}(p)\Pi(P) = \frac{p_{nA}(X_{n+1})}{p^*(X_{n+1})}\int_A R_n(p)\Pi(P)$$

as can be verified easily. This in turn implies that

$$E_0[(\int_A R_{n+1}(p)\Pi(P))^\alpha|X_1,\ldots,X_n] = (\int_A R_n(p)\Pi(P))^\alpha E_0(\frac{p_{nA}}{p^*})^\alpha \tag{4}$$

which means that $\int_A R_n(p)\Pi(P)$ is a supermartingale when $E_0(p_{nA}/p^*)^\alpha < 1$.

3 RATES OF CONVERGENCE

To study rates of convergence, for a sequence $\epsilon_n \to 0$, we let $A_n = \{P \in \mathcal{P} : d(P, P^*) \geq M\epsilon_n\}$, and let $A_{n,j}$ be an α-covering of A_n, i.e., $\{A_{n,j}\}$ are convex sets that covers A_n and

$$\inf_{P\in A_{n,j}} -\log E_0(\frac{p}{p^*})^\alpha \geq \frac{M^2\epsilon_n^2}{4}$$

Define

$$L_{k,j}^{(n)} = \int_{A_{n,j}} R_k(p)\Pi(P) \tag{5}$$

and

$$I_n = \int_{\mathcal{P}} R_n(p)\Pi(P)$$

To obtain a lower bound for I_n, which is the denominator in (1), we also need a condition on the prior mass for a Kullback-Leibler neighborhood of p^*, which is defined as

$$B(\epsilon, P^*; P_0) = \{P \in \mathcal{P} : -E_0(\log \frac{p}{p^*}) \leq \epsilon^2, E_0(\log \frac{p}{p^*})^2 \leq \epsilon^2\}$$

Theorem 1 *Assume that P^* is the unique minimizer in $\mathcal{P}$ of the Kullback-Leibler divergence to the true distribution with $E_0(\log(p_0/p^*)) < \infty$. For a sequence ϵ_n such that $\epsilon_n \to 0$ and $n\epsilon_n^2 \to \infty$, and $A_n, A_{n,j}$ defined as above. If the following conditions hold*

1) $e^{-n\epsilon_n K} \sum_j (A_{n,j})^\alpha$ for a sufficiently large constant K

2) $\Pi(B(\epsilon_n, P^*; P_0)) \geq e^{-Ln\epsilon_n^2}$ for a sufficiently large constant L

then $\Pi^n(P : d(P, P^*) \geq M\epsilon_n) \to 0$ in P_0^n probability.

Proof. First we observe that $P_{nA_{n,j}} \in A_{n,j}$ by the convexity of $A_{n,j}$. From the definition of α-covering, $\inf_{P \in A_{n,j}} - \log E_0(p/p^*)^\alpha \geq M^2\epsilon_n^2/4$, so the predictive density satisfies

$$E_0(\frac{p_{nA_{n,j}}}{p^*})^\alpha \leq e^{-M^2\epsilon_n^2/4}$$

Taking expectations in (4), with A replaced by $A_{n,j}$, we get

$$E_0(L_{k+1,j}^{(n)})^\alpha \leq E_0(L_{k,j}^{(n)})^\alpha e^{-M^2\epsilon_n^2/4}$$

and hence

$$E_0(L_{n,j}^{(n)})^\alpha \leq e^{-nM^2\epsilon_n^2/4}(\Pi(A_{n,j}))^\alpha$$

The posterior distribution can be bounded as follows:

$$\begin{aligned} \Pi^n(A_n) &\leq \sum_j \Pi^n(A_{n,j}) \\ &\leq \sum_j [\Pi^n(A_{n,j})]^\alpha = \sum_j \frac{(L_{n,j}^{(n)})^\alpha}{I_n^\alpha} \end{aligned}$$

Lemma 7.1 in Kleijn & van der Vaart (2006) shows that when $n\epsilon_n^2 \to \infty$, for every $C > 0$, on a set Ω_n with probability converging to 1, we have $I_n \geq \Pi(B(\epsilon_n, P^*; P_0))e^{-n\epsilon_n^2(1+C)}$, so we can write

$$\begin{aligned} E_0(\Pi^n(A_n)) &= E_0(\Pi^n(A_n)1_{\Omega_n}) + E_0(\Pi^n(A_n)1_{\Omega_n^c}) \\ &\leq \frac{E_0 \sum_j (L_{n,j}^{(n)})^\alpha}{\Pi(B(\epsilon_n, P^*; P_0))^\alpha e^{-\alpha n\epsilon_n^2(1+C)}} + P_0(\Omega_n^c) \\ &\leq e^{-nM^2\epsilon_n^2/4 + \alpha n\epsilon_n^2(1+C) + \alpha n\epsilon_n^2 L} \sum_j \Pi(A_{n,j})^\alpha + P_0(\Omega_n^c) \end{aligned}$$

which converges to zero by condition 1) if M is sufficiently large. □

For a compact set of models $\mathcal{P}$, we can use the trivial bound

$$\sum_j \Pi(A_{n,j})^\alpha \leq N_t(\epsilon_n, \alpha, A_n)$$

, which gives the following result similar to Theorem 2.1 in Kleijn & van der Vaart (2006), while they used a local version of the entropy instead.

Theorem 2 *If instead of condition 1) in Theorem 1, we assume $N_t(\epsilon_n, \alpha, A_n) \leq e^{n\epsilon_n^2}$, then for sufficiently large constant M,*

$$\Pi^n(P : d(P, P^*) \geq M\epsilon_n) \to 0 \text{ in probability.}$$

In order to get optimal rate for parametric models, Kleijn & van der Vaart (2006) used a more refined assumption. In place of condition 2) in Theorem 1 above, they assumed

$$\frac{\Pi(P : J\epsilon_n < d(P, P^*) < 2J\epsilon_n^2)}{\Pi(B(\epsilon_n, P^*; P_0))} \leq e^{n\epsilon_n^2 J^2/8} \tag{6}$$

for all natural numbers n and J. In order to recover this result, we need a more careful analysis.

First, we define $A_n^J = \{P \in \mathcal{P} : M_n J\epsilon_n \leq d(P, P^*) < 2M_n J\epsilon_n\}$, with $\alpha-$covering $\{A_{n,j}^J\}$ defined similarly as before with the property: $\inf_{P\in A_{n,j}^J} -\log E_0(p/p^*)^\alpha \geq M_n^2 J^2 \epsilon_n^2/4$. Let $\tilde{A}_{n,j}^J = A_{n,j}^J \cap A_n^J$, note that $\tilde{A}_{n,j}^J$ might not be convex even though $A_{n,j}^J$ is constrained to be so. Similarly, we can define $\tilde{L}_{k,j}^{(n),J}$ as in (5) with $A_{n,j}$ replaced by $\tilde{A}_{n,j}^J$. It is easy to see that the following still holds:

$$E_0(\tilde{L}_{k+1,j}^{(n),J})^\alpha = E_0(\tilde{L}_{k,j}^{(n),J})^\alpha E_0(\frac{p_{n\tilde{A}_{n,j}^J}}{p*})^\alpha \leq E_0(\tilde{L}_{k,j}^{(n),J})^\alpha e^{-M_n^2 J^2 \epsilon_n^2/4}$$

even though $\tilde{A}_{n,j}^J$ might be nonconvex, since $P_{n\tilde{A}_{n,j}^J}$ is still contained in $A_{n,j}^J$ though not necessarily in $\tilde{A}_{n,j}^J$.

With A_n^J playing the role of A_n before, the same strategy in the proof of Theorem 1 can be followed to show that

$$\begin{aligned} E_0(\Pi^n(A_n^J) 1_{\Omega_n}) &\leq \frac{E_0 \sum_j (\tilde{L}_{n,j}^{(n),J})^\alpha}{\Pi(B(\epsilon_n, P^*; P_0))^\alpha e^{-\alpha n\epsilon_n^2(1+C)}} \\ &\leq e^{-nM_n^2 J^2 \epsilon_n^2/4 + \alpha n\epsilon^2(1+C)} \frac{\sum_j \Pi(\tilde{A}_{n,j}^J)^\alpha}{\Pi(B(\epsilon_n, P^*; P_0))^\alpha} \end{aligned}$$

We will use the notation N_t^J to denote the α-covering number for A_n^J. We are now ready to prove the following:

Theorem 3 *Assume that P^* is the unique minimizer of the Kullback-Leibler divergence to the true distribution with $E_0(\log(p_0/p^*)) < \infty$. For a sequence ϵ_n such that*

$\epsilon_n \to 0$ and $n\epsilon_n^2$ bounded away from zero, and $A_n^J, \{A_{n,j}^J\}_{j=1}^{N_t^J}$ defined as above. If the following conditions hold

1) $N_t^J \le e^{n\epsilon_n^2}$ for all $J \ge 1$

2) (6) is satisfied

Then we have

$$\Pi^n(P : d(P, P^*) \ge M_n\epsilon_n) \to 0$$

in probability for any sequence $M_n \to \infty$

Proof. We start by writing

$$\begin{aligned} E_0(\Pi^n(A_n)) &= \sum_{J=1}^{\infty} E_0(\Pi^n(A_n^J)) \\ &\le \sum_J E_0(\Pi^n(A_n^J) 1_{\Omega_n}) + P_0(\Omega_n^c) \\ &\le \sum_J e^{-nM_n^2 J^2 \epsilon_n^2/4 + \alpha n\epsilon_n^2(1+C)} \frac{\sum_j \Pi(\tilde{A}_{n,j}^J)^\alpha}{\Pi(B(\epsilon_n, P^*; P_0))^\alpha} + P_0(\Omega_n^c) \qquad (7) \end{aligned}$$

we can bound the inner sum for each fixed J as

$$\sum_j \Pi(\tilde{A}_{n,j}^J)^\alpha \le N_t^J \Pi(A_n^J)^\alpha \le e^{n\epsilon_n^2} \Pi(A_n^J)^\alpha$$

since $\tilde{A}_{n,j}^J \subset A_n^J$ and using condition 1). Plugging this into (7) and using condition 2), we get

$$E_0(\Pi^n(A_n)) \le \sum_{J\ge 1} e^{-n\epsilon_n^2 M_n^2 J^2/4 + \alpha n\epsilon_n^2(1+C) + n\epsilon_n^2 + \alpha n\epsilon_n^2 M_n^2 J^2/8} + P_0(\Omega_n^c)$$

By Lemma 7.1 of Kleijn & van der Vaart (2006), $P_0(\Omega_n^c)$ can be made arbitrarily small by choosing C sufficiently large, under the condition that $n\epsilon_n^2$ is bounded away from zero. For any C, the sum above converges to zero since $M_n \to \infty$. □

4 DISCUSSION

We demonstrated that rates of convergence of posterior distribution under misspecification can be established without construction of a sequence of tests. Theorem 3 we derived above is slightly weaker than Theorem 2.2 in Kleijn & van der Vaart (2006) due to our use of assumption (2), which is stronger than (3). This said, we are not aware of any examples where the weaker condition (3) provides any advantage over (2). In Walker (2007), the authors demonstrated that using the martingale approach can improve on the rates slightly for some problems. Theorem 3 shows that the results by Kleijn & van der Vaart (2006) is implied by our result, this is precisely true for well-specified problem, while for misspecified problem this is not conclusive due to the reason stated above. Unfortunately, we have not been able to construct an example that this approach provides a faster rate.

The extension to the case that the prior Π depends on n, and the case that there exists a finite number of points at minimal Kullback-Leibler divergence to the true distribution should be straightforward.

References

BARRON, A., SCHERVISH, M.J. & WASSERMAN, L. (1999). The consistency of posterior distributions in nonparametric problems. *Ann. Statist.* 27, 536-561.

BLACKWELL, D. & DUBINS, L. (1962). Merging of opinions with increasing information. *Ann. Math. Statist.* 33, 882-886.

DIACONIS, P. & FREEDMAN, D. (1986). On the consistency of Bayes estimates (with discussion). *Ann. Statist.* 14, 1-67.

GHOSAL, S., GHOSH, J.K. & VAN DER VAART, A.W. (2000). Convergence rates of posterior distributions. *Ann. Statist.* 28, 500-531.

KLEIJN, B.J.K. & VAN DER VAART, A.W. (2006). Misspecification in infinite-dimensional Bayesian statistics. *Ann. Statist.* 34, 837-877.

SHEN, X. & WASSERMAN, L. (2001). Rates of convergence of posterior distributions. *Ann. Statist.* 29, 687-714.

WALKER, S.G. (2003). On sufficient conditions for Bayesian consistency. *Biometrika* 90, 482-488.

WALKER, S.G. (2004). New approaches to Bayesian consistency. *Ann. Statist.* 32, 2028-2043

WALKER, S.G., LIJOI, A. & Prunster, I. (2007). On rates of convergence for posterior distributions in infinite-dimensional models. *Ann. Statist.* In press.

SCHWARTZ, L. (1965). On Bayes Procedures. *Z. Wahrsch. Verw. Gabiete* 4, 10-26