Asymptotically optimal empirical Bayes inference in a piecewise constant sequence model

Ryan Martin* and Weining Shen†

May 22, 2018

Abstract

Inference on high-dimensional parameters in structured linear models is an important statistical problem. This paper focuses on the piecewise constant Gaussian sequence model, and we develop a new empirical Bayes solution that enjoys adaptive minimax posterior concentration rates and, thanks to the conjugate form of the empirical prior, relatively simple posterior computations.

Keywords and phrases: Adaptive estimation; change-point; high-dimensional; minimax; posterior concentration rate.

1 Introduction

Consider a Gaussian sequence model

\[ Y_i \sim N(\theta_i, \sigma^2), \quad i = 1, \ldots, n, \]

where \( Y = (Y_1, \ldots, Y_n) \) are independent, the variance \( \sigma^2 > 0 \) is known, and inference on the unknown mean vector \( \theta = (\theta_1, \ldots, \theta_n) \) is desired. It is common to assume that \( \theta \) satisfies a sparsity structure, i.e., most \( \theta_i \)'s are zero, and work on these problems goes back at least to Donoho and Johnstone (1994), and more recently in Johnstone and Silverman (2004), Jiang and Zhang (2009), Castillo and van der Vaart (2012), Martin and Walker (2014), and van der Pas et al. (2017).

There has also been recent interest in the piecewise constant structure, in which there is a simple partition \( B \) of the set \( \{1, 2, \ldots, n\} \) into consecutive blocks \( B(s) \subseteq \{1, 2, \ldots, n\}, \ s = 1, \ldots, |B| \), where

\[ \theta_i = \theta_{B(s)}, \quad i \in B(s), \quad s = 1, \ldots, |B|. \]

This structure is parametrized by \( (B, \theta_B) \), where \( B \) is the blocking configuration and \( \theta_B \) is the blocking configuration-specific mean parameters. This piecewise constant sequence model is the canonical example in the change-point literature, e.g., Frick et al. (2014) and Fryzlewicz (2014). The former paper describes applications of this piecewise constant

*Department of Statistics, North Carolina State University, rgmarti3@ncsu.edu
†Department of Statistics, University of California Irvine, weinings@uci.edu
sequence model in DNA copy number profiling and photo emission spectroscopy, and the latter in finance. A related model is piecewise constant regression, investigated by Hutter (2007) and van der Pas and Rockova (2017), the former cites applications in seismology, tomography, biology, and economics, while the latter focuses primarily on machine learning applications.

Our goal in this paper is to develop a Bayesian or, more precisely, empirical Bayesian approach for posterior inference on $\theta$ in the piecewise constant sequence model. Our proposed empirical prior makes use of genuine prior information that says the vector $\theta$ is not too complex, i.e., that the block configuration $B$ has size $|B|$ relatively small, but is data-dependent, or non-informative, on the actual values $\theta_B$ corresponding to the block configuration. Section 2 describes how data is incorporated in the prior in two ways: the first is via a prior centering and the second is a mild regularization. Our theoretical results in Section 3 demonstrate that the corresponding empirical Bayes posterior distribution enjoys adaptive concentration at the minimax rate, recently worked out in Gao et al. (2017), even adjusting to phase transitions. To our knowledge, this is the first adaptive minimax posterior concentration rate result for the piecewise Gaussian sequence model. For the related piecewise constant regression problem, similar results are obtained by van der Pas and Rockova (2017) but ours are stronger and more comprehensive; see Remark 1. Moreover, since the proposed empirical prior for $\theta_B$ are conjugate, the posterior is relatively easy to compute, and numerical examples in Section 5 show the remarkable accuracy of our posterior inference even in finite samples.

## 2 Empirical Bayes model

### 2.1 Prior

In light of the representation $(B, \theta_B)$ of the mean vector $\theta$ in terms of a block configuration and block-specific parameters, a hierarchical prior is appealing. That is, first specify a prior for $B$, then a conditional prior for $\theta_B$, given $B$. Here we follow this general prior specification strategy, but with a twist to make it an empirical prior.

Intuitively, there is no reason to introduce a piecewise constant structure if not for the belief that there are not too many constant blocks, i.e., that $|B|$ is relatively small compared to $n$; see Section 3. This belief can be incorporated into the prior for $B$ in the following way. Set $b = |B|$, and introduce a marginal prior

$$f_n(b) \propto n^{-\lambda(b-1)}, \quad b = 1, \ldots, n,$$

where $\lambda > 0$ is a constant to be specified. Note that this is effectively a truncated geometric distribution with parameter $p = n^{-\lambda}$, which puts most of its mass on small values of the block configuration size, hence incorporating the prior information that $\theta$ is not too complex. Next, if the configuration size $b$ is given, the blocks correspond to a simple partition of $\{1, 2, \ldots, n\}$ into $b$ consecutive chunks, and there are $\binom{n-1}{b-1}$ such partitions. So, for the conditional prior distribution of $B$, given $|B|$, we can simply take a discrete uniform distribution. Therefore, the prior distribution for $B$ is given by

$$\pi_n(B) = f_n(|B|)\left(\binom{n-1}{|B|-1}\right)^{-1},$$

where $\lambda > 0$ is a constant to be specified. Note that this is effectively a truncated geometric distribution with parameter $p = n^{-\lambda}$, which puts most of its mass on small values of the block configuration size, hence incorporating the prior information that $\theta$ is not too complex. Next, if the configuration size $b$ is given, the blocks correspond to a simple partition of $\{1, 2, \ldots, n\}$ into $b$ consecutive chunks, and there are $\binom{n-1}{b-1}$ such partitions. So, for the conditional prior distribution of $B$, given $|B|$, we can simply take a discrete uniform distribution. Therefore, the prior distribution for $B$ is given by

$$\pi_n(B) = f_n(|B|)\left(\binom{n-1}{|B|-1}\right)^{-1},$$

where $\lambda > 0$ is a constant to be specified. Note that this is effectively a truncated geometric distribution with parameter $p = n^{-\lambda}$, which puts most of its mass on small values of the block configuration size, hence incorporating the prior information that $\theta$ is not too complex. Next, if the configuration size $b$ is given, the blocks correspond to a simple partition of $\{1, 2, \ldots, n\}$ into $b$ consecutive chunks, and there are $\binom{n-1}{b-1}$ such partitions. So, for the conditional prior distribution of $B$, given $|B|$, we can simply take a discrete uniform distribution. Therefore, the prior distribution for $B$ is given by

$$\pi_n(B) = f_n(|B|)\left(\binom{n-1}{|B|-1}\right)^{-1},$$

where $\lambda > 0$ is a constant to be specified. Note that this is effectively a truncated geometric distribution with parameter $p = n^{-\lambda}$, which puts most of its mass on small values of the block configuration size, hence incorporating the prior information that $\theta$ is not too complex. Next, if the configuration size $b$ is given, the blocks correspond to a simple partition of $\{1, 2, \ldots, n\}$ into $b$ consecutive chunks, and there are $\binom{n-1}{b-1}$ such partitions. So, for the conditional prior distribution of $B$, given $|B|$, we can simply take a discrete uniform distribution. Therefore, the prior distribution for $B$ is given by

$$\pi_n(B) = f_n(|B|)\left(\binom{n-1}{|B|-1}\right)^{-1},$$

where $\lambda > 0$ is a constant to be specified. Note that this is effectively a truncated geometric distribution with parameter $p = n^{-\lambda}$, which puts most of its mass on small values of the block configuration size, hence incorporating the prior information that $\theta$ is not too complex. Next, if the configuration size $b$ is given, the blocks correspond to a simple partition of $\{1, 2, \ldots, n\}$ into $b$ consecutive chunks, and there are $\binom{n-1}{b-1}$ such partitions. So, for the conditional prior distribution of $B$, given $|B|$, we can simply take a discrete uniform distribution. Therefore, the prior distribution for $B$ is given by

$$\pi_n(B) = f_n(|B|)\left(\binom{n-1}{|B|-1}\right)^{-1},$$

where $\lambda > 0$ is a constant to be specified. Note that this is effectively a truncated geometric distribution with parameter $p = n^{-\lambda}$, which puts most of its mass on small values of the block configuration size, hence incorporating the prior information that $\theta$ is not too complex. Next, if the configuration size $b$ is given, the blocks correspond to a simple partition of $\{1, 2, \ldots, n\}$ into $b$ consecutive chunks, and there are $\binom{n-1}{b-1}$ such partitions. So, for the conditional prior distribution of $B$, given $|B|$, we can simply take a discrete uniform distribution. Therefore, the prior distribution for $B$ is given by

$$\pi_n(B) = f_n(|B|)\left(\binom{n-1}{|B|-1}\right)^{-1},$$

where $\lambda > 0$ is a constant to be specified. Note that this is effectively a truncated geometric distribution with parameter $p = n^{-\lambda}$, which puts most of its mass on small values of the block configuration size, hence incorporating the prior information that $\theta$ is not too complex. Next, if the configuration size $b$ is given, the blocks correspond to a simple partition of $\{1, 2, \ldots, n\}$ into $b$ consecutive chunks, and there are $\binom{n-1}{b-1}$ such partitions. So, for the conditional prior distribution of $B$, given $|B|$, we can simply take a discrete uniform distribution. Therefore, the prior distribution for $B$ is given by

$$\pi_n(B) = f_n(|B|)\left(\binom{n-1}{|B|-1}\right)^{-1},$$

where $\lambda > 0$ is a constant to be specified. Note that this is effectively a truncated geometric distribution with parameter $p = n^{-\lambda}$, which puts most of its mass on small values of the block configuration size, hence incorporating the prior information that $\theta$ is not too complex. Next, if the configuration size $b$ is given, the blocks correspond to a simple partition of $\{1, 2, \ldots, n\}$ into $b$ consecutive chunks, and there are $\binom{n-1}{b-1}$ such partitions. So, for the conditional prior distribution of $B$, given $|B|$, we can simply take a discrete uniform distribution. Therefore, the prior distribution for $B$ is given by

$$\pi_n(B) = f_n(|B|)\left(\binom{n-1}{|B|-1}\right)^{-1},$$

where $\lambda > 0$ is a constant to be specified. Note that this is effectively a truncated geometric distribution with parameter $p = n^{-\lambda}$, which puts most of its mass on small values of the block configuration size, hence incorporating the prior information that $\theta$ is not too complex. Next, if the configuration size $b$ is given, the blocks correspond to a simple partition of $\{1, 2, \ldots, n\}$ into $b$ consecutive chunks, and there are $\binom{n-1}{b-1}$ such partitions. So, for the conditional prior distribution of $B$, given $|B|$, we can simply take a discrete uniform distribution. Therefore, the prior distribution for $B$ is given by

$$\pi_n(B) = f_n(|B|)\left(\binom{n-1}{|B|-1}\right)^{-1},$$
where $B$ ranges over all simple partitions of $\{1, 2, \ldots, n\}$ into consecutive blocks.

It remains to propose a conditional prior for $\theta_B$, given $B$. Here we take a prior which assigns independent normal distributions to each $\theta_{B(s)}$, $s = 1, \ldots, |B|$, but we let the data inform the prior location. According to Martin and Walker (2014, 2017) and Martin et al. (2017), this data-driven prior centering reduces the potentially-problematic influence of the prior tails on the asymptotic behavior of the posterior. In this case, define the $B$-specific maximum likelihood estimator $\hat{\theta}_B = \{\hat{\theta}_{B(s)} : s = 1, \ldots, |B|\}$, where $\hat{\theta}_{B(s)}$ is the average of $\{Y_i : i \in B(s)\}$. Then we center the normal prior on $\hat{\theta}_B$, i.e., the conditional prior for $\theta_B$, given $B$, is given by

$$\theta_{B(s)} \sim N(\hat{\theta}_{B(s)}), \quad s = 1, \ldots, |B|,$$

where the variance $v_{B(s)} = v|B(s)|^{-1}$ and $v > 0$ is a constant to be specified. Denote the density function, with respect to Lebesgue measure on $\mathbb{R}^{|B|}$, of this proposed prior distribution for $\theta_B$ as $\pi_n(\cdot | B)$. A related empirical prior construction for a monotone density was given in Martin (2017).

As usual, the prior distribution for the mean vector $\theta$ under this hierarchical formulation is the finite mixture of $\pi_n(\theta_B | B)$ with respect to the prior distribution $\pi_n(B)$ for the configuration $B$. That mixture prior distribution will be denoted as $\Pi_n$.

The reader may be anticipating that the combination of an empirical prior with the likelihood amounts to double-use of data. To avoid potentially over-fitting, we propose the following very mild additional regularization. Let $L_n(\theta)$ denote the likelihood function based on the model (1), i.e., $L_n(\theta) \propto \exp\{-\frac{1}{2\sigma^2}\|Y - \theta\|^2\}$, where $\|\cdot\|$ denotes the $\ell_2$-norm on $\mathbb{R}^n$. For a fixed $\alpha \in (0, 1)$, define a regularized empirical prior

$$\Pi^\text{reg}_n(d\theta) \propto L_n(\theta)^{-(1-\alpha)} \Pi_n(d\theta).$$

Dividing by a fractional power of the likelihood effectively down-weights those parameter values that are too consistent with the data, hence discouraging over-fitting. Typically, one would take $\alpha$ to be close to 1—e.g., we take $\alpha = 0.99$ in our simulation examples in Section 5—so this additional regularization is very mild indeed.

### 2.2 Posterior

For the posterior distribution, we proposed to combine the regularized empirical prior $\Pi^\text{reg}_n$ with the likelihood $L_n$ according to Bayes’s formula:

$$\Pi^n(A) \propto \int_A L_n(\theta) \Pi^\text{reg}_n(d\theta), \quad A \subseteq \mathbb{R}^n. \quad (4)$$

The following sections investigate the theoretical properties and practical performance of this empirical Bayes posterior distribution.

Of course, the above posterior can be rewritten as

$$\Pi^n(A) \propto \int_A L_n(\theta)^\alpha \Pi_n(d\theta),$$

which is particular well-suited for theoretical analysis; see Appendix 5. This sort of generalized Bayes posterior has received considerable attention recently, e.g., Grünwald.
and van Ommen (2017), Miller and Dunson (2015), Holmes and Walker (2017), Syring and Martin (2018), and Bhattacharya et al. (2017), though not specifically for regularization purposes. One might ask if \( \alpha = 1 \) is a valid choice, since this makes the above display look more like the familiar Bayesian update, but the answer is unclear. Likely, plugging in \( \alpha = 1 \), would result in a posterior distribution with the desired theoretical properties, but this does not follow from the analysis presented here, which takes advantages of having \( \alpha < 1 \). This is not a shortcoming of our approach, we are actually able to improve upon the existing Bayesian rate results for this problem; see Remark 1. While it should be possible to prove results analogous to those in Section 3 below for the \( \alpha = 1 \) case, such an extension is of little practical consequence. Indeed, there is virtually no effect on numerical results since \( \alpha \) can be arbitrarily close to 1, there can be no improvement to the theory since the rates in Theorem 1 are optimal, and, finally, the empirical Bayes posterior \( \Pi^n \) in (4) is already derived by combining the likelihood with an empirical prior that is no more or less justifiable than that with \( \alpha = 1 \).

3 Posterior concentration rates

Consider a true piecewise constant mean vector \( \theta^* \in \mathbb{R}^n \) having block configuration \( B_{\theta^*} \); here and throughout, \( B_{\theta} \) will denote the block configuration of a vector \( \theta \in \mathbb{R}^n \). Define the target rate

\[
\varepsilon_n(\theta^*) = \begin{cases} 
1 & \text{if } |B_{\theta^*}| = 1 \\
|B_{\theta^*}| \log \frac{n}{|B_{\theta^*}|} & \text{if } |B_{\theta^*}| \geq 2.
\end{cases}
\]  

(5)

Note that, in the case \( |B_{\theta^*}| = 1 \), the model is iid, and the best estimator of the constant vector \( \theta^* \) would be the \( n \)-vector with each entry equal to \( \bar{Y} \), the sample mean, and its risk is constant, as in (5). More generally, according to Corollary 3.1 in Gao et al. (2017), the rate (5) is close to the minimax optimal; see Remark 2 below. Theorem 1 says that \( \Pi^n \) attains the target rate in (5). Proofs are deferred to the appendix.

**Theorem 1.** Under the model (1), with known \( \sigma^2 > 0 \), let \( \Pi^n \) be the corresponding empirical Bayes posterior distribution for \( \theta \in \mathbb{R}^n \) described above. If \( \varepsilon_n(\theta^*) \) is the target rate in (5) and \( M_n \) is any sequence satisfying \( M_n \to \infty \), then

\[
\sup_{\theta^*} \mathbb{E}_{\theta^*} \Pi^n(\{\theta \in \mathbb{R}^n : \|\theta - \theta^*\|^2 > M_n \varepsilon_n(\theta^*)\}) \to 0, \quad n \to \infty,
\]

with \( \| \cdot \| \) the \( \ell_2 \)-norm on \( \mathbb{R}^n \). If \( \theta^* \) is such that \( |B_{\theta^*}| \geq 2 \), the above convergence property holds with \( M_n \) replaced by a sufficiently large constant \( M > 0 \).

The prior described above does not have knowledge of the block configuration size \( |B_{\theta^*}| \), hence the rate is adaptive to the unknown complexity level.

**Remark 1.** A result very similar to that in Theorem 1 is presented in van der Pas and Rockova (2017), with a rate of \( |B_{\theta^*}| \log(n/|B_{\theta^*}|) \). However, translating their notation to ours, they assume bounds on both \( \theta^* \) and on \( |B_{\theta^*}| \), which we do not require. Also, their rates do not detect any phase transitions; see Remark 2. It turns out that results similar to that in Theorem 1 can be proved with \( \alpha = 1 \) if we assume suitable bounds on \( \theta^* \) but, since we prefer not to make uncheckable assumptions, we opt to take \( \alpha < 1 \), which is entirely within our control.
Remark 2. The target rate \( \varepsilon_n(\theta^*) = \log \log n \) for all \( \theta^* \) with \( |B_{\theta^*}| = 2 \). Following the proof of Theorem 1, accommodating this phase transition would require \( f_n(2) \asymp n(\log n)^{-c} \) for some constant \( c > 0 \). However, \( f_n \) is a probability mass function, so it cannot blow up with \( n \) like this. Since the concentration rate proofs for other posterior distributions would follow along similar lines, we expect that our inability to handle this additional phase transition is a general phenomenon for Bayesian approaches to this problem, not unique to our approach here.

Next, we show that the posterior mean \( \hat{\theta} = \int \theta \Pi_n(d\theta) \) is an adaptive, asymptotically minimax estimator.

**Theorem 2.** Under the setup in Theorem 1, \( \sup_{\theta^*} \varepsilon_n(\theta^*)^{-1} \| \hat{\theta} - \theta^* \|^2 \lesssim 1 \).

In addition to recovery rate results, the complexity of the posterior itself is interesting. Theorem 3 says that the effective dimension of the posterior is no larger than a multiple of the true block configuration size, i.e., the posterior is of roughly the correct complexity. van der Pas and Rockova (2017) do not have an analogous complexity result.

**Theorem 3.** Under the setup in Theorem 1, there exists a constant \( C > 1 \) such that
\[
\sup_{\theta^*} \mathbb{E}_{\theta^*} \Pi_n(\{ \theta : |B_\theta| > C |B_{\theta^*}| \}) \to 0, \quad n \to \infty.
\]

4 Practical considerations

This section details several practical considerations pertaining to the proposed empirical Bayes solution. This includes posterior computation, tuning parameter selection, etc.

Genuine Bayesian solutions to high-dimensional problems, ones for which optimal posterior rates are available, tend to be based on non-conjugate priors, making computation non-trivial. Our empirical Bayes solution, on the other hand, is based on a conjugate prior, making computations relatively simple. Indeed, the marginal posterior for \( B \) is available in closed-form,
\[
\pi_n(B) \propto \pi_n(B) e^{-\alpha_2 \|B - \hat{B}\|^2} (1 + \frac{v \sigma^2}{\alpha_2})^{-|B|/2}.
\]

Sampling from \( \pi_n(B) \) via Metropolis–Hastings is easy, as is sampling from the conditional posterior of \( \theta_B \), given \( B \). The result is posterior samples of \( (B, \theta_B) \) from which all sorts of relevant quantities can be evaluated, including the posterior mean \( \hat{\theta} \) for estimation and the marginal posterior for \( |B| \) which is relevant for assessing model complexity. We can also extract various types of credible regions for the vector \( \theta \). For example, we can easily get marginal credible intervals for each \( \theta_i \) based on quantiles of its marginal posterior distribution. R code for carrying out this posterior sampling is available at [http://www4.stat.ncsu.edu/~rmartin](http://www4.stat.ncsu.edu/~rmartin).

Next, we need to set the tuning parameters \( \alpha, v, \) and \( \lambda \). As mentioned before, it is reasonable to take \( \alpha \) close to 1 and our simulations suggest that \( \alpha = 0.99 \) is a reasonable choice. Next, it makes sense to take \( v \) to be larger than \( \sigma^2 \) and, for the examples below, with relatively small \( \sigma \), we have found that \( v = 1 \) works well. Finally, \( \lambda \) controls the penalty against large \( |B| \) and, in the examples considered here, \( \lambda = 1 \) is satisfactory.
As for the model variance, the theory in Section 3 assumes $\sigma^2$ to be known. However, in real applications, this may not be known and, therefore, must be estimated. Of course, one can take a prior for $\sigma^2$ and get a corresponding joint posterior for $(\theta, \sigma^2)$. Here, in keeping with the empirical Bayes approach, we opt for a plug-in estimator. We consider the estimator

$$\hat{\sigma}^2 = \frac{1}{2(n-1)} \sum_{i=1}^{n-1} (Y_{i+1} - Y_i)^2,$$  \hspace{1cm} (6)

which, as [Hutter (2007)] argues, will work well when $|B|$ is relatively small. Indeed, in the simulated data examples below, this estimator is shown to be quite accurate.

5 Numerical results

Example 1. First is a standard test example described in [Frick et al. (2014), p. 561; see, also, Fryzlewicz (2014), Appendix B(2)]. This one has $n = 497$ mean parameters, but only $|B^*| = 7$ unique values. This true signal is depicted by the thin black line in Figure 1(a). Data $Y$ is sampled from a normal distribution, centered at the true signal, with standard deviation $\sigma = 0.2$. Here we use the known variance to construct our posterior, but it is worth pointing out that the estimate (6) of the variance in this case is $\hat{\sigma}^2 = 0.05$, which is very close to the true $\sigma^2 = 0.04$, so the results would be virtually the same had we used the estimated variance. The empirical Bayes model described above is fit to these data, using $v = 1$ and $\lambda = 1$, and the posterior mean based on 10,000 Monte Carlo samples from $\Pi^\theta$ is depicted by the black dots in the figure. The posterior mean does an excellent job estimating the true signal, and the marginal 95% credible intervals capture the true $\theta$. Note also that the wider blocks have narrower credible intervals, so a sort of “borrowing information” is achieved. Figure 1(b) shows the posterior distribution for the block configuration size $|B|$ and, in this case, all the mass is assigned to the true $|B^*| = 7$.

Example 2. Next, we follow [Fan and Guan (2017), Section 6.3, and consider a mean vector of size $n = 1000$ with $|B^*| = 20$ unique mean parameters, equally spaced. The specific mean values are randomly sampled from a uniform distribution on $(-2, 2)$. Data are simulated from a normal distribution centered at the signal with $\sigma = 0.5$. Again, we use the known variance in our construction, but the results using estimated variance, $\hat{\sigma}^2 = 0.26$, would be indistinguishable from those presented here. Figure 1(b) summarizes the output of the empirical Bayes posterior, again based on $v = 1$ and $\lambda = 1$. Again, we see that the posterior mean does a good job estimating the true signal, and the marginal 95% credible intervals capture the true $\theta$. The credible intervals are all roughly the same width here because of the equally spaced structure. Finally, according to Panel (b), the posterior for $|B|$ concentrates just below the true $|B^*| = 20$.

Example 3. Finally, we consider a real data example following up on the DNA copy number data analysis in [Hutter (2007)]. In these applications, it is of biological importance to identify the change points, hence the proposed method would be of practical use. Data on the copy number for a particular gene are displayed in Figure 3(a). We fit the proposed empirical Bayes model to these data, using the plug-in estimator for the model variance, which in this case is $\hat{\sigma}^2 = 0.093$, just like in Table 2 of [Hutter (2007)]. Plots of the posterior mean estimate and marginal 95% credible intervals are also shown. The fit
Figure 1: Output for Example 1. Panel (a) shows the data (gray dots), true signal (green line), posterior mean (black line), and marginal 95% credible intervals (red line). Panel (b) shows the posterior distribution of the block configuration size $|B|$. Here appears to be quite good, perhaps with the exception of around Index = 600, and arguably the reason for this is the within-group variance seems to be much larger here than in other regions. Interestingly, the distribution of $|B|$ in Panel (B) is concentrated on much smaller values than in [Hutter (2007)], who estimates about 15 piecewise constant blocks. But a visual inspection of the data suggests much fewer blocks, and roughly 6–7 seems much more reasonable than 15.

6 Discussion

This paper presents an empirical Bayes approach for inference on a high-dimensional normal mean vector under a piecewise constant constraint. We show that the posterior adaptively achieves the asymptotic minimax concentration rate under weaker conditions than those assumed in [van der Pas and Rockova (2017)] for a related piecewise constant regression model. Moreover, by centering the prior distribution on the data in a suitable way, we can take its shape to be a conjugate normal, which makes computation relatively simple compared to that with a fixed-center but heavy-tailed prior.

An interesting possible extension of the work here is related to the formulation in [Fan and Guan (2017)]. Consider a graph $G = (V, E)$ and, at each vertex $i \in V$, there is a response $Y_i \sim N(\theta_i^*, \sigma^2)$, but only a small number of edges $(i, j) \in E$ have $\theta^*_i \neq \theta^*_j$. That paper derives minimax bounds on the recovery rate that are analogous to those achieved here in the sequence model. The only obstacle preventing us from extending our analysis to this more general graph setting is the need to assign a prior distribution for the block structure $B$ in this more complex setting. For example, in a two-dimensional lattice graph, as might be used in imaging applications, one would need a prior on all possible ways that the lattice can be carved up into connected chunks, which is non-trivial. But given such a prior, the theoretical results described here would carry over directly.
Figure 2: Output for Example 2. Panel (a) shows the data (gray dots), true signal (green line), posterior mean (black line), and marginal 95% credible intervals (red line). Panel (b) shows the posterior distribution of the block configuration size $|B|$.

Another interesting and practically important extension is to the case where the mean vector is monotone in addition to piecewise constant. In such a case, our proposed prior for the block configuration $B$ is fine, but a prior for $\theta_B$ that respects the monotonicity is not so straightforward. A natural idea is to project the prior presented in Section 2 onto the cone of monotone sequences, but this projection forces positive prior mass on the boundary of the cone which introduces some additional technicalities that affect the convergence rate proofs. We hope to present results on this elsewhere.

Acknowledgment

This work is partially supported by the National Science Foundation, DMS–1737933, and the Simons Foundation Award 512620. The authors thank Marcus Hutter for sharing the DNA copy number data analyzed in Example 3 above.

A. Proofs

A.0. Preliminary results

For our theoretical analysis, it will help to rewrite the posterior distribution $\Pi^n$ as the ratio $\Pi^n(A) = N_n(A)/D_n$, where the numerator and denominator are

$$N_n(A) = \sum_B \pi_n(B) \int_{A \cap \Theta_B} R_n(\theta_B^+)^\alpha \pi_n(\theta_B \mid B) \, d\theta_B, \quad A \subseteq \mathbb{R}^n,$$

$$D_n = \sum_B \pi_n(B) \int_{\Theta_B} R_n(\theta_B^+)^\alpha \pi_n(\theta_B \mid B) \, d\theta_B,$$
Figure 3: Output for Example 3. Panel (a) shows the data (gray dots), posterior mean (black line), and marginal 95% credible intervals (red line). Panel (b) shows the posterior distribution of the block configuration size $|B|$.

$$R_n(\theta) = \frac{L_n(\theta)}{L_n(\theta^*)}$$ is the likelihood ratio, $\Theta_B \subseteq \mathbb{R}^{|B|}$ consists of all $|B|$-vectors $\theta_B = \{\theta_{B(s)} : s = 1, \ldots, |B|\}$ corresponding to block configuration $B$, and $\theta_B^+$ is the $n$-vector that satisfies $(\theta_B^+)_i = \theta_{B(s)}$ for $i \in B(s)$, with $s = 1, \ldots, |B|$. Since $\theta^*$ is fixed in our calculations, we will abbreviate $B_{\theta^*}$ by $B^*$.

**Lemma 1.** There exists $c > 0$ such that

$$D_n \gtrsim \pi_n(B^*)e^{-c|B^*|} \text{ for all large } n.$$

**Proof.** Define the set $\mathcal{L}_{n,B^*} = \{\theta_{B^*} : L_n(\theta_{B^*}) \geq e^{-|B^*|L_n(\hat{\theta}_{B^*})}\}$. Then

$$D_n \geq \pi_n(B^*) \int_{\mathcal{L}_{n,B^*}} \left\{ \frac{L_n(\theta_{B^*})}{L_n(\hat{\theta}_{B^*})} \right\}^\alpha \pi_n(\theta_{B^*} \mid B^*) d\theta_{B^*} \geq \pi_n(B^*)e^{-|B^*|} \pi_n(\mathcal{L}_{n,B^*} \mid B^*).$$

Under the empirical prior for $\theta_{B^*}$, $Z := \frac{1}{\nu} \sum_{s=1}^{|B^*|} |B^*(s)|(|\theta_{B^*(s)} - \hat{\theta}_{B^*(s)})|^2$ has a chi-square distribution with $|B^*|$ degrees of freedom, and the event $\{\theta_B \in \mathcal{L}_{n,B^*}\}$ is precisely $\{Z \leq 2|B^*|\}$. Using standard bounds on the chi-square distribution function, and Stirling’s approximation of the gamma function, the claim follows, with $c = \alpha + \frac{1}{2} - \log 2$. \qed

Next, for the numerator, we consider a particular sequence of subsets, namely,

$$A_n = A_{M\varepsilon_n} = \{\theta \in \mathbb{R}^n : \|\theta - \theta^*\|^2 > M\varepsilon_n\},$$

where $M > 0$ is a sufficiently large constant, and $\varepsilon_n$ is the target rate.

**Lemma 2.** Take $q > 1$ such that $aq < 1$. Then $\mathbb{E}_{\theta^*}\{N_n(A_n)\} \lesssim e^{-Mk\varepsilon_n}$, for all large $n$, where $k = \alpha(1 - \alpha q)/2\sigma^2$. 

9
Proof. Towards an upper bound, we interchange expectation with the finite sum over $B$ and the integral over $\theta_B$, the latter step justified by Tonelli’s theorem, so that

$$E_{\theta^*}\{N_n(A_n)\} = \sum_B \pi_n(B) \int_{\theta_n} E_{\theta^*}\{R_n(\theta_B^+)^{\alpha} \pi_n(\theta_B \mid B)\} d\theta_B. \quad (7)$$

Next, we work with each of the $B$-dependent integrands separately. Take an arbitrary $p > 1$ and set $q = p/(p - 1)$ to be its Hölder conjugate. Then Hölder’s inequality gives

$$E_{\theta^*}\{R_n(\theta_B^+)^{\alpha} \pi_n(\theta_B \mid B)\} \leq E_{\theta^*}^{1/q}\{R_n(\theta_B^+)^{aq}\} E_{\theta^*}^{1/p}\{\pi_n(\theta_B \mid B)^p\}.$$ 

On the set $A_n$, since $\alpha q < 1$, the first term above is uniformly bounded by $e^{-Mk\varepsilon_n}$. To see this, note that, for a general $\theta \in A_n$, if $p_\theta^0$ denotes the joint density of $Y$ under (1), and $D_{\alpha q}$ the Rényi $\alpha q$-divergence of one normal distribution from another (e.g., van Erven and Harremoës 2014, p. 3800), then

$$E_{\theta^*}\{R_n(\theta)^{aq}\} = \int \{p_\theta^n(y)\}^{aq} \{p_\theta^0(y)\}^{1-aq} dy = e^{-\frac{aq(1-aq)}{2\alpha q} \|\theta-\theta^*\|^2}.$$ 

For the second term, we show that it simplifies to a suitable constant times a normal density function in $\theta_B$. Using the fact that $\hat{\theta}_{B(s)}$ is normally distributed, a simple-but-tedious moment generating function calculation gives

$$E_{\theta^*}^{1/p}\{\pi_n(\theta_B \mid B)^p\} = \left(\frac{v}{v + p\sigma^2}\right)^{|B|/p} (1 - 1)^{|B|} \prod_{s=1}^{|B|} N(\theta(1) \mid E_{\theta^*}\hat{\theta}_{B(s)}, (v + p\sigma^2)|B(s)|^{-1}),$$

and the latter product is a normal density in $\theta_B$. Integrating the upper bound with respect to $\theta_B$ gives

$$E_{\theta^*}\{N_n(A_n)\} \leq e^{-Mk\varepsilon_n} \sum_B \zeta^{|B|} \pi_n(B) = e^{-Mk\varepsilon_n} \sum_{b=1}^n \zeta^b f_n(b),$$

where $\zeta = (1 + p\sigma^2/v)^{(p-1)/2p} > 1$. Using the formula (2) for $f_n$ and standard properties of a geometric series, it is easy to see that, for all large $n$, the summation term in the above upper bound is uniformly bounded in $n$, proving the claim. \qed

A.1. Proof of Theorem 1

Then the proof of Theorem 1 follows by simply combining the bounds in Lemmas 1–2. Since the lower bound on $D_n$ in Lemma 1 is non-stochastic, we have

$$E_{\theta^*}\{\Pi^n(A_n)\} \leq e^{c|B^*|} \pi_n(B^*) E_{\theta^*}\{N_n(A_n)\}.$$ 

Plug in the bound from Lemma 2, with $M = M_n$ to get

$$E_{\theta^*}\{\Pi^n(A_n)\} \leq e^{-M_nk\varepsilon_n+c|B^*|} \frac{(n-1)!}{f_n(|B^*|)}.$$
If $|B_\theta^*| = 1$, then both $\varepsilon_n$ and the ratio in the above display are constant. Therefore, the upper bound vanishes if $M_n \to \infty$. On the other hand, if $|B_\theta^*| \geq 2$, then $\varepsilon_n$ is diverging. Also, using the formula for $f_n(|B^*|)$ in \(2\) and the standard bound, \(\binom{n}{b} \leq e^{b \log(en/b)}\), on the binomial coefficient, we get

$$E_{\theta^*} \{ \Pi^n(A_n) \} \lesssim e^{-M_n k \varepsilon_n + \lambda |B^*| \log n + c |B^*|}.$$ 

Since $|B^*| \log n$ is roughly $\varepsilon_n + |B^*| \log |B^*| = O(\varepsilon_n)$, it follows that, for fixed $M_n$ sufficiently large, the upper bound above vanishes as $n \to \infty$, as was to be proved.

**A.2. Proof of Theorem 2**

Start with $\int \|\theta - \theta^*\|^2 \Pi^n(d\theta)$. Write $R_n$ as $A \cup A^c$, where $A = A_{M\varepsilon_n}$ is as defined above. Then

$$E_{\theta^*} \int \|\theta - \theta^*\|^2 \Pi^n(d\theta) \leq M\varepsilon_n + E_{\theta^*} \int_A \|\theta - \theta^*\|^2 \Pi^n(d\theta).$$

That remaining integral can be expressed as a ratio of numerator to denominator, where the denominator $D_n$ is just as in Lemma 1 and the numerator $\widetilde{N}_n(A)$ is

$$\widetilde{N}_n(A) = \int_A \|\theta - \theta^*\|^2 R_n(\theta)^\alpha \Pi_n(d\theta)$$

$$= \sum_B \pi_n(B) \int_{A \cap \Theta_B} \|\theta_B^+ - \theta^*\|^2 R_n(\theta_B^+)^\alpha \pi_n(\theta_B | B) d\theta_B.$$ 

Take expectation of the numerator to the inside of the integral and apply Hölder’s inequality just like in the proof of Lemma 2. This gives the following upper bound on each $B$-specific integral:

$$\int_{A \cap \Theta_B} \|\theta - \theta^*\|^2 e^{-h\|\theta - \theta^*\|^2} E_{\theta^*} \{ \pi_n(\theta_B | B)^p \} d\theta_B,$$

where $h > 0$ is a constant that depends only on $\alpha$, $\sigma^2$, and the Hölder constant $q > 1$. Since the function $x \mapsto xe^{-hx}$ is eventually monotone decreasing, for sufficiently large $M$ we get a trivial upper bound on the above display, i.e.,

$$M\varepsilon_n e^{-Mh\varepsilon_n} \int_{\Theta_B} E_{\theta^*} \{ \pi_n(\theta_B | B)^p \} d\theta_B.$$ 

The same argument as above bounds the remaining integral by $\zeta^{\|B\|}$, and the prior $\pi_n(B)$ takes care of its contribution. Making $M$ large enough in $e^{-Mh\varepsilon_n}$ will also take care of the bound on $D_n$ from Lemma 1. Therefore, the second term on the right-hand side of \(8\) is also bounded by a multiple of $\varepsilon_n$. Finally, Jensen’s inequality gives $\|\theta - \theta^*\|^2 \leq \int \|\theta - \theta^*\|^2 \Pi^n(d\theta)$, and the claim follows.

**A.3. Proof of Theorem 3**

The lower bound on the denominator in Lemma 1 holds here, and we can employ a similar argument to that in Lemma 2 to upper-bound the numerator. Indeed, we need
to upper-bound

\[ \sum_{B : |B| > C|B^*|} \pi_n(B) \int_{\Theta_B} E_{\theta_B} \left\{ R_n(\theta_B^+)^\alpha \pi_n(\theta_B \mid B) \right\} d\theta_B. \]

The same Hölder’s inequality approach leads to an upper bound like \(\zeta |B|\). Then

\[ E_{\theta_B} \Pi_n(\{ \theta : |B_\theta| > C|B^*| \}) \lesssim \frac{\epsilon^{|B^*|}}{\pi_n(B^*)} \sum_{b=C|B^*|+1}^n \zeta^b f_n(b). \]

From (2), factor out a common \((\zeta n^{-\lambda})^{C|B^*|}\) from the summation, which will be the dominant term. Indeed, like in the proof of Theorem 1, the ratio in the above display is of order \(e^{\epsilon n + \lambda |B^*| \log n}\). Then the right-hand side above is of order \(\exp\{\epsilon n + \lambda |B^*| \log n + C|B^*| \log \zeta - C\lambda |B^*| \log n\}\). For sufficiently large \(C > 1\), the negative term dominates, so the upper bound vanishes, proving the claim.

**References**

Bhattacharya, A., Pati, D., and Yang, Y. (2017). Bayesian fractional posteriors. Unpublished manuscript, arXiv:1611.01125.

Castillo, I. and van der Vaart, A. (2012). Needles and straw in a haystack: posterior concentration for possibly sparse sequences. *Ann. Statist.*, 40(4):2069–2101.

Donoho, D. L. and Johnstone, I. M. (1994). Minimax risk over \(l_p\)-balls for \(l_q\)-error. *Probab. Theory Related Fields*, 99(2):277–303.

Fan, Z. and Guan, L. (2017). Approximate \(\ell_0\)-penalized estimation of piecewise constant signals on graphs. *Ann. Statist.*, to appear; arXiv:1703.01421.

Frick, K., Munk, A., and Sieling, H. (2014). Multiscale change point inference. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 76(3):495–580. With 32 discussions by 47 authors and a rejoinder by the authors.

Fryzlewicz, P. (2014). Wild binary segmentation for multiple change-point detection. *Ann. Statist.*, 42(6):2243–2281.

Gao, C., Han, F., and Zhang, C.-H. (2017). Minimax risk bounds for piecewise constant models. Unpublished manuscript, arXiv:1705.06386.

Grünwald, P. and van Ommen, T. (2017). Inconsistency of Bayesian inference for misspecified linear models, and a proposal for repairing it. *Bayesian Anal.*, 12(4):1069–1103.

Holmes, C. C. and Walker, S. G. (2017). Assigning a value to a power likelihood in a general Bayesian model. *Biometrika*, 104(2):497–503.

Hutter, M. (2007). Exact Bayesian regression of piecewise constant functions. *Bayesian Anal.*, 2(4):635–664.
Jiang, W. and Zhang, C.-H. (2009). General maximum likelihood empirical Bayes estimation of normal means. *Ann. Statist.*, 37(4):1647–1684.

Johnstone, I. M. and Silverman, B. W. (2004). Needles and straw in haystacks: empirical Bayes estimates of possibly sparse sequences. *Ann. Statist.*, 32(4):1594–1649.

Martin, R. (2017). Empirical priors and posterior concentration rates for a monotone density. Unpublished manuscript, arXiv:1706.08567.

Martin, R., Mess, R., and Walker, S. G. (2017). Empirical Bayes posterior concentration in sparse high-dimensional linear models. *Bernoulli*, 23(3):1822–1847.

Martin, R. and Walker, S. G. (2014). Asymptotically minimax empirical Bayes estimation of a sparse normal mean vector. *Electron. J. Stat.*, 8(2):2188–2206.

Martin, R. and Walker, S. G. (2017). Empirical priors for target posterior concentration rates. Unpublished manuscript, arXiv:1604.05734.

Miller, J. W. and Dunson, D. B. (2015). Robust Bayesian inference via coarsening. Unpublished manuscript, arXiv:1506.06101.

Syring, N. and Martin, R. (2018). Calibrating general posterior credible regions. Unpublished manuscript, arXiv:1509.00922.

van der Pas, S. and Rockova, V. (2017). Bayesian dyadic trees and histograms for regression. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30*, pages 2089–2099. Curran Associates, Inc.

van der Pas, S., Szabó, B., and van der Vaart, A. (2017). Adaptive posterior contraction rates for the horseshoe. *Electron. J. Stat.*, 11(2):3196–3225.

van Erven, T. and Harremoës, P. (2014). Rényi divergence and Kullback-Leibler divergence. *IEEE Trans. Inform. Theory*, 60(7):3797–3820.