Electronic Journal of Statistics
Vol. 4 (2010) 239–253
ISSN: 1935-7524
DOI: 10.1214/09-EJS477

Posterior convergence and model estimation in Bayesian change-point problems

Heng Lian[*]

Division of Mathematical Sciences
School of Physical and Mathematical Sciences
Nanyang Technological University
Singapore, 637371
e-mail: henglian@ntu.edu.sg

Abstract: We study the posterior distribution of the Bayesian multiple change-point regression problem when the number and the locations of the change-points are unknown. While it is relatively easy to apply the general theory to obtain the $O(1/\sqrt{n})$ rate up to some logarithmic factor, showing the parametric rate of convergence of the posterior distribution requires additional work and assumptions. Additionally, we demonstrate the asymptotic normality of the segment levels under these assumptions. For inferences on the number of change-points, we show that the Bayesian approach can produce a consistent posterior estimate. Finally, we show that consistent posterior for model selection necessarily implies that the parametric rate for posterior estimation stated previously cannot be uniform over the class of models we consider. This is the Bayesian version of the same phenomenon that has been noted and studied by other authors.

AMS 2000 subject classifications: Primary 62G20; secondary 62G08.
Keywords and phrases: Change-point problems, posterior distribution, rate of convergence.

Received September 2009.

1. Introduction

We consider the regression problem of estimating a piece-wise constant function when the number of segments as well as the locations of its change-points is unknown. This is an old problem that has attracted much attention recently [14, 2, 6]. Applications of multiple change-point models surged after efficient computations using reversible jump MCMC was discovered [15]. [15] applied piece-wise constant function in the study of the coal mining disaster data in the context of Poisson process. A more recent trend of analysis that dispenses with the usage of MCMC for the change-point problem starts with the paper [23] where a dynamic programming approach is utilized to marginalize over segment levels and change-point locations. Their original motivation comes from the problem of partitioning DNA sequences into homogeneous segments. This dynamic programming approach is later extended by [5] and [21].

[*] Supported by Singapore Ministry of Education Tier 1 Grant.

Unlike the above studies, in this paper we are only concerned with the asymptotic properties of Bayesian multiple change-point problems and investigate from the frequentist view the posterior contraction characteristics of a simplified model. Although a piece-wise constant function involves only a finite number of parameters, as we will only consider the case where an upper bound on the number of change-points is available a priori, it is nevertheless best studied from a infinite-dimensional viewpoint and put the estimation problem in the context of function spaces. Until recently, little is known about the behavior of the posterior distribution of infinite-dimensional models. For consistency issues, [27] shows that the posterior is consistent when certain tests can be established for the true distribution versus the complement of its neighborhood. [1] further developed the theory by sieve construction and bracketing entropy bounds. Similar results using only metric entropy are given in [7]. Convergence rates are studied in two independent and to some extent overlapping but complementary works [10, 29]. In particular, [10] extends the idea of constructing suitable tests in order to bound the convergence rates for both nonparametric and parametric problems, and [13] further extends the approach to non-i.i.d. observations. Berstein-von Mises theorems have also been obtained in some cases [18, 3, 26]. The existence of tests for many specific problems can be found in the existing literature although sometimes new tests need to be carefully designed.

In nonparametric Bayesian analysis, we have an i.i.d. sample $Z_1, \ldots, Z_n$ from the distribution P_0 with density p_0 with respect to some measure on the sample space $(\mathcal{Z}, \mathcal{B})$. The model space is denoted by $\mathcal{P}$ which is known to contain the true distribution P_0. Given some prior distribution Π on $\mathcal{P}$, the posterior is a random measure given by

$$\Pi^n(A|Z_1, \ldots, Z_n) = \frac{\int_A \prod_{i=1}^n p(Z_i) d\Pi(P)}{\int \prod_{i=1}^n p(Z_i) d\Pi(P)} .$$

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

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

for any $\epsilon > 0$, where 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 any $M_n \to \infty$,

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

In our regression problem, we observe an i.i.d. sample $Z = (Z_1, \ldots, Z_n)$ with the distribution of $Z_i = (X_i, Y_i)$ defined structurally by

$$Y_i = \theta_0(X_i) + \epsilon_i$$

for i.i.d Gaussian noise $\epsilon_i \sim N(0, 1)$. and θ_0 is a piece-wise constant function on $[0, 1)$ with unknown locations of change-points. We can write $\theta_0(t) =$

$\sum_{j=1}^{k_0} a_j^0 I(t_{j-1}^0 \leq t < t_j^0), t_0^0 = 0 < t_1^0 < t_2^0 < \cdots < t_{k_0}^0 = 1$ using the indicator function and thus θ_0 is parameterized by (a^0, t^0), $a^0 = (a_1^0, \ldots, a_{k_0}^0), t^0 = (t_0^0, \ldots, t_{k_0}^0)$. For simplicity, we assume the marginal distributions for $\{X_i\}$ are i.i.d uniform on $[0, 1)$, and note that it is straightforward to extend all the following results to any distribution of X with density bounded away from zero and infinity. Note P_0 is fully determined by θ_0 under these assumptions, and thus we also use the space of piece-wise constant functions as our model space which is equivalent to using $\mathcal{P}$. The measure induced by θ is denoted by P_θ, and thus P_{θ_0} is the same as P_0, the true distribution.

Consistency of the above model was investigated in [19] with the exception that there X_i's are deterministically chosen on a grid. In that paper, consistency is proved for the case that the true regression function is in the Lipschitz class as well. Another related work is [28] where the Bayesian density estimation problem is studied with density approximated by piece-wise constant functions. Besides the fact that they are interested in density estimation instead of regression, the focus of that paper is very different from the current one. They are mostly concerned with the case of approximating a smooth density using step functions and aim to achieve the optimal rates up to a logarithmic factor. For density functions that are piece-wise constant, they prove parametric rate of convergence also with an extra logarithmic factor. A diverging number of grid points is used and thus this approach cannot be used to estimate the number of segments when the density is truly piece-wise constant.

In finite-dimensional problems, there exists a relatively complete answer to the question of consistency and convergence of posterior based on the set up of [16], for both regular and non-regular problems where the densities have discontinuities or singularities at certain points. A series of papers [12, 8, 9] give asymptotic approximations to the posterior distributions, with special emphasis on non-regular models including change-point problems. Other related works include [24] and [25]. All these works demonstrated qualitatively similar $1/n$ rates for change-point locations, but do not apply to our present problem where the number and locations of discontinuities are unknown.

In this paper, we focus on the case that θ_0 is piece-wise constant and aim to achieve the parametric $O(1/\sqrt{n})$ rate of convergence and also study the posterior consistency in the estimation of the number of change-points, which we refer to as the model selection problem. The proofs for the estimation rates involve direct application of general theorems in [10] but the calculation of the covering number is nontrivial in this case. In order to achieve the parametric rate, an additional assumption needs to be made to exclude functions with segments that are too short.

One simplification of our model compared with those works mentioned at the beginning of this section that focus on the computational issues is that the variance of the noise σ^2 is assumed to be known here (and actually 1 without loss of generality). This simplification is assumed in the next section for clarity of presentation. Consistency of a general regression problem with unknown noise level is addressed in [4]. In Section 3 we treat the case where the noise level is unknown but known to lie between two positive numbers and show that all the results proved for known noise variance still hold.

2. Main results

Consider the case where we have a priori bounds for the number of change-points as well as for the segment levels $\{a_j\}$. The model space, which contains the true θ_0, is defined as

$$\begin{aligned}\Theta &= \{\theta : \theta(t) = \sum_{j=1}^{k} a_j I(t_{j-1} \le t \le t_j), t_0 = 0 < t_1 < \cdots < t_k = 1, \\ &\qquad k \le k_{max}, |a_j| \le K\}.\end{aligned}$$

For identifiability, we also impose the constraint that $a_{j-1} \ne a_j, j = 1, \ldots, k$, so that θ has a unique representation. By convention, we say θ with $t_k = 1$ has k change-points, which is the same as the number of segments.

Another equivalent representation of Θ is $\Theta = \{(a, t) \in [-K, K]^k \times T_k : 1 \le k \le k_{max}\}$, where T_k is the set of $(k + 1)$-tuples $(t_0, \ldots, t_k)$ with $t_j < t_{j+1}$. We will not distinguish between these two different representations and θ can denote either a function or the tuple (a, t). This ambiguity can always be resolved by the context.

For rates of convergence, the distance d we use is the L_2 norm of the function $||\theta|| = \left(\int_0^1 \theta^2(x)dx\right)^{1/2}$. Since we only consider uniformly bounded functions, the L_2 norm is equivalent to the Hellinger distance (e.g. [13], section 7.7).

We now specify a prior on Θ using a hierarchical approach. Let Θ_k be the subspace of Θ that consists of functions with k change-points, and the prior Π is specified as a mixture

$$\Pi = \sum_{k=1}^{k_{max}} p(k)\Pi_k,\ p(k) > 0, \sum_k p(k) = 1$$

with Π_k the prior measure on Θ_k. We assume that Π_k has a density $\pi_k(\theta)$ which can be further decomposed as

$$\pi_k(a, t) = \pi_k^a(a|t)\pi_k^t(t)\,.$$

The assumption we make on the prior is that

(A) The density $\pi_k^a(a|t)$ and $\pi_k^t(t)$ are bounded away from zero and infinity on $[-K, K]^k$ and T_k respectively.

This assumption is satisfied, for example, when $t_1, t_2, \ldots, t_{k-1}$ are distributed as the order statistic of $k - 1$ points uniform distributed on $[0, 1)$ while segment levels are independent and uniformly distributed on $[-K, K]$.

The first simple result shows that the posterior rate of convergence is $n^{-1/2}$ up to a logarithmic factor.

Theorem 1. *Under assumption (A), the posterior rate of convergence is at least $\epsilon_n = \sqrt{\log n/n}$. That is, $\Pi^n(\theta : ||\theta - \theta_0|| > M\epsilon_n) \to 0$ in P_0^n probability for a sufficiently large M.*

Throughout this paper, we assume that the model space contains the true θ_0 so that the model is correctly specified. Theorem 1 considers the convergence rates of the estimation problem. A different problem is the convergence of the posterior for the number of change-points. Under no additional assumptions, we can show that the posterior probability will concentrate on the true number of change-points with probability converging to 1.

Theorem 2. *Under the same assumption (A), we have $\Pi^n(k = k_0) \to 1$ in P_0^n probability, where k_0 is the number of change-points for the true function θ_0.*

Remark 1. From the proof of the theorem in the appendix, it is clear that the under-estimation error $\Pi^n(k < k_0)$ converges to zero exponentially fast since $k < k_0$ implies the estimation error $||\theta - \theta_0||$ is bound away from zero and exponential convergence is well-known in this case [1]. However the over-estimation error $\Pi^n(k > k_0)$ only converges to zero at polynomial rate, as is clear from the proof and Theorem 1 of [29]. This difference in rates for under-estimation and over-estimation errors in Bayesian models is also present in other contexts. For example, it was shown in [17] for model selection in Markov random fields that the over-estimation error only converges at polynomial rate while the under-estimation error has an exponential decay.

Nonparametric Bayesian model selection has been investigated in [11]. The focus of that paper is on conditions under which the adaptive rates are achieved when simultaneously considering models with different rates of contraction. Thus it seems the results presented there cannot be applied to our problem.

To get rid of the extra logarithmic factor in Theorem 1, we consider the smaller model space

$$\Theta^\delta = \{\theta \in \Theta : \min |t_j - t_{j-1}| \geq \delta\}\,.$$

We can define Θ_k^δ in a similar way and assumption (A) can be modified accordingly. Theorem 1 and Theorem 2 are still true on Θ^δ with few modifications required on the proofs. In practice, specification of a prior on Θ^δ poses some difficulties or inconvenience at least. Conceptually, we can just restrict $\pi_k^t(t)$ to be supported on Θ^δ and renormalize the density. Reversible jump algorithms can be easily modified to take into account the constraint by rejecting samples violating the constraint. Dynamic programming can also incorporate the pre-specified shortest possible segment length [20, 22]. However, determination of δ in practice is problematic. This is a reflection of existing gap between Bayesian theory and practice and is similar to the fact that people will use unboundedly supported density for regression function while the parametric rate is achieved when one considers bounded functions in our present problem.

The practical advantage of avoiding short steps was also noticed by [15]. They proposed using even-numbered order statistics from $2k-1$ uniformly distributed points so that short segment lengths are better penalized.

As shown in the appendix, putting some lower bound on the segment lengths makes the local covering number bounded by a constant. This requires a very detailed argument to construct the covering. Using this more refined bound on the covering number, we can achieve the parametric rate without the extra logarithmic term.

Proposition 1. *For any $\delta > 0$, under assumption (A), the posterior rate of convergence on Θ^δ is $\epsilon_n = 1/\sqrt{n}$. That is, for every $M_n \to \infty$, we have that $\Pi^n(\theta \in \Theta^\delta : ||\theta - \theta_0|| > M_n\epsilon_n) \to 0$ in P_0^n probability.*

Combination of Theorem 2 and Proposition 1 immediately gives us the rates of convergence for the change-point locations:

Theorem 3. *Under the same assumptions as above, the posterior convergence rate for the change-point locations is at least $\epsilon_n^2 = O(1/n)$. That is, for any sequence $M_n \to \infty$, $\Pi^n(\max_{1\le j\le k_0} |t_j - t_j^0| > M_n\epsilon_n^2) \to 0$ in P_0^n probability. This rate of convergence agrees with many frequentist estimators, say using the cumulative sum approach.*

It is well-known that the posterior distribution in regular parametric models conditionally converges to a Gaussian distribution under weak conditions. Since our previous results show that the number and locations of the change-points can be consistently estimated, one would naturally conjecture that the posterior distribution for segment levels will converge to a multivariate Gaussian distribution. This is indeed the case as stated in the following theorem:

Theorem 4. *Suppose the true segment lengths are $l_j = t_j^0 - t_{j-1}^0, j = 1, \ldots, k_0$. Denote the posterior distribution of $a = (a_1, \ldots, a_k)$ restricted on the event $\{k = k_0\}$ (which has a posterior probability converging to 1) by $\Pi_{a|Z}^n$ and set the covariance matrix $I_0 = \mathrm{diag}(1/(l_j \cdot n))$. Then we have*

$$E_{Z|\theta_0}||\Pi_{a|Z}^n - N(\hat{a}(t_0), I_0)||_{TV} \to 0\,,$$

where $||P - Q||_{TV}$ is the total variation distance between probability measures P and Q, $\hat{a}(t_0)$ is the maximum likelihood estimator for a, assuming the true locations of the change-points are known.

The above theorems show that the Bayesian procedure possesses very good properties. On the one hand, the parametric rate is achieved for the estimation problem in the function space. On the other hand, the number of change-points can be consistently estimated. However, Theorem 2 and Proposition 1 only apply to a fixed true piece-wise constant function and thus the convergence as stated is point-wise in nature in this sense. It is not difficult to see from the proof of Proposition 1 that the $1/\sqrt{n}$ rate is not actually uniform over the model space Θ^δ. The reason is that in order to obtain the bound for the local covering number (Lemma 4.4 in the appendix), the constant involved does depend on θ_0. In particular, the derivation of the lemma requires a lower bound on the size of the jumps of the neighboring segments and thus the convergence is not uniform over Θ^δ. Intuitively, small jumps makes the estimation more difficult and heavier penalization by the prior must be incorporated (possibly by using a prior that depends on the sample size) to achieve model selection consistency at the cost of losing estimation accuracy. As seen in the proof of Theorem 5, the difficulty occurs when the size of the jump is of order $O(1/\sqrt{n})$, in which case there appears a conflict between change-point detection and efficient estimation.

Nevertheless, as discussed above, the convergence is uniform if we further restrict our attention to the sub-class:

$$\Theta^{\delta_1,\delta_2} = \{\theta \in \Theta^{\delta_1}, \min |a_j - a_{j-1}| \geq \delta_2\}.$$

We state the uniform convergence as a proposition without proof:

Proposition 2. *For any fixed $\delta_1, \delta_2 > 0$, the rate of convergence is uniformly at least $\epsilon_n = O(1/\sqrt{n})$. That is, for any $M_n \to \infty$, $\sup_{\bar{\theta}\in\Theta^{\delta_1,\delta_2}} E_{Z|\bar{\theta}}\Pi^n(\theta \in \Theta^{\delta_1,\delta_2} : ||\theta - \bar{\theta}|| > M_n\epsilon_n) \to 0$. The property of model selection consistency is still satisfied in this case.*

On the other hand, the following result confirms that we cannot expect the posterior to converge uniformly over the class Θ^δ if the method can adapt to the number of change-points. Note that the theorem applies for any Bayesian posterior distribution for the change-point problem, not just the specific prior we constructed.

Theorem 5. *Suppose the posterior distribution satisfies the model consistency condition: $\Pi^n(k = k_0) \to 1$ in P_0^n probability, then the maximal L_2 convergence of θ is necessarily slower than the parametric rate $\epsilon_n = O(1/\sqrt{n})$. That is, for some $M_n \to \infty$,*

$$\sup_{\bar{\theta}\in\Theta^\delta} E_{Z|\bar{\theta}}\Pi^n(\theta \in \Theta^\delta : ||\theta - \bar{\theta}|| > M_n\epsilon_n) \to 1\,.$$

The above theorem demonstrated the trade-off between function estimation and model selection for our Bayesian multiple change-point problems.

3. Unknown noise variance

Here we briefly discuss the case where the variance of the noise in the regression problem is unknown. Thus we will extend our previous notation and use $P_{\theta,\sigma}$ to denote the measure induced by θ and σ, with $p_{\theta,\sigma}$ denoting the corresponding density. We only deal with the case where σ is known to be bounded by two positive values, that is $\sigma \in [b_1, b_2], b_2 > b_1 > 0$, and we impose a prior distribution on σ with density bounded away from zero and infinity, independent of the prior on θ. We will consider the convergence rate in terms of $\max\{||\theta - \theta_0||, |\sigma - \sigma_0|\}$ where σ_0 is the true unknown noise standard deviation. In this simple situation, we can show that the Hellinger distance $h(P_{\theta_1,\sigma_1}, P_{\theta_2,\sigma_2})$ between P_{θ_1,σ_1} and P_{θ_2,σ_2} is equivalent to $\max\{||\theta_1 - \theta_2||^2, |\sigma_1 - \sigma_2|^2\}$ and the Kullback-Leibler divergence of P_{θ_0,σ_0} and $P_{\theta,\sigma}$ as well as the second moment of $\log p_{\theta,\sigma}/p_{\theta_0,\sigma_0}$ is bounded by a multiple of $\max\{||\theta - \theta_0||^2, |\sigma - \sigma_0|^2\}$. These facts are stated in the following proposition. Its proof involves relatively simple although somewhat cumbersome calculations and is presented at the author's website (`https://edventure.ntu.edu.sg/bbcswebdav/users/henglian/onlineprop.pdf`).

Proposition 3. *Denote* $d^2((\theta_1,\sigma_1),(\theta_2,\sigma_2)) = \max\{||\theta_1-\theta_2||^2, |\sigma_1-\sigma_2|^2\}$. *Then*

$$\begin{aligned} P_{\theta_0,\sigma_0}\log\frac{p_{\theta_0,\sigma_0}}{p_{\theta,\sigma}} &\lesssim d^2((\theta,\sigma),(\theta_0,\sigma_0)), \\ P_{\theta_0,\sigma_0}\left(\log\frac{p_{\theta_0,\sigma_0}}{p_{\theta,\sigma}}\right)^2 &\lesssim d^2((\theta,\sigma),(\theta_0,\sigma_0)), \\ d((\theta_1,\sigma_1),(\theta_2,\sigma_2)) &\lesssim h(P_{\theta_1,\sigma_1},P_{\theta_2,\sigma_2}) \lesssim d((\theta_1,\sigma_1),(\theta_2,\sigma_2)), \end{aligned}$$

where $a \lesssim b$ *means* $a \le Cb$ *for some constant* $C > 0$.

Using the proposition, it is easy to see that all the results in the previous section still hold for the case of unknown noise variance. In particular, the covering number calculations in Lemmas 4.3 and 4.4 are still valid although the constants involved become larger. For prior concentration results in Lemmas 4.1 and 4.2, the only change is an additional factor of ϵ in the bounds, coming from the constraint $|\sigma-\sigma_0| < \epsilon$. Theorem 2 on model selection consistency can be shown along the same lines with only minor modifications. Theorems 4 and 5 only depend on the root-n convergence rate of θ and on Theorem 2 respectively and are thus still valid.

4. Discussion

In this paper, we investigated in detail some asymptotic properties of Bayesian multiple change-point problems when the noise level is assumed known. We proved estimation rate of convergence as well as model selection consistency of the posterior distribution. For simplicity, we only treated the case with random covariate X, but the results also applies to the case with deterministic covariates using the corresponding theory in [13].

The main contribution of the paper is to show that the parametric rate is achieved for a restricted class of piece-wise constant functions, that the posterior distributions of segment levels are asymptotically normal, and that the optimal rate cannot be achieved uniformly over the class.

Our theory still leaves some gaps in between. For example, it is still unknown whether it is absolutely necessary to restrict the functions to have not too short segment lengths in order to achieve the optimal rate. This additional restriction makes the local covering number bounded in order to apply Theorem 2.4 in [10]. Besides, as pointed out by the associate editor during the review process, it is natural to expect that one can relax the condition in Proposition 2 to $\delta_1 \gg n^{-1}$ and $\delta_2 \gg n^{-1/2}$. However, the calculation of covering number seems to be technically challenging and our current method used in Lemma 4.4 cannot produce the desired covering number bound.

Acknowledgements

The author wants to thank an Associate Editor and a reviewer for their comments that greatly improved the presentation of this paper as well as for the suggestion to study the case with unknown noise variance which is not touched upon in the first version of the paper. This research is funded by the Singapore Ministry of Education Tier 1 Grant.

Appendix

Some Lemmas

In preparation for the proofs of the main results, we first collect some lemmas here. The constant C is used to denote a generic constant which might not be the same at different places. Note that since we are only considering uniformly bounded class of functions, the Hellinger distance, the Kullback-Leibler divergence, as well as the second moment of the likelihood ratio are all equivalent to the L_2 norm of the regression function. In the following, we set $\delta_0 = \min\{\min_j |t_j^0 - t_{j-1}^0|, \min_j |a_j^0 - a_{j-1}^0|\} > 0$, which bounds the segment lengths as well as the jump size from below.

Lemma 4.1. *Under condition (A), we have the lower bound for the prior concentration when $\epsilon_n \to 0$,*

$$\Pi(\theta : ||\theta - \theta_0|| \leq \epsilon_n) \geq Cp(k_0)\epsilon_n^{3k_0-2},$$

where C is a constant depending on the lower bound of the prior density in a neighborhood of θ_0.

Proof. When $\theta = \sum_{j=1}^{k_0} a_j I(t_{j-1} \leq t < t_j) \in \Theta_{k_0}$ with $|a_j - a_j^0| < \epsilon_n/2, 1 \leq j \leq k_0$ and $|t_j - t_j^0| < \frac{\epsilon_n^2}{8K^2 k_{max}}, 1 \leq j \leq k_0 - 1$, it is easy to show that $||\theta - \theta_0||^2 < \epsilon_n^2$. Since the prior density for (a, t) is bounded away from zero, we get

$$\Pi(\theta : ||\theta - \theta_0|| \leq \epsilon_n) \geq p(k_0)\Pi_{k_0}(\theta : ||\theta - \theta_0|| \leq \epsilon_n) \geq Cp(k_0)\epsilon_n^{3k_0-2}.$$

□

Lemma 4.2. *Let $\delta' = \sqrt{\frac{\delta_0^3}{4k_{max}}}$ (δ_0 is defined immediately before Lemma 4.1). When $\epsilon < \delta'$, we have that $\Pi_k(\theta \in \Theta_k : ||\theta - \theta_0|| \leq \epsilon) \leq C\epsilon^{3k_0-2}, k = 1, \ldots, k_{max}$, where C is a constant that depends on K, k_{max} and δ_0. For $k > k_0$, the bound can be refined to $C\epsilon^{3k_0-1}$.*

Proof. First we consider the case $k < k_0$ and $\theta \in \Theta_k$. By the definition of δ_0, the $k_0 - 1$ intervals $(t_j^0 - \delta_0/2, t_j^0 + \delta_0/2), j = 1, \ldots k_0 - 1$ are nonoverlapping. Thus there is at least one segment of θ that contains one of these $k_0 - 1$ intervals. Thus the distance between θ and θ_0 is at least $\sqrt{\delta_0(\delta_0/2)^2} \geq \delta'$, and thus $\Pi_k(\theta \in \Theta_k : ||\theta - \theta_0|| \leq \epsilon) = 0$.

When $k \geq k_0, \theta \in \Theta_k$ and $||\theta - \theta_0|| < \epsilon$, for any j, let $s(j)$ be the index of the interval $[t_{s(j)-1}, t_{s(j)})$ which has the largest overlap with $[t^0_{j-1}, t^0_j)$. Obviously the length of the overlap is at least δ_0/k_{max}. This implies $|a_{s(j)} - a^0_j| \leq \epsilon\sqrt{\frac{k_{max}}{\delta_0}}$ (otherwise the squared L_2 distance between θ and θ_0 is at least $(a_{s(j)} - a^0_j)^2 \frac{\delta_0}{k_{max}} > \epsilon^2$). Similarly, let $t(j)$ be the index of the change-point of θ that is closest to t^0_j, we have $|t_{t(j)} - t^0_j| \leq \frac{4\epsilon^2}{\delta_0^2}$ (otherwise the squared distance will be bigger than $\frac{4\epsilon^2}{\delta_0^2}(\frac{\delta_0}{2})^2 = \epsilon^2$). The above considerations give us k_0 constraints on the segments levels of θ as well as $k_0 - 1$ constraints on the change-point locations. Thus under assumption (A), the prior probability $\Pi_k(\theta \in \Theta_k : ||\theta - \theta_0|| \leq \epsilon)$ is bounded by $C\epsilon^{3k_0-2}$. For the refined bound when $k > k_0$, we only consider $k = k_0 + 1$ for simplicity. In this case, $||\theta - \theta_0|| \leq \epsilon$ implies an additional restriction $(a^0_{k_0} - a_{k_0+1})^2(1 - t_{k_0}) \leq \epsilon^2$. This gives us an additional factor of ϵ in the bound. □

Lemma 4.3. *$\log D(\epsilon, \Theta) \leq b\log(1/\epsilon) + c$, for some constants $b, c > 0$ that depends on K and k_{max}, where $D(\epsilon, \Theta)$ is the $\epsilon-$covering number of Θ, defined as the minimal number of balls of radius ϵ needed to cover Θ and the metric used is the L_2 distance.*

Proof. Choose a grid on the domain $[0, 1)$ and another grid on $[-K, K]$

$$\Delta_t = \left\{\frac{\epsilon^2}{8K^2k_{max}} \cdot i, i \in \mathbb{N}\right\} \cap [0, 1],\ \Delta_y = \{\epsilon \cdot i, i \in \mathbb{Z}\} \cap [-K, K]\,.$$

Let $\tilde{\Theta} = \{\theta \in \Theta, \theta$ jumps only at points in Δ_t and takes segment levels in $\Delta_y\}$. It is then easy to show that $\tilde{\Theta}$ is an $\epsilon-$covering of Θ with covering number bounded by

$$\binom{\lfloor \frac{8K^2k_{max}}{\epsilon^2} \rfloor + 1}{k_{\max}} \left(\frac{2K}{\epsilon} + 1\right)^{k_{max}}.$$

□

Lemma 4.4. *For $2\epsilon < \delta' = \sqrt{\frac{\delta_0^3}{4k_{max}}}$,*

$$\log D(\epsilon/2, \{\theta \in \Theta^\delta, \epsilon \leq ||\theta - \theta_0|| \leq 2\epsilon\}) \leq C$$

for some constant C that depends on δ, δ_0, K and k_{max} but does not depend on ϵ.

Proof. Suppose that $||\theta - \theta_0|| \leq 2\epsilon$. From the proof of Lemma 4.2, we know that each change-point of θ_0 has a corresponding change-point of θ that satisfies $|t_{t(j)} - t^0_j| \leq 16\epsilon^2/\delta_0^2$. For any segment level a^0_j of θ_0, denote the corresponding index of the segment of θ that has an overlap of at least $\delta/2$ by $r(j)$, by a similar argument as in Lemma 4.2, $|a_j - a^0_{r(j)}| \leq 2\sqrt{2}\epsilon/\sqrt{\delta}$.

To construct a covering, we partition $[0, 1)$ into nonoverlapping intervals. In the following, M, B, N are sufficiently large integers to be chosen later. First, each interval $[t^0_j - 16\epsilon^2/\delta_0^2, t^0_j + 16\epsilon^2/\delta_0^2]$ is partitioned into M subintervals with equal lengths. For the rest of $[0, 1)$ we partition it into segments of lengths between $\delta/2B$ and δ/B. Obviously the total number of subintervals does not depend on ϵ. These subintervals falls into two types: (i) the subinterval that contains some change-point of θ_0; (ii) the subinterval that is entirely contained in some segment of θ_0. The function class F that forms a covering is defined as the set of functions which is piece-wise constant with respect to the partition, takes a value of 0 on type (i) subintervals and takes values of the form $a_j^0 + \frac{2\sqrt{2}\epsilon}{N\sqrt{\delta}} i$, $i = -N, -(N-1), \dots, N$, on type (ii) subintervals if the subinterval is contained in segment j of θ_0. The size of F is a constant independent of ϵ and we show next that it is indeed a $\epsilon/2$-covering.

On subintervals of type (i), the squared L_2 distance between F and θ restricted on these intervals are at most $\frac{16\epsilon^2}{M\delta^2} k_{max} K^2$. Type (ii) subintervals can be further divided into three types: (iii) it contains a change-point of θ which is closest to some change-point of θ_0; (iv) it contains a change-point of θ other than those closest to some change-point of θ_0; (v) it is entirely contained in some segment of θ. On subintervals of type (iii) the squared distance is at most $\frac{16\epsilon^2}{M\delta^2} k_{max} K^2$. On subintervals of type (iv) the squared distance is at most $\frac{\delta}{B} k_{max} (\frac{2\sqrt{2}\epsilon}{\sqrt{\delta}})^2$. On subintervals of type (v) the squared distance is at most $(\frac{2\sqrt{2}\epsilon}{N\sqrt{\delta}})^2$. Thus when M, B, N is large enough, we have a $\epsilon/2$-covering. □

Proofs of the main results

Proof of Theorem 1. We apply Theorem 2.1 in [10] with $\epsilon_n = C\sqrt{\log n/n}$. Condition (2.2) for that theorem is verified in Lemma 4.3, condition (2.3) is trivially satisfied and condition (2.4) is verified in Lemma 4.1. □

Proof of Theorem 2. Theorem 1 immediately implies that the under-estimation probablity $\Pi^n(k < k_0) \to 0$ in P_0^n probability. For over-estimation, it is sufficient to show that $P_0^n(\int_{\Theta_{k_0}} \frac{p_\theta^n(Z)}{p_0^n(Z)} d\pi_{k_0}(\theta) < Cn^{-(3k_0-2+2\xi)/2}) \to 0$ for some $0 < \xi < 1/2$, and $P_0^n(\int_{\Theta_k} \frac{p_\theta^n(Z)}{p_0^n(Z)} d\pi_k(\theta) > (\log n)^{-1} n^{-(3k_0-2+2\xi)/2}) \to 0$, when $k > k_0$, since these two statements together imply that the posterior mass on $k > k_0$ is asymptotically ignorable compared to $k = k_0$.

step 1. Let $U_n = \{t \in T_{k_0} : t = t^0 + u, u \in R^{k_0+1}, u_0 = u_{k_0} = 0, |u_i| < c/n\}$ with $\Pi_{k_0}^t(U_n) \geq c' n^{-k_0+1}$, where $\Pi_{k_0}^t$ is the prior measure on the locations of change-points. For any fixed $t \in U_n$, with probability converging to 1, by considering a small neighborhood of the maximum likelihood estimator $\hat{a}(t)$ for the given t as in Laplace approximation, we have

$$\int \frac{p_\theta^n(Z)}{p_0^n(Z)} d\pi_{k_0}(a|t) \geq \frac{C}{n^{k_0/2}} \frac{p_{(\hat{a}(t),t)}^n(Z)}{p_0^n(Z)} \geq \frac{C}{n^{k_0/2}} \frac{p_{(a_0,t)}^n(Z)}{p_0^n(Z)} .$$

For any $t \in U_n$, and conditional on $\{X_i\}$, $\log \frac{p_{(a_0,t)}^n(Z)}{p_0^n(Z)}$ is normally distributed with mean $-\frac{1}{2} f(t)^2$ and variance $f(t)^2$, where $f(t)^2 = \sum_{j=1}^{k_0-1} (a_{j+1}^0 - a_j^0)^2 \cdot n_j$, and n_j is the number of X_i that falls into the subinterval $[t_j^0, t_j)$ or $[t_j, t_j^0)$ (depending on the sign of u_j). Since n_j is Binomial distributed with mean less than

c, $f(t)^2 = O_p(\xi \log n)$ and thus $\log \frac{p^n_{(a_0,t)}(Z)}{p_0(Z)} \geq -\xi \log n$ with probability converging to 1. Thus, with probability converging to 1, we have $\int_{\Theta_{k_0}} \frac{p^n_\theta(Z)}{p^n_0(Z)} d\pi_{k_0}(\theta) \geq \Pi^t_{k_0}(U_n) \frac{C}{n^{k_0/2}} n^{-\xi} = Cn^{-(3k_0-2+2\xi)/2}$.

step 2. Letting $\delta_n = \frac{1}{2\log n n^{(3k_0-2(1-\xi))/2}}$, and $\epsilon_n = C\log n/\sqrt{n}$, we have that

$$\begin{aligned} & P^n_0\left(\int_{\Theta_k} \frac{p^n_\theta(Z)}{p^n_0(Z)} d\pi_k(\theta) > (\log n)^{-1} n^{-(3k_0-2+2\xi)/2}\right) \\ \leq \quad & P^n_0\left(\int_{\{||\theta-\theta_0||\leq\epsilon_n\}\cap\Theta_k} \frac{p^n_\theta(Z)}{p^n_0(Z)} d\pi_k(\theta) > \delta_n\right) \\ & \quad + P^n_0\left(\int_{\{||\theta-\theta_0||>\epsilon_n\}\cap\Theta_k} \frac{p^n_\theta(Z)}{p^n_0(Z)} d\pi_k(\theta) > \delta_n\right). \end{aligned}$$

By the Markov inequality and Fubini's theorem, the first term above is bounded by $\frac{1}{\delta_n}\pi_k(||\theta-\theta_0|| \leq \epsilon_n) \leq \frac{1}{\delta_n}\epsilon_n^{3k_0-1} \to 0$, where we have made use of Lemma 4.2.

For the second term, we apply Theorem 1 of [29] with ϵ in that theorem replaced by ϵ_n defined above. Using

$$\int_{\epsilon^2/2^8}^{\sqrt{2}\epsilon} \sqrt{b\log(1/\epsilon)+c} \leq \sqrt{b\log\frac{2^8}{\epsilon^2}+c} \cdot \sqrt{2}\epsilon\,,$$

the entropy condition in that theorem can be verified for $\epsilon = \epsilon_n$. Thus when C is large enough, the second term also converges to 0. □

Proof of Proposition 1. We apply Theorem 2.4 in [10] using $\epsilon_n = A/\sqrt{n}$ with A sufficiently large. Condition (2.7) for that theorem is verified in Lemma 4.4, and (2.8) is trivially satisfied. Now we verify (2.9), for which we need to bound $\frac{\Pi(j\epsilon_n \leq ||\theta-\theta_0|| \leq 2j\epsilon_n)}{\Pi(||\theta-\theta_0|| \leq \epsilon_n)}$. When $j < \delta'\sqrt{n}/2A$, $2j\epsilon_n < \delta'$ and Lemma 4.2 can be directly applied to obtain that $\Pi_k(\theta \in \Theta^\delta_k : ||\theta-\theta_0|| \leq 2j\epsilon_n) \leq C(j\epsilon_n)^{3k_0-2}$, and we get $\frac{\Pi(j\epsilon_n \leq ||\theta-\theta_0|| \leq 2j\epsilon_n)}{\Pi(||\theta-\theta_0|| \leq \epsilon_n)} \leq Cj^{3k_0-2} \leq Cexp(A^2j^2/2)$. For $j \geq \delta'\sqrt{n}/2A$, we bound the numerator by 1, and $\frac{\Pi(j\epsilon_n \leq ||\theta-\theta_0|| \leq 2j\epsilon_n)}{\Pi(||\theta-\theta_0|| \leq \epsilon_n)} \leq C(1/\epsilon_n)^{3k_0-2} \leq Cexp(A^2j^2/2)$ for this range of j. □

Proof of Theorem 3. By Theorem 2, we can assume the number of change-points of θ is also k_0. Then $\max_{1\leq i\leq k_0}|t_i - t^0_i| > M_n\epsilon^2_n$ implies that $||\theta-\theta_0||^2 > (\delta_0/2)^2 M_n\epsilon^2_n$. Thus $\Pi_n(\max_{1\leq i\leq k_0}|t_i - t^0_i| > M_n\epsilon^2_n) \leq \Pi_n(\theta \in \Theta^\delta : ||\theta-\theta_0|| > (\delta_0/2)\sqrt{M_n}\epsilon_n) \to 0$. □

Proof of Theorem 4. Fixing one $t \in T^C_{k_0} = \{t \in T_{k_0} : \max_j |t_j - t^0_j| \leq C/n\}$, denote the maximum likelihood estimator for a by $\hat{a}(t)$. Let $\Pi^n_{a|t,Z}$ and $\Pi^n_{t|Z}$ be the posterior measure for a conditioning on t and the posterior measure for t respectively. The classical Bernstein-von Mises Theorem implies that $E_0||\Pi^n_{a|t,Z} -$

$N(\hat{a}(t), I_0)||_{TV} \to 0$ ([30], Theorem 10.1). We have that

$$\begin{aligned} & E_0||\Pi^n_{a|Z} - N(\hat{a}(t_0), I_0)||_{TV} \\ \leq\ & ||\int_{T^C_{k_0}} \Pi^n_{a|t,Z} d\Pi^n_{t|Z} - N(\hat{a}(t_0), I_0)||_{TV} \\ & \quad + ||\int_{(T^C_{k_0})^c} \Pi^n_{a|t,Z} d\Pi^n_{t|Z} - N(\hat{a}(t_0), I_0)||_{TV} \\ =\ & (I) + (II). \end{aligned}$$

(I) can be bounded by

$$\begin{aligned} & E_0||\int_{T^C_{k_0}} \Pi^n_{a|t,Z} d\Pi^n_{t|Z} - N(\hat{a}(t_0), I_0)||_{TV} \\ \leq\ & E_0 \left[\int_{T^C_{k_0}} ||\Pi^n_{a|t,Z} - N(\hat{a}(t), I_0)||_{TV} d\Pi^n_{t|Z} \right] \\ & \quad + E_0 \left[\int_{T^C_{k_0}} ||N(\hat{a}(t), I_0) - N(\hat{a}(t_0), I_0)||_{TV} d\Pi^n_{t|Z} \right]. \end{aligned}$$

The first term converges to zero by the boundedness of the TV norm and the Fubini's theorem. The second term converges to zero since $||\hat{a}(t) - \hat{a}(t_0)|| = o_p(1/\sqrt{n})$. Letting n goes to infinity for a fixed C first, we see that $\limsup_n E_0 \times ||\Pi^n_{a|Z} - N(\hat{a}(t_0), I_0)||_{TV}$ is upper bounded by expression (II), which can be made arbitrarily small when C is big enough by Theorem 3 as well as the fact that TV-norm is bounded. Thus letting n goes to infinity an then C goes infinity, we see that $E_0||\Pi^n_{a|Z} - N(\hat{a}(t_0), I_0)||_{TV} \to 0$. □

Proof of Theorem 5. Fix any number $M > 0$ and $\gamma > 2M$. Define $\theta_0 = 0$ and $\theta_n = \frac{\gamma}{\sqrt{n}} I(\frac{1}{2} \leq t < 1)$, a function with a single change-point and jump size $\frac{\gamma}{\sqrt{n}}$. We trivially have $||\theta - \theta_n|| \geq \frac{\gamma}{2\sqrt{n}} > \frac{M}{\sqrt{n}}$ for all $\theta \in \Theta_1$ (i.e. θ is a constant function). Under θ_0, the posterior probability on Θ_1 converges to 1 by Theorem 2. This gives us

$$\begin{aligned} E_{Z|\theta_0}\Pi^n(\theta : ||\theta - \theta_n|| > M/\sqrt{n}) &\geq E_{Z|\theta_0}\Pi^n(\theta : ||\theta - \theta_n|| > M/\sqrt{n}, \theta \in \Theta_1) \\ &\geq E_{Z|\theta_0}\Pi^n(\Theta_1) \to 1. \end{aligned}$$

Since the measure P^n_0 induced by θ_0 and the measure $P^n_{\theta_n}$ induced by θ_n are mutually contiguous (this is a straightforward extension of Theorem 7.2 in [30]), we have

$$\sup_{\bar{\theta} \in \Theta^\delta} E_{Z|\bar{\theta}}\Pi^n(\theta \in \Theta^\delta : ||\theta - \bar{\theta}|| > M\epsilon_n) \geq E_{Z|\theta_n}\Pi^n(\theta \in \Theta^\delta : ||\theta - \theta_n|| > M\epsilon_n) \to 1.$$

Since this is true for any M, it is also true for some slowly diverging sequence M_n as in the statement of the theorem. □

References

[1] Barron, A., Schervish, M. J., and Wasserman, L. The consistency of posterior distributions in nonparametric problems. *Annals of Statistics*, 27(2):536–561, 1999. MR1714718

[2] S. Ben Hariz, Wylie, J. J., and Zhang, Q. Optimal rate of convergence for nonparametric change-point estimators for nonstationary sequences. *Annals of Statistics*, 35(4):1802–1826, 2007. MR2351106

[3] Castillo, I. A semi-parametric Bernstein-von Mises theorem. *Preprint*, 2009.

[4] Choi, T. and Schervish, M. J. On posterior consistency in nonparametric regression problems. *Journal of Multivariate Analysis*, 98(10):1969–1987, 2007. MR2396949

[5] Fearnhead, P. Exact and efficient bayesian inference for multiple changepoint problems. *Statistics and Computing*, 16(2):203–213, 2006. MR2227396

[6] Fearnhead, P. Computational methods for complex stochastic systems: a review of some alternatives to mcmc. *Statistics and Computing*, 18(2):151–171, 2008. MR2390816

[7] Ghosal, S., Ghosh, J. K., and Ramamoorthi, R. V. Posterior consistency of Dirichlet mixtures in density estimation. *Annals of Statistics*, 27(1):143–158, 1999. MR1701105

[8] Ghosal, S., Ghosh, J. K., and Samanta, T. On convergence of posterior distributions. *Annals of Statistics*, 23(6):2145–2152, 1995. MR1389869

[9] Ghosal, S., Ghosh, J. K., and Samanta, T. Approximation of the posterior distribution in a change-point problem. *Annals of the Institute of Statistical Mathematics*, 51(3):479–497, 1999. MR1722841

[10] Ghosal, S., Ghosh, J. K., and Van der Vaart, A. W. Convergence rates of posterior distributions. *Annals of Statistics*, 28(2):500–531, 2000. MR1790007

[11] Ghosal, S., Lember, J., and Van Der Vaart, A. Nonparametric Bayesian model selection and averaging. *Electronic Journal of Statistics*, 2:63–89, 2008. MR2386086

[12] Ghosal, S. and Samanta, T. Asymptotic behaviour of Bayes estimates and posterior distribution in multiparameter nonregular cases. *Mathematical Methods of Statistics*, 4(4):361–388, 1995. MR1372011

[13] Ghosal, S. and Van Der Vaart, A. Convergence rates of posterior distributions for noniid observations. *Annals of Statistics*, 35(1):192–223, 2007. MR2332274

[14] Goldenshluger, A., Tsybakov, A., and Zeevi, A. Optimal change-point estimation from indirect observations. *Annals of Statistics*, 34(1):350–372, 2006. MR2275245

[15] Green, P. J. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(4):711–732, 1995. MR1380810

[16] Ibragimov, I. A. and Khasminskii, R. Z. *Statistical estimation–asymptotic theory*. Applications of mathematics. Springer-Verlag, New York, 1981. MR0620321

[17] Ji, C. and Seymour, L. A consistent model selection procedure for Markov random fields based on penalized pseudolikelihood. *Annals of Applied Probability*, 6(2):423–443, 1996. MR1398052

[18] Kim, Y. and Lee, J. A Bernstein-von Mises theorem in the nonparametric right-censoring model. *Annals of Statistics*, 32(4):1492–1512, 2004. MR2089131

[19] Lian, H. On the consistency of Bayesian function approximation using step functions. *Neural Computation*, 19(11):2871–2880, 2007. MR2352968

[20] Lian, H. *Some topics on statistical theory and applications*. PhD thesis, Brown University, 2007.

[21] Lian, H. Bayes and empirical Bayes inference in change-point problems. *Communications in Statistics - Theory and Methods*, 38(3):419–430, 2009. MR2510794

[22] Lian, H., Thompson, W. A., Thurman, R., Stamatoyannopoulos, J. A., Noble, W. S., and Lawrence, C. E. Automated mapping of large-scale chromatin structure in encode. *Bioinformatics*, 24(17):1911–1916, 2008.

[23] Liu, J. S. and Lawrence, C. E. Bayesian inference on biopolymer models. *Bioinformatics*, 15(1):38–52, 1999.

[24] Pflug, G. C. The limiting log-likelihood process for discontinuous density families. *Zeitschrift Fur Wahrscheinlichkeitstheorie Und Verwandte Gebiete*, 64(1):15–35, 1983. MR0710646

[25] Polfeldt, T. Minimum variance order when estimating the location of an irregularity in the density. *Annals of Mathematical Statistics*, 41(2):673–679, 1970. MR0256500

[26] Rivoirard, V. and Rousseau, J. Bernstein-von Mises theorem for linear functionals of the density. *Preprint*, 2009.

[27] Schwartz, L. On Bayes procedures. *Z. Wahrsch. Verw. Gabiete*, 4:10–26, 1965. MR0184378

[28] Scricciolo, C. On rates of convergence for Bayesian density estimation. *Scandinavian Journal of Statistics*, 34(3):626–642, 2007. MR2368802

[29] Shen, X. T. and Wasserman, L. Rates of convergence of posterior distributions. *Annals of Statistics*, 29(3):687–714, 2001. MR1865337

[30] van der Vaart, A. W. *Asymptotic statistics*. Cambridge University Press, Cambridge, UK; New York, 1998. MR1652247