Statistica Sinica **35** (2025), 225-247
doi:https://doi.org/10.5705/ss.202022.0335

EMPIRICAL PRIORS AND POSTERIOR CONCENTRATION IN A PIECEWISE POLYNOMIAL SEQUENCE MODEL

Chang Liu[1], Ryan Martin[1] and Weining Shen*[2]

[1]*North Carolina State University and* [2]*University of California, Irvine*

Abstract: Inference on high-dimensional parameters in structured linear models is an important statistical problem. Focusing on the case of a piecewise polynomial Gaussian sequence model, we develop a new empirical Bayes solution that enjoys adaptive minimax posterior concentration rates and improved structure learning properties than existing methods. Moreover, the conjugate form of the empirical prior means the posterior computations are fast and easy. Numerical examples highlight the method's strong finite-sample performance compared with that of existing methods in various scenarios.

Key words and phrases: Bayesian estimation, change-point detection, high dimensional inference, structure learning, trend filtering

1. Introduction

Consider a Gaussian sequence model

$$Y_i \sim \mathsf{N}(\theta_i, \sigma^2), \quad i = 1, \ldots, n, \tag{1.1}$$

where $Y = (Y_1, \ldots, Y_n)^\top$ are independent, the variance $\sigma^2 > 0$ is known, and we desire to conduct inference on the unknown mean vector $\theta = (\theta_1, \ldots, \theta_n)^\top$. It is common to assume that θ satisfies a *sparsity structure*, that is, most θ_i are zero. Works on these problems include that of Donoho and Johnstone (1994), and more recently those of Johnstone and Silverman (2004), Jiang and Zhang (2009), Castillo and van der Vaart (2012), Martin and Walker (2014), van der Pas, Szabó and van der Vaart (2017), and Martin and Ning (2020).

There has also been recent interest in imposing low-dimensional structures on high-dimensional parameters, namely, *piecewise constant* and, more generally, *piecewise polynomial* structures. For a fixed positive integer K, we say that the n-vector θ has a piecewise degree-K polynomial structure if there exists a simple partition B of the index set into consecutive blocks $B(s) \subseteq \{1, \ldots, n\}$, with $s = 1, \ldots, |B|$, such that, for each block $B(s)$, the corresponding sub-vector $\{\theta_j : j \in B(s)\}$ can be expressed as a degree-K polynomial of the indices $j \in B(s)$. This piecewise polynomial form is determined by the degree K and the complexity $|B|$ of the block, that is, its dimension is $(K+1)|B|$. When this number is smaller

*Corresponding author. E-mail: weinings@uci.edu

than n, then a θ of this form clearly has a relatively low-dimensional structure. For example, the piecewise constant case corresponds to $K = 0$, so the complexity is determined completely by the number of blocks $|B|$.

Compared with sparse Gaussian signals, few studies examine piecewise constant and piecewise polynomial Gaussian sequence models. Regularization methods, such as trend filtering (Kim et al., 2009) and locally adaptive regression splines (Mammen and van de Geer, 1997), have been proposed to estimate the signal adaptively and recover the underlying block partitions. For piecewise constant problems, Tibshirani et al. (2005) introduce a fused lasso based on a penalized least squares problem using the total variation penalty. Rinaldo (2009) and Qian and Jia (2016) investigate the convergence rate of the fused lasso estimator and the asymptotic properties of pattern recovery. For signals with a more general piecewise polynomial structure, Tibshirani (2014) proposes an adaptive piecewise polynomial estimation using trend filtering that minimizes a penalized least squares criterion, in which the penalty term sums the absolute Kth-order discrete derivatives over the input points. Guntuboyina et al. (2020) show that, under a strong sparsity setting and a minimum length condition, the trend filtering estimator achieves an n^{-1}-rate, up to a logarithmic multiplicative factor. In the Bayesian domain, methods such as the Bayesian fused lasso (Kyung et al., 2010) and Bayesian trend filtering (Roualdes, 2015) have been proposed. However, to the best of our knowledge, no Bayesian studies have examined the posterior contraction as it relates to adaptive estimation and asymptotic structure recovery for such piecewise polynomial Gaussian sequence models. Our goal here is to fill this gap in the literature.

Given the relatively low-dimensional representation of the high-dimensional θ, the now-standard Bayesian approach would be to assign a prior for the unknown block configuration B, and a conditional prior on the block-specific $(K + 1)$-dimensional parameters that determine the polynomial form. For the prior on B, the goal is to induce "sparsity" in the sense that the prior concentrates on block configurations B, with $|B|$ relatively small. For this, one can mostly follow the existing Bayesian literature on sparsity structures, such as Castillo and van der Vaart (2012), Castillo, Schmidt-Hieber and van der Vaart (2015), Martin, Mess and Walker (2017), and Liu et al. (2021), among others. However, for the quantities that determine the polynomial form on a given block configuration, the situation is quite different. In classical sparsity settings, it is reasonable to assume that signals that are not exactly zero are still relatively small, in which case, a conditional prior centered around zero is effective. In this piecewise polynomial setting, there is no obvious fixed center around which a prior should be concentrated. Of course, one option is to choose a fixed center and a wide spread, but then the tails of the prior distribution become particularly relevant. In particular, Theorem 2.8 in Castillo and van der Vaart (2012) shows that if the fixed-center prior has tails thinner than Laplace, then the posterior

contraction rates are sub-optimal, thus excluding the computationally convenient conjugate Gaussian priors. An alternative is to follow Martin and Walker (2019), building on Martin and Walker (2014) and Martin, Mess and Walker (2017), using an *empirical prior* that lets the data help with correctly centering the prior distribution, relieving the computational burden from the restrictions on the tails of the fixed-center prior.

Details of this empirical prior construction are presented in Section 2. Our theoretical results in Section 3 demonstrate that the corresponding empirical Bayes posterior distribution enjoys adaptive concentration at the same rate of trend filtering, adjusting to phase transitions, but requires weaker conditions than those in Guntuboyina et al. (2020). In addition, we establish structure learning consistency results that, to the best of our knowledge, are the first for piecewise polynomial sequence models in the Bayesian literature. Furthermore, because the proposed empirical priors are conjugate, the posterior is relatively easy to compute. The numerical simulations in Section 5 compare our method with trend filtering, showing the advantageous performance of our method in terms of signal estimation and structure recovery under finite-sample settings. In Section 6, we apply our method to two real-world applications, where the underlying truths are considered to be piecewise constant and piecewise linear, respectively. Finally, Section 7 concludes the paper. All technical details and proofs are presented in the Supplementary Material.

2. Empirical Bayes Formulation

2.1. Piecewise polynomial model

Before we introduce our proposed prior and corresponding empirical Bayes model, we precisely formulate the within-block polynomial. Start with the case $|B| = 1$, corresponding to there being only one block. A vector θ being a degree-K polynomial with respect to B corresponds to $\theta \in \mathcal{S}$, where

$$\mathcal{S} = \text{span}\{v_0, v_1, \ldots, v_K\}, \tag{2.1}$$

and $v_k = (1^k, 2^k, \ldots, n^k)^\top \in \mathbb{R}^n$, with $k = 0, 1, \ldots, K$. In other words, if $Z \in \mathbb{R}^{n\times(K+1)}$ is a matrix with columns that form a basis for $\mathcal{S}$, then θ can be expressed as $Z\beta$, for some vector $\beta \in \mathbb{R}^{K+1}$. More generally, for a generic simple partition B, if θ is a piecewise degree-K polynomial on the block configuration B, then it can be expressed as $Z^B\beta^B$, where

$$Z^B = \begin{pmatrix} Z_{B(1)} & 0 & \ldots & 0 \\ 0 & Z_{B(2)} & \ldots & 0 \\ \vdots & \vdots & \ldots & \vdots \\ 0 & 0 & \ldots & Z_{B(|B|)} \end{pmatrix} \in \mathbb{R}^{n\times|B|(K+1)}, \tag{2.2}$$

$Z_{B(s)}$ is the sub-matrix of Z with its row indices included in $B(s)$, and

$$\beta^B = \begin{pmatrix} \beta_1^B \\ \vdots \\ \beta_{|B|}^B \end{pmatrix} \in \mathbb{R}^{|B|(K+1)}, \quad \beta_s^B \in \mathbb{R}^{K+1}, \quad s = 1, \ldots, |B|. \tag{2.3}$$

The following two examples illustrate the piecewise polynomial formulation.

- When $K = 0$, the vector θ formed by $Z^B\beta^B$ is piecewise constant. For a specific block segment $B(s)$, we can write $Z_{B(s)} = I_{|B(s)|}$ and, therefore,

  $$\theta_i \equiv \beta_s^B \in \mathbb{R}, \quad i \in B(s).$$

  Note that in this case, the Gaussian sequence model can be rewritten in the form of a one-way analysis of variance model with $|B|$ treatments and $|B(s)|$ number of replications in each treatment, for $s = 1, \ldots, |B|$.

- When $K = 1$, the vector θ formed by $Z^B\beta^B$ is piecewise linear. For a specific block segment $B(s)$, we can write

  $$Y_{B(s)} = Z_{B(s)}\beta_s^B + \varepsilon, \quad \varepsilon \sim \mathsf{N}_{|B(s)|}(0, \sigma^2 I),$$

  where $Y_{B(s)}$ is a sub-vector of Y with its indices in $B(s)$, and β_s^B is a two-dimensional vector. Hence, within each segment, the observed data can be viewed as a random sample generated from a block-specific simple linear regression model with intercept and slope $\beta_{s,1}^B$ and $\beta_{s,2}^B$, respectively.

To summarize, if θ is an n-vector that is assumed to have a piecewise degree-K polynomial structure, then we can reparametrize θ as (B, θ^B), where θ^B is expressed as $Z^B\beta^B$, for some $\beta^B \in \mathbb{R}^{|B|(K+1)}$, and Z^B is as in (2.2) for some generator matrix Z with columns that form a basis for $\mathcal{S}$ in (2.1). The matrix Z is not unique and, therefore, β^B is not unique either. However, our interest is in the projection θ^B, which is independent of the choice of basis, so this non-uniqueness is not a problem in what follows.

2.2. Empirical prior

Given our representation of a piecewise polynomial mean vector θ using (B, θ^B) or (B, β^B), a hierarchical representation of the prior distribution would be most convenient. That is, we first specify a prior for B, and then a conditional prior for β^B, given B; this, in turn, induces a conditional prior for θ^B. Here, we follow this general prior specification strategy, but allow the conditional prior for β^B to depend on the data in a particular way. Then, this empirical prior for (B, β^B) immediately induces a corresponding empirical prior for (B, θ^B) and θ.

Intuitively, there is no reason to introduce a piecewise polynomial structure unless we believe there are not too many blocks, that is, 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

$$f_n(b) \propto n^{-\lambda(b-1)}, \quad b = 1, \ldots, n, \tag{2.4}$$

where $\lambda > 0$ is a specified constant that controls the severity of the prior's penalty against large $|B|$. 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 θ 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. Therefore, for the conditional prior distribution of B, given $|B|$, we can use a discrete uniform distribution. Therefore, the prior distribution for B is given by

$$\pi_n(B) = \binom{n-1}{|B|-1}^{-1} f_n(|B|), \tag{2.5}$$

where B ranges over all simple partitions of $\{1, 2, \ldots, n\}$ into consecutive blocks. Next we give the conditional prior for β^B, given B. We propose assigning independent, conjugate normal priors to each β_s^B corresponding to a segment $B(s)$. In light of the results of Castillo and van der Vaart (2012), assuming this thin-tailed prior for β_s^B has a fixed center risks sub-optimal posterior contraction rates. To avoid this, we make a notable departure from the traditional Bayesian formulation by following Martin, Mess and Walker (2017) and letting the data inform the prior center. Specifically, our conditional prior for β^B, given B, is taken to be

$$\beta_s^B \sim \mathsf{N}_{K+1}\left(\hat{\beta}_s^B, v\left(Z_{B(s)}^\top Z_{B(s)}\right)^{-1}\right), \quad s = 1, \ldots, |B|, \tag{2.6}$$

independently, where $\hat{\beta}_s^B$ is the least-squares estimator

$$\hat{\beta}_s^B = (Z_{B(s)}^\top Z_{B(s)})^{-1} Z_{B(s)}^\top Y_{B(s)},$$

and $v > 0$ is a constant controlling the prior spread. Write the conditional density of β^B, given B, with respect to the Lebesgue measure on $\mathbb{R}^{|B|(K+1)}$, as

$$\tilde{\pi}_n(\beta^B \mid B) = \prod_{s=1}^{|B|} \mathsf{N}_{K+1}(\beta_s^B \mid \hat{\beta}_s^B, v\{Z_{B(s)}^\top Z_{B(s)}\}^{-1}),$$

which is a product of individual $(K+1)$-variate normal densities. This induces a prior on θ^B through the mapping $\theta^B = Z^B \beta^B$ that defines it. However, because this is generally not a bijection, there is no density function with respect to the Lebesgue measure on $\mathbb{R}^n$. To see this, let $\theta_{B(s)}^B$ denote the sub-vector of θ^B with

indices included in $B(s)$. Then we observe that the induced conditional prior on $\theta^B_{B(s)}$ is $\mathsf{N}_{|B(s)|}(P_{B(s)}Y_{B(s)}, vP_{B(s)})$, where

$$P_{B(s)} = Z_{B(s)}\{Z_{B(s)}^\top Z_{B(s)}\}^{-1} Z_{B(s)}^\top \tag{2.7}$$

is the matrix that projects onto the space spanned by the columns of $Z_{B(s)}$. Because $P_{B(s)}$ is a projection, it is not full rank and, therefore, the prior for $\theta^B_{B(s)}$ is a degenerate normal. Despite this degeneracy, the conditional prior for θ^B, given B, still exists; it is just more convenient to express in terms of the conditional prior for β^B. That is, we define the conditional empirical prior for θ^B, given B, as

$$\Pi_n(\mathcal{A} \mid B) = \int_{\{\beta^B : Z^B\beta^B \in \mathcal{A}\}} \tilde{\pi}_n(\beta^B \mid B)\, d\beta^B, \quad \mathcal{A} \subseteq \mathbb{R}^n.$$

Note that while the prior for β^B depends on the particular basis in Z, the prior for θ^B depends only on the projection, which does not depend on the choice of basis. Finally, our empirical prior for θ is defined as

$$\begin{aligned}\Pi_n(\mathcal{A}) &= \sum_B \pi_n(B)\, \Pi_n(\mathcal{A} \mid B) \\ &= \sum_B \pi_n(B) \int_{\{\beta^B : Z^B\beta^B \in \mathcal{A}\}} \tilde{\pi}_n(\beta^B \mid B)\, d\beta^B.\end{aligned}$$

Although we refer to the object Π_n defined above as a "prior," it is of course not a prior in the traditional Bayesian sense, owing to the data-driven centering. This also differs from the traditional empirical Bayes formulation, where the prior depends on the data only through the choice of a few hyperparameters; here, the "prior" is directly and heavily dependent on the data. Despite these differences, there is nothing stopping us from treating this formally as a "prior" and combining it (see below) with the likelihood to get a corresponding "posterior." There are significant practical advantages to this unorthodox approach. For example, we can enjoy the theoretically optimal concentration rate properties using a computationally simple thin-tailed conjugate prior for β^B, whereas an orthodox Bayesian would require a computationally burdensome heavy-tailed prior for β^B.

2.3. Posterior

Let $L_n(\theta)$ denote the likelihood function based on the model (1.1), that is, $L_n(\theta) \propto \exp\{-\|Y-\theta\|^2/2\sigma^2\}$, where $\|\cdot\|$ denotes the ℓ_2-norm on $\mathbb{R}^n$. We propose combining the empirical data-driven prior Π_n, defined above, with the data as encoded in L_n using (almost) the usual Bayes's formula. Specifically, we define

our corresponding empirical Bayes posterior as

$$\Pi^n(\mathcal{A}) \propto \int_{\mathcal{A}} L_n(\theta)^\alpha \, \Pi_n(d\theta), \tag{2.8}$$

where $\alpha \in (0,1)$ is an additional regularizing factor that down-weights the influence of the data in the likelihood portion of the posterior; see below. (Of course, because Π_n is a proper prior, Π^n is a proper posterior.) This sort of *generalized* or *pseudo posterior* has received considerable attention; see, for example, Grünwald and Van Ommen (2017), Miller and Dunson (2019), Holmes and Walker (2017), Syring and Martin (2019), and Bhattacharya, Pati and Yang (2019), though not specifically for the purpose of regularization. To examine the role α plays, we consider an equivalent formulation. Define a *regularized* empirical prior

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

and then the corresponding more-Bayesian-looking posterior

$$\Pi^n(\mathcal{A}) \propto \int_{\mathcal{A}} L_n(\theta) \, \Pi_n^{\text{reg}}(d\theta). \tag{2.9}$$

By comparing the equivalent expressions (2.8) and (2.9), the role of α becomes clear. The power α on the likelihood in (2.8) is equivalent to an ordinary-looking Bayesian update with a regularized prior that effectively down-weights parameter values with an especially large likelihood, hence discouraging overfitting. The point is that using a data-driven prior blurs the line between what is the "prior" part and what is the "likelihood" part of the posterior. We understand that this might make the reader uncomfortable, but remember the practical motivations for incorporating the data into the prior. The following sections investigate the theoretical convergence properties and practical performance of Π^n in (2.8).

Whether $\alpha = 1$ is a valid choice for an analysis depends on what, if anything, we are willing to give up. In general, the asymptotic consistency of posterior distributions can be established under weaker conditions when using $\alpha < 1$ than when using $\alpha = 1$; this was the motivation behind the results in Walker and Hjort (2001). Similarly, in general, faster rates can be achieved with $\alpha < 1$ than with $\alpha = 1$, under the same conditions. It may be that in a particular application, such as the one considered here, the additional conditions needed to close the gap between $\alpha < 1$ and $\alpha = 1$ can be checked without introducing any practical restrictions, but this is not true in all cases. For example, in generalized linear models, Jeong and Ghosal (2021) argue that the conditions needed to establish optimal concentration rates with $\alpha = 1$ are much stronger than those needed to get the same rates with $\alpha < 1$. Our perspective is that

(a) there is nothing to lose by taking $\alpha < 1$, because it can be taken very close to one; for example, we take $\alpha = 0.99$ in our simulation examples (Sec. 5),

(b) and there is nothing to gain by insisting on $\alpha = 1$, because our concentration rates are improvements on the existing Bayesian rates for this problem (Remark 3) and are optimal in certain cases.

Therefore we embrace the simplicity and flexibility that $\alpha < 1$ affords, rather than apologizing for it and/or insisting on $\alpha = 1$, solely because it makes the posterior distribution "look more Bayesian."

A practical benefit of the simplicity of our formulation is that the posterior distribution is not complicated. Indeed, by combining (2.6) and (2.8), the posterior distribution Π^n for θ is given by

$$\Pi^n(\mathcal{A}) = \sum_B \pi^n(B) \int_{\{\beta^B : Z^B\beta^B \in \mathcal{A}\}} \left\{ \prod_{s=1}^{|B|} f_n(\beta_s^B; s, B) \right\} d\beta^B, \tag{2.10}$$

where

$$f_n(\beta_s^B; s, B) = \mathsf{N}_{K+1}\left(\beta_s^B \mid \hat{\beta}_s^B, \frac{\sigma^2 v}{\sigma^2 + \alpha v}\{Z_{B(s)}^\top Z_{B(s)}\}^{-1}\right),$$

and the marginal posterior for B has mass function

$$\pi^n(B) \propto \pi_n(B)\left(1 + \frac{v\alpha}{\sigma^2}\right)^{-(K+1)|B|/2} e^{-(\alpha/2\sigma^2)\sum_{s=1}^{|B|} \|(I - P_{B(s)})Y_{B(s)}\|^2}, \tag{2.11}$$

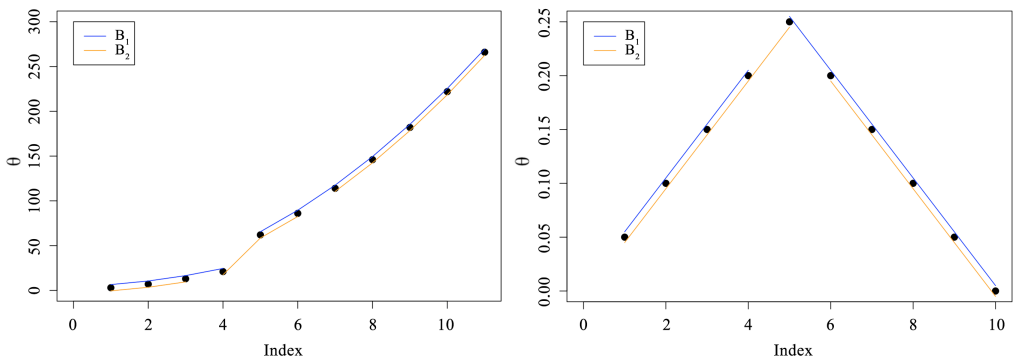
with $P_{B(s)}$ the projection in (2.7). From the latter expression, there are three major factors contributing to the log-marginal posterior distribution of B: the prior distribution of block configuration $\log \pi_n(B)$, a penalty term on the model complexity proportional to $-|B|$, and a model-fitting measure proportional to the negative sum of the squared residuals. Therefore, our posterior distribution prefers models with fewer blocks and better fitting, given the observed data Y. Details about how we compute the posterior distribution are given in Section 4 and in the Supplementary Material.

3. Asymptotic Properties

3.1. Setup

For a vector $\theta \in \mathbb{R}^n$ that has a piecewise degree-K polynomial structure, write B_θ for its block configuration, and let $|B_\theta|$ denote its cardinality. Then, our parameter space corresponds to $\Theta_n(K)$, the set of all n-vectors with a piecewise degree-K polynomial structure and $|B_\theta| = o(n)$. The latter condition on the size of the block configuration ensures that there are not too many blocks, that is, that the signal is not too complex.

When $K \geq 1$, it is possible that a vector θ has multiple block configurations B_θ. That is, there could be multiple B and β^B such that $\theta = Z^B\beta^B$. This does not present a problem when estimating θ, but it does create identifiability



(a) Can be resolved (b) Cannot be resolved

Figure 1. Two examples pertaining to the identifiability of B_θ for a given θ, one where the non-uniqueness of B_θ can be resolved, and one where it cannot. The two lines on each plot both pass through the marked points; the small jitter is to help distinguish the groupings corresponding to the two block configurations.

concerns in the context of structure learning, that is, recovering the underlying block structure. In some cases, the non-uniqueness can be resolved by defining B_θ as the "most economical" of the candidate B's. For example, for an arbitrary signal, any $(K+1)$-tuple of consecutive points can be fit perfectly by a degree-K polynomial. Therefore, blocks of size $K+1$ or smaller are meaningless, and should be ruled out. Figure 1(a) shows an illustration of this for the case $K=2$. However, there are other cases where the non-uniqueness cannot be resolved by ruling out blocks that are too small. Figure 1(b) shows an example of this, where the two candidate block configurations cannot be distinguished by the data. This is not relevant for the results in Section 3.2 below, so we postpone our discussion of how to resolve this problem to Section 3.3.

3.2. Posterior concentration rates

For $x \in \mathbb{R}^n$, define the scaled ℓ_2-norm $\|x\|_n = n^{-1/2}\|x\|_2$ and, for $\theta^\star \in \Theta_n(K)$, define

$$\varepsilon_n^2(\theta^\star) = \begin{cases} n^{-1} & \text{if } |B_{\theta^\star}| = 1, \\ n^{-1}|B_{\theta^\star}| \log n & \text{if } |B_{\theta^\star}| \geq 2. \end{cases} \tag{3.1}$$

Note that, in the case $|B_{\theta^\star}| = 1$, the best estimator of $\theta^\star$ is $P_{\mathcal{S}}Y$, where $P_{\mathcal{S}}$ is the projection matrix onto $\mathcal{S}$ in (2.1), and its expected sum-of-squared-error is $O(n^{-1})$, consistent with (3.1). For the case with $|B_{\theta^\star}| \geq 2$, the rate (3.1) is consistent with others obtained in the literature; see Remark 2 below. Theorem 1 states that the Π^n constructed above attains the rate defined in (3.1). Because the prior can achieve the rates $\varepsilon_n^2(\theta^\star)$ defined above, without knowledge of $\theta^\star$ or $|B_{\theta^\star}|$, it follows that our posterior concentration results are *adaptive* to the unknown complexity of $\theta^\star$.

Theorem 1. *Consider the model* (1.1) *with known* $\sigma^2 > 0$ *and, assume that* $\theta^\star$ *has a piecewise polynomial structure of degree* $K \geq 0$*, with* K *known. Let* Π^n *be the corresponding empirical Bayes posterior distribution for* $\theta \in \mathbb{R}^n$ *described above. If* $\varepsilon_n^2(\theta^\star)$ *is as in* (3.1)*, then for any sequence* M_n *with* $M_n \to \infty$*, there exists a constant* $G > 0$ *such that*

$$\mathsf{E}_{\theta^\star} \Pi^n(\{\theta \in \mathbb{R}^n : \|\theta - \theta^\star\|_n^2 > M_n \varepsilon_n^2(\theta^\star)\}) \lesssim e^{-G M_n n \varepsilon_n^2(\theta^\star)},$$

for all large n*, uniformly in* $\theta^\star \in \Theta_n(K)$*. For the latter case in* (3.1)*, the sequence* M_n *can be replaced by a sufficiently large constant* $M > 0$*.*

Remark 1. Given data $Y \sim \mathsf{N}_n(\theta^\star, \sigma^2 I)$, an oracle who has access to $B_{\theta^\star}$ would fit a polynomial of degree K in each of the partitions given by $B_{\theta^\star}$. This would be a linear estimator, and its corresponding oracle risk is $O(n^{-1}|B_{\theta^\star}|)$. Note that the rate achieved in Theorem 1 is comparable to the oracle risk. Indeed, our method adaptively learns the underlying block structure of $\theta^\star$ and, in the case $|B_{\theta^\star}| = 1$, we can exactly match the oracle rate; otherwise, the price we pay in terms of the rate is only logarithmic.

Remark 2. The minimax rate, $n^{-1}|B_{\theta^\star}| \log(en/|B_{\theta^\star}|)$, can be achieved if we assume more control on the complexity of $\theta^\star$, that is, if $|B_{\theta^\star}| = O(n^t)$, for some $t \in [0, 1)$. The only way this extra assumption fails is if the signal is extremely complex, for example, if $|B_{\theta^\star}| = O(n/\log n)$. Such cases effectively have no low-dimensional block structure, and should be rare in practice. This minimax rate can be achieved by using trend filtering (see Guntuboyina et al., 2020, Cor. 2.3), but this too requires additional assumptions. Indeed, their result holds only when their minimum length condition is satisfied and the tuning parameter is properly chosen within an unspecified "ideal" range. The former—see Equation (13) in Guntuboyina et al. (2020)—restricts the length of the minimal block to be no smaller than $O(n|B_{\theta^\star}|^{-1})$, which cannot be checked in practice. They also make a strong sparsity assumption that requires $|B_{\theta^\star}|$ to be "much smaller than n." This surely excludes extremely high-complexity cases, such as $|B_{\theta^\star}| = O(n/\log n)$. Therefore, our empirical Bayes posterior concentration rate result is no weaker than the results for trend filtering in Guntuboyina et al. (2020), which the authors argue are stronger than any existing results in the literature. Chatterjee and Goswami (2021) present some risk-bound results for multivariate piecewise polynomial estimation based on a dyadic decision tree approach. Their rate (e.g., their Cor. 3.2 and Thm. 3.4) agrees with ours in Theorem 1, but also requires conditions on the tuning parameter and the complexity of the tree partition space.

Remark 3. A similar result to Theorem 1 is presented in van der Pas and Ročková (2017) for the piecewise constant case $K = 0$, with a rate of $|B_{\theta^\star}| \log(n/|B_{\theta^\star}|)$. However, translating their notation to ours, they assume bounds on both $\|\theta^\star\|_\infty$ and $|B_{\theta^\star}|$, which we do not require. In addition, from

Theorem 2.8 of Castillo and van der Vaart (2012), we do not expect that optimal concentration rates can be achieved using their fixed-center normal prior for θ^B, given B, without some assumptions on the magnitude of $\theta^\star$. Another related work is that of Gao, van der Vaart and Zhou (2020), who consider a structured Bayesian linear model and establish oracle inequalities based on elliptical Laplace priors on the coefficients. Their result (e.g., Theorem 4) is applicable to the piecewise polynomial model considered here, and it implies a posterior concentration rate of $n^{-1}\{(K+1)|B_{\theta^\star}| + |B_{\theta^\star}|\log(en/|B_{\theta^\star}|)\}$, which is virtually the same rate obtained in our Theorem 1. In addition to the concentration rates in terms of θ, we address the structure learning problem in Section 3.3, which is not discussed by Gao, van der Vaart and Zhou (2020). Moreover, because the data-driven prior formulation allows us to achieve optimal concentration rates while using a convenient conjugate prior, we also enjoy straightforward posterior computation, as shown in Section 4. Although Gao, van der Vaart and Zhou (2020) do not discuss computational considerations, posterior computations based on their non-conjugate prior are more difficult than using our proposed empirical prior-based method.

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*,*

$$\mathsf{E}_{\theta^\star} \|\hat{\theta} - \theta^\star\|_n^2 \lesssim M_n \varepsilon_n^2(\theta^\star),$$

for all large n, uniformly in $\theta^\star \in \Theta_n(K)$. In the latter case of (3.1)*, the diverging sequence M_n can be replaced by a constant M, which can be absorbed into "$\lesssim$" above.*

3.3. Structure learning

In addition to estimation consistency, it is interesting to consider when the posterior is able to recover the unknown block structure of the true piecewise polynomial signal $\theta^\star$. To the best of our knowledge, this is the first Bayesian (or empirical Bayesian) investigation into structure learning in the piecewise polynomial Gaussian sequence model. When $K = 0$, that is, the true signal is piecewise constant, learning the underlying block structure can be viewed as detecting the "change points" or "jump points," which has many real-world applications. In the non-Bayesian literature, structure recovery for piecewise constant and piecewise polynomial signals has received some attention. Next we compare our results with those available for trend filtering and binary segmentation, among others.

As a first result in this direction, Theorem 3 states that the effective dimension of the posterior is no larger than a multiple of the true block configuration size; in other words, the posterior is of roughly the correct complexity. Note that this result pertains only to the size $|B_\theta|$ of the block configurations, which can

be determined uniquely, and thus there are no identifiability issues here. Finally, for this and the other results of this section, the statements are formulated in terms of the marginal posterior distribution π^n for the block configuration B, as defined in (2.11).

Theorem 3. *Under the setup in Theorem* 1*, for any* $C > 1 + \lambda^{-1}$*, where* λ *is as in* (2.4)*, there exists a constant* $G > 0$ *such that*

$$\mathsf{E}_{\theta^\star}\pi^n(\{B : |B| > C|B_{\theta^\star}|\}) \lesssim e^{-G|B_{\theta^\star}|\log n},$$

for all large n*, uniformly in* $\theta^\star \in \Theta_n(K)$*.*

Block configuration size is important, but we need to identify the underlying block structure. However, first, we need to address the potential non-identifiability of $B_{\theta^\star}$. As mentioned before, there are no such issues in the piecewise constant case with $K = 0$, but non-identifiability is possible for $K \geq 1$. On the one hand, if $\theta^\star$ is such that non-uniqueness can be resolved simply by taking the most economical of the equally well-fitting block configurations, then that is how $B_{\theta^\star}$ is defined. On the other hand, if $\theta^\star$ has multiple block configurations of the same size, as in Figure 1(b), then it is not possible to distinguish between these. In such cases, the best we can hope for is that the posterior distribution will concentrate on the set $\mathbb{B}^\star = \{B_{\theta^\star}\}$ of equivalent block configurations corresponding to $\theta^\star$. We establish that this is the case in the results below.

The first result concerns the event that B is a *refinement* of $B_{\theta^\star}$, denoted by $B \sqsupset B_{\theta^\star}$, for some $B_{\theta^\star} \in \mathbb{B}^\star$. That is, if $B \sqsupset B_{\theta^\star}$, then every block in $B_{\theta^\star}$ can be expressed as a union of blocks in B or, equivalently, no block in B intersects with more than one block in $B^\star$. Because refinements or unnecessary splits of $B_{\theta^\star}$ are a sign of inefficiency, we hope the posterior will discourage such cases. Indeed, Theorem 4 shows that the posterior distribution assigns a vanishing probability to the event "$B \sqsupset B_{\theta^\star}$," which means that the posterior for B asymptotically avoids those inefficient refinements. This is analogous to the "no supersets" theorems in Castillo, Schmidt-Hieber and van der Vaart (2015, Thm. 4) and Martin, Mess and Walker (2017, Thm. 4) for variable selection in a linear regression context. The only additional requirement here is that the power λ in the prior for $|B|$ in (2.4) is not too small; otherwise, the prior does not sufficiently penalize block configurations that are too complex, leaving open the possibility for overfitting. Similar conditions appear in the regression setting, for example, the conditions of Theorem 4 in Castillo, Schmidt-Hieber and van der Vaart (2015).

Theorem 4. *Under the setup of Theorem* 1*,*

$$\mathsf{E}_{\theta^\star}\pi^n(\{B : B \sqsupset B_{\theta^\star} \text{ for some } B_{\theta^\star} \in \mathbb{B}^\star\}) \to 0, \quad n \to \infty,$$

uniformly in $\theta^\star \in \Theta_n(K) \cap \{\theta : |B_\theta| = o(n^\lambda)\}$*, with* $\lambda > 0$ *as in* (2.4)*.*

If $\lambda \geq 1$, then the above condition on $|B_{\theta^\star}|$ is satisfied for all $\theta^\star \in \Theta_n(K)$. However, for smaller values of λ, such as those with good empirical performance in Section 5, restricting to a proper subset of the parameter space is required, but is not severe.

Next we discuss how to exactly recover the true block configuration $B_{\theta^\star}$ or, more generally, the set $\mathbb{B}^\star$ of equivalent true block configurations. First, we need some additional notation. Define the 0^{th}- and 1^{st}-order difference operators as $\Delta^0 x = x$ and

$$\Delta^1 x = (x_2 - x_1, x_3 - x_2, \ldots, x_n - x_{n-1})^\top,$$

respectively, where $x = (x_1, \ldots, x_n)^\top \in \mathbb{R}^n$. For a generic order $K \geq 2$, the K^{th}-order difference, $\Delta^K : \mathbb{R}^n \to \mathbb{R}^{n-K}$, is defined recursively as $\Delta^K x = \Delta^1(\Delta^{K-1} x)$. Second, a change in the signal $\theta^\star$ from one block to another can only be detected if the change is sufficiently large, and the definitions of "change" and "sufficiently large" are related to the properties of the difference operators applied to $\theta^\star$. In particular, the set of indices where a change in the $(K+1)^{\text{st}}$-order occurs is defined as

$$J_{\theta^\star} = \{j = 1, \ldots, n - K - 1 : (\Delta^{K+1}\theta^\star)_j \neq 0\}.$$

In the piecewise constant case, with $K = 0$, the set $\{j + 1 : j \in J_{\theta^\star}\}$ consists of those indices at which the signal jumps from one value to another. Then, both the minimal change in $\theta^\star$ on $J_{\theta^\star}$ and the minimal spacing between changes are relevant to determining whether a change is sufficiently large to be detectable. These are defined, respectively, as

$$\delta_n(\theta^\star) = \min_{j \in J_{\theta^\star}} |(\Delta^{K+1}\theta^\star)_j| \quad \text{and} \quad \gamma_n(\theta^\star) = \min_{j, j' \in J_{\theta^\star}, j \neq j'} |j - j'|.$$

Then, the following theorem states that the block configuration $B_{\theta^\star}$ can be recovered exactly if $\gamma_n(\theta^\star)\delta_n^2(\theta^\star)$ is sufficiently large, analogous to the so-called *beta-min* condition in linear regression (e.g., Bühlmann and Van De Geer, 2011, Chap. 2).

Theorem 5. *Under the setup in Theorem* 1*, suppose that*

$$\gamma_n(\theta^\star)\delta_n^2(\theta^\star) \geq \frac{4^{K+1} M \sigma^2}{\alpha(1-\alpha)} \log n, \tag{3.2}$$

with $M > 4 + \lambda$ *and* $\lambda \geq 3$*, where* λ *controls the prior* (2.4)*. Then,*

$$\mathsf{E}_{\theta^\star} \pi^n(\mathbb{B}^\star) \to 1, \quad n \to \infty. \tag{3.3}$$

To the best of our knowledge, only the piecewise constant ($K = 0$) case, where the true $B_{\theta^\star}$ is unique, has been considered in the literature, so we focus on that version here in our discussion of Theorem 5. In that case, $\gamma_n(\theta^\star)$ and $\delta_n(\theta^\star)$ represent the smallest number of indices between jumps and the smallest

signal jump in $\theta^\star$, respectively. To draw a parallel between the piecewise constant signal problem and a one-way analysis of variance, $\gamma_n(\theta^\star)$ is like the minimum number of replications across all the treatment groups, and $\delta_n(\theta^\star)$ is like the minimum effect size. In that classical analysis of variance context, where the number of treatment groups and group memberships are fixed and known, the F-test has power converging to one if $\gamma_n(\theta^\star)\delta_n^2(\theta^\star)$ is bounded away from zero. The condition $\gamma_n(\theta^\star)\delta_n^2(\theta^\star) \gtrsim \log n$ in (3.2) is only slightly stronger, that is, we pay only a logarithmic price for not knowing the number of groups or group memberships. Returning to the general piecewise constant case, if the minimum block size $\gamma_n(\theta^\star)$ is fixed as n and $|B_{\theta^\star}|$ go to infinity, the result in Theorem 5 matches the pattern recovery property of the fused lasso in Qian and Jia (2016), and is stronger than the corresponding results in Lin et al. (2017) and Dalalyan, Hebiri and Lederer (2017). We can also allow the minimum block size $\gamma_n(\theta^\star)$ to grow. For example, the minimum block length condition in Guntuboyina et al. (2020) states that $\gamma_n(\theta^\star)$ can be of order $O(n|B_{\theta^\star}|^{-1})$, corresponding to equally partitioning over blocks. In this case, the minimum jump size simply needs to satisfy $\delta^2(\theta^\star) \gtrsim n^{-1}|B_{\theta^\star}| \log n$, which is mild, because the right-hand side typically vanishes. This flexibility makes our result preferable to those for the fused lasso, and comparable to thoese for the wild binary segmentation in Theorem 3.2 of Fryzlewicz (2014), which is the best result available in the literature that we are aware of. Finally, note that Theorems 4 and 5 are, to the best of our knowledge, the first results of their kind in the Bayesian literature.

Recently Fang and Ghosh (2024) considered a high-dimensional linear regression model with an inverse gamma prior on σ^2 and an empirical prior on the coefficients. They obtained model selection consistency and a posterior contraction rate for the coefficients. Therefore, we expect our rate convergence results (e.g., Thm. 1 and 2) can be extended to treat unknown σ^2. The prior on the coefficients can be changed to

$$(\beta_s^B \mid B, \sigma^2) \sim \mathsf{N}_{K+1}\big(\hat{\beta}_s^B, \sigma^2 v\{Z_{B(s)}^\top Z_{B(s)}\}^{-1}\big),$$

which allows for easy computation (Lee, Lee and Lin, 2019). Whether the structure learning results hold with unknown σ^2 remains an open question.

4. Computation

Genuine Bayesian solutions to high-dimensional problems, ones for which optimal posterior rates are available, tend to be based on non-conjugate, heavy-tailed priors, making computation nontrivial. Our empirical Bayes solution, on the other hand, is based on a conjugate prior for θ^B, making computations relatively simple.

Recall that the marginal posterior for B is available in closed form, up to proportionality, as in (2.11). Furthermore, recall from (2.10) that the conditional distribution of θ^B, given B, is determined by a linear transformation of a normal random variable, which is easy to simulate. Together, these two observations suggest a Metropolis–Hastings algorithm to draw Markov chain Monte Carlo (MCMC) samples from the proposed posterior Π^n for θ. We provide more details in Section S4 of the Supplementary Material.

5. Simulated data examples

5.1. Methods

In this section, we compare the numerical performance of our proposed method with that of the adaptive piecewise polynomial trend filtering of Tibshirani (2014). We use the R package `genlasso` to implement trend filtering, and choose the tuning parameter using five-fold cross-validation or the "one-standard error" rule; see Hastie, Tibshirani and Friedman (2009, Chap. 7).

In order to implement the above sampling procedures, we need to specify some additional hyperparameters in (2.11). As mentioned before, because $\alpha = 0.99$ has little practical difference to the $\alpha = 1$ case, which corresponds to the genuine Bayesian model, we plug $\alpha = 0.99$ into the posterior distribution functions for practical implementation. Next, for model variance σ^2, although the theory in Section 3 assumes it is known, in practice, it may need to be estimated. Of course, one can take a prior for σ^2 and get a corresponding joint posterior for (θ, σ^2); see Martin and Tang (2020). Here, in keeping with the spirit of our empirical Bayes approach, we opt for a plug-in estimator. Specifically, we consider

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{\theta}_i^{\text{TF}})^2, \tag{5.1}$$

where $\hat{\theta}^{\text{TF}}$ is the trend filtering/lasso estimate based on cross-validation. For the prior variance v, it makes sense to take v to be larger than σ^2 and, for the examples below, with relatively small σ^2, we find that $v = 1$ works well. Finally, λ controls the penalty against large $|B|$. In the examples considered here, we conduct a sensitivity analysis in which $\lambda = 0.2, 0.5, 1$ are considered. For every data set, 50,000 iterations of the aforementioned MCMC algorithm, with an additional 50,000 burn-in runs, are used to generate posterior samples.

5.2. Scenarios

For data generation, we consider six models for the true signal $\theta^\star$. More details are given in Section S2. For model fitting, we use the true K value for Models 1–4. For Models 5 and 6, because the data are generated from trigonometric functions, there is no "true value" of K. Therefore, we use $K = 2$,

because it already provides an accurate curve approximation.

5.3. Results

We investigate the numerical performance of the two methods in terms of their estimation error and block selection accuracy. For Models 1–6, we compute the squared estimation error loss, $\|\hat{\theta} - \theta^\star\|^2$, where $\hat{\theta}$ is either our posterior mean or the trend filtering estimator obtained using cross-validation; see Table S1 in the Supplementary Material. In addition, the estimated signal $\hat{\theta}$ and the true $\theta^\star$ are plotted in Figures 2 and S7. In these graphical comparisons, the trend filtering estimator is computed using the "one-standard error" rule, because it is usually smoother than that chosen using cross-validation, although it typically suffers from a higher mean squared error; see Hastie, Tibshirani and Friedman (2009, Chap. 7) for details.

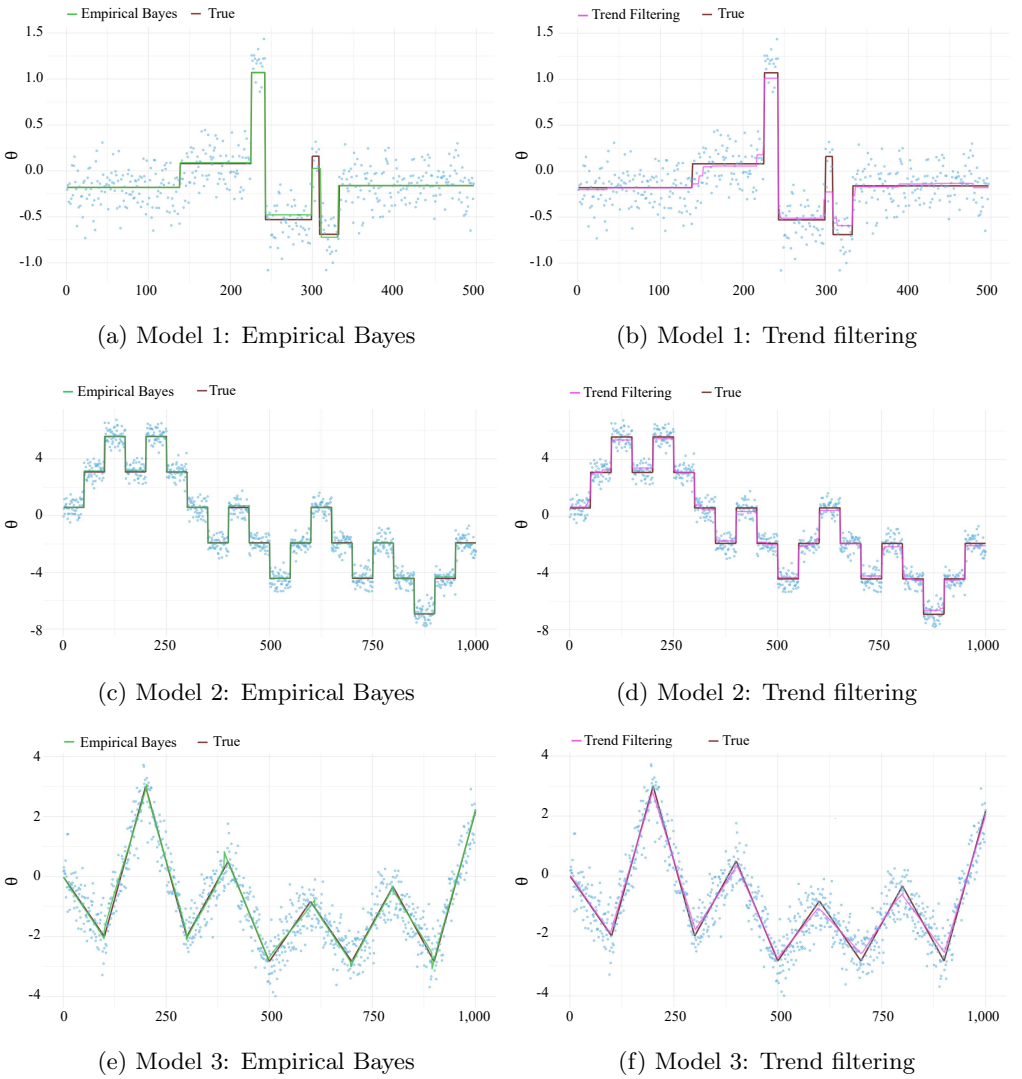
The estimated block partition $\hat{B}$ for trend filtering is obtained from the nonzero entries of $D^{(K+1)}\hat{\theta}$, that is, the K^{th}-order "knots" of $\hat{\theta}$; see Guntuboyina et al. (2020). For our empirical Bayes method, $\hat{B}$ is the maximizer of the marginal posterior probability $\pi^n(B)$. Because structure recovery is most meaningful for lower-order polynomials, we focus on the piecewise constants, namely, Models 1 and 2. The results are displayed in Tables S2 and S3 and Figures S8 and S9 in the Supplementary Material.

To gain a better understanding of the performance of the two methods in terms of structure learning, we use multiple criteria to measure the change-point detection/block selection accuracy. From the 100 replications for each model, we estimate the probability that $\hat{B}$ is equal to the true $B^\star$ and covers the true $B^\star$, denoted as $\mathsf{P}(\hat{B} = B^\star)$ and $\mathsf{P}(\hat{B} \supset B^\star)$, respectively. We also estimate $\mathsf{E}|\hat{B}|$, the mean size of $\hat{B}$. In addition, as discussed in Section 3.3, an equivalent representation of the block partition is the set of jump locations J, defined in Theorem 5. We can calculate the Hausdorff distance between J and $J^\star$ using the following formula

$$H(J \mid J^\star) = \max_{j^\star \in J^\star} \min_{j \in J} |j - j^\star| + \max_{j \in J} \min_{j^\star \in J^\star} |j - j^\star|.$$

Finally, we consider an $(n - K - 1)$-dimensional binary vector S, with $S_i = 1$ if and only if $i \in J$. The Hamming distance between $\hat{S}$ and $S^\star$ is reported as a measure of how close $\hat{J}$ and $J^\star$ are to each other.

For the estimation accuracy for $\theta^\star$, in Table S1, our method achieves a smaller squared error loss than trend filtering, except for Models 3 and 5, which are the two that are continuous; the Doppler wave function in Model 6 is continuous too, but the high frequency oscillation in $[0, 100]$ makes it "almost discontinuous." Therefore, our method tends to have an advantage in terms of estimation performance when the underlying $\theta^\star$ has jump discontinuities, particularly for the piecewise constant signals. Furthermore, our method demonstrates stronger



(a) Model 1: Empirical Bayes

(b) Model 1: Trend filtering

(c) Model 2: Empirical Bayes

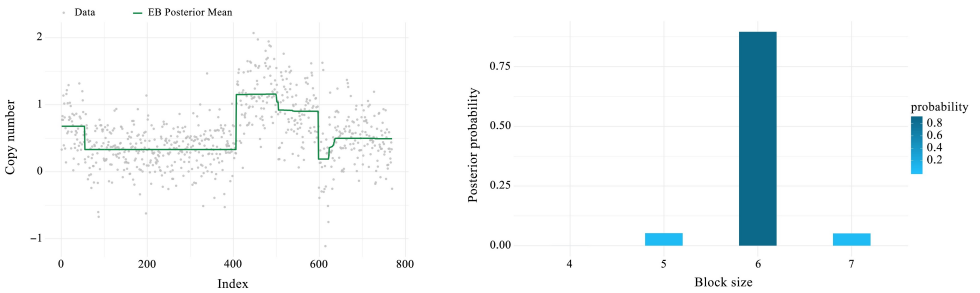
(d) Model 2: Trend filtering

(e) Model 3: Empirical Bayes

(f) Model 3: Trend filtering

Figure 2. Plots of the empirical Bayes and trend filtering estimates of the signal for representative cases under Models 1–3.

structure recovery for piecewise constant Models 1 and 2 as shown in Tables S2 and S3 in the Supplementary Material. Compared with trend filtering, which tends to select more blocks, when $\lambda = 0.5$, our method detects the exact block number for Model 1. In terms of the Hamming and Hausdorff distances, our method also outperforms trend filtering for both models. However, the probabilities of identifying the true block partition are relatively low for both methods. This is likely because the jump size, $\delta^\star = \delta(\theta^\star)$, is borderline too small to be detected. To investigate this, we redo the simulations for Model 2, but with $\delta^\star$-values ranging over $[0.5, 4.0]$; see Figure S8 in the Supplementary Material. On



(a) Posterior mean for θ (b) Posterior distribution for $|B|$

Figure 3. DNA copy number analysis results

the one hand, trend filtering has a rapidly increasing probability of covering $B^\star$, but the probability of identifying $B^\star$ is effectively zero. On the other hand, both probabilities for the empirical Bayes method are increasing at about the same rate. We conclude that trend filtering is rather conservative in the context of structure learning, tending to pick too many blocks, whereas the empirical Bayes method is more aggressive, and hence more efficient. Furthermore, the plots of the Hamming and Hausdorff distances versus $\delta^\star$ in Figure S9 in the Supplementary Material confirm that the more aggressive approach of the empirical Bayes method leads to more accurate structure learning than when using trend filtering.

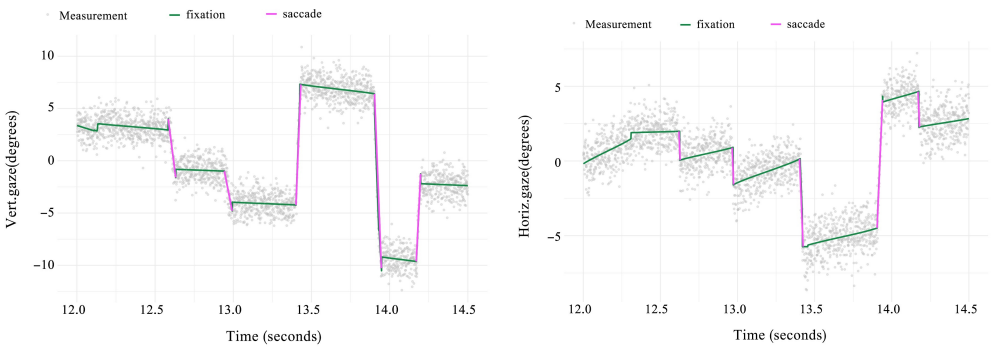
6. Real-Data Examples

6.1. DNA copy number analysis

We consider a real-data example based on the DNA copy number analysis in Hutter (2007). In these applications, it is of biological importance to identify the change points, so the proposed method is useful. Data on the copy number for a particular gene are displayed in grey dots 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). A plot of the posterior mean estimate is also shown. The fit here appears to be reasonably good, perhaps with the exception around 600, where the within-group variance seems to be much larger than in other regions. Interestingly, the distribution of $|B|$ in Figure 3(b) is concentrated on much smaller values than in Hutter (2007), who estimates about 15 piecewise constant blocks. However, a simple visual inspection of the data suggests much fewer blocks, perhaps six or seven, rather than 15.

6.2. Eye movement signal analysis

Another interesting application of our method when $K = 1$ is eye movement signal denoising. Eye movement of human and other foveate animals when



(a) Vertical gaze position (b) Horizontal gaze position

Figure 4. Eye movement signal denoising results

scanning scenes is characterized by a fixate-saccade-fixate pattern. During the fixation phase, gaze position is stable on the order of 0.2–0.3 seconds; in the saccadic phase, the eye moves quickly on the order of 0.01–0.1 seconds. The time series of gaze position in terms of the vertical and horizontal visual angle degree can be well approximated by piecewise linear functions, assuming the eye moves at an approximately constant velocity during each phase; see Pekkanen and Lappi (2017).

Noise in eye-movement recording is usually inevitable, ranging from around 0.01° with laboratory optical equipment to well over 1° in mobile recording with moving cameras. Here, we consider the gaze position data set in Vig, Dorr and Cox (2012), in which participants watch a movie clip. The noise level is not reported in Vig, Dorr and Cox (2012), so we adopt the procedure in Pekkanen and Lappi (2017), who investigate the same dataset, and add a simulated measurement noise with standard deviation 1°; see the Supplementary Information in Pekkanen and Lappi (2017). Then, for both vertical and horizontal gaze position data in a 2.5 second excerpt of the full recording, we fit an empirical Bayes estimator for the mean gaze trajectory, with $\lambda = 1$ and 50,000-length MCMC after burn-in. The posterior mean estimate and the measurements mimicking mobile recording using a moving camera are plotted in Figure 4. Based on the fitted vertical gaze position and horizontal gaze position, an estimated mean gaze path is plotted in Figure 5.

Our method helps to identify and understand the segmentation of the fixate-saccade-fixate pattern in eye movements. As shown in Figures 4–5, in the fixation segments (green), the eye moves slowly and steadily, and hence the gaze position appears to be linear with a slope close to zero. In the saccade segments (magenta), the gaze position is still linear, but much steeper, showing a jump pattern. In addition, segmentation of eye movements is consistent between vertical gaze signal and horizontal gaze signal.

Figure 5. Estimated gaze path

7. Conclusion

We have considered inference on a piecewise polynomial signal, where the degree is known, but the block structure is unknown. We have developed an empirical Bayes posterior that is simple and fast to compute, and exhibits several desirable theoretical results, including optimal posterior concentration rates and block selection consistency. Our general results are new and, when applied to cases that have been investigated previously in the literature, in general, our assumptions are weaker and/or our conclusions are stronger than those currently available. In addition, as our numerical results demonstrate, the strong theoretical properties of the proposed method carry over to real applications, particularly when the underlying function being estimated is discontinuous, or approximately so, as in Model 6 above.

There has been recent interest in cases where the signal is both piecewise constant and *monotone*; see for example, Gao, Han and Zhang (2020) and Guntuboyina and Sen (2018). Of course, the proposed method can be applied to such cases, but it is not immediately clear how to incorporate monotonicity into the prior formulation directly. An alternative strategy is to force the monotonicity constraint by projecting the posterior samples of θ from Π^n onto the space of monotone sequences. That is, if $\theta \sim \Pi^n$, then set $\text{proj}(\theta) = \text{argmin}_{z \in \Theta^\uparrow} \|z - \theta\|$, where $\Theta^\uparrow \subset \mathbb{R}^n$ is the set of monotone sequences. This projection operation is just a function of θ, albeit implicit, so there is a corresponding posterior distribution for the projection, called a *projection posterior*. General details about the projection posterior can be found in Chakraborty and Ghosal (2021). Aside from inheriting many of the desirable properties of the original posterior, the projection posterior is also relatively simple to compute. The R package "Iso" (Turner, 2015) contains an implementation of the "pool adjacent violators algorithm," or *PAVA*. Thus we would need to generate samples of the piecewise

constant θ from the posterior Π^n, and then apply the `pava` function to project it onto the space of monotone sequences. Figure S10 in the Supplementary Material shows the results of sampling from this projected posterior for a simulated data set, and the corresponding estimate appears to be quite accurate.

Another interesting possible extension of our work is related to the formulation in Fan and Guan (2018). Consider a graph $G = (V, E)$ and, at each vertex $i \in V$, there is a response $Y_i \sim \mathsf{N}(\theta_i^\star, \sigma^2)$, but only a small number of edges $(i, j) \in E$ have $\theta_i^\star \neq \theta_j^\star$. They derive bounds on the recovery rate analogous to those achieved here in the chain graph/sequence model. The only obstacle preventing us from extending our analysis to this more general setting is the need to assign a prior distribution for the block structure B in this more complex graph. 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 seems nontrivial. However, given such a prior, we expect that our theoretical results would hold.

Supplementary Material

Additional technical details, numerical results, and proofs are presented in the Supplementary Material.

Acknowledgments

CL and RM were partially supported by the U.S. National Science Foundation, DMS–1737933, and WS was supported by the Simons Foundation, Award 512620. The authors thank Marcus Hutter for sharing the DNA copy number data, and the anonymous reviewers for their helpful comments.

References

Bhattacharya, A., Pati, D. and Yang, Y. (2019). Bayesian fractional posteriors. *The Annals of Statistics* **47**, 39–66.

Bühlmann, P. and Van De Geer, S. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer Science & Business Media.

Castillo, I., Schmidt-Hieber, J. and van der Vaart, A. (2015). Bayesian linear regression with sparse priors. *The Annals of Statistics* **43**, 1986–2018.

Castillo, I. and van der Vaart, A. (2012). Needles and straw in a haystack: Posterior concentration for possibly sparse sequences. *The Annals of Statistics* **40**, 2069–2101.

Chakraborty, M. and Ghosal, S. (2021). Coverage of Bayesian credible intervals in monotone regression. *The Annals of Statistics* **49**, 1011–1028.

Chatterjee, S. and Goswami, S. (2021). Adaptive estimation of multivariate piecewise polynomials and bounded variation functions by optimal decision trees. *The Annals of Statistics* **49**, 2531–2551.

Dalalyan, A. S., Hebiri, M. and Lederer, J. (2017). On the prediction performance of the Lasso. *Bernoulli* **23**, 552–581.

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

Fan, Z. and Guan, L. (2018). Approximate ℓ_0-penalized estimation of piecewise-constant signals on graphs. *The Annals of Statistics* **46**, 3217–3245.

Fang, X. and Ghosh, M. (2024). High-dimensional properties for empirical priors in linear regression with unknown error variance. *Statistical Papers* **65**, 237–262.

Fryzlewicz, P. (2014). Wild binary segmentation for multiple change-point detection. *The Annals of Statistics* **42**, 2243–2281.

Gao, C., Han, F. and Zhang, C.-H. (2020). On estimation of isotonic piecewise constant signals. *The Annals of Statistics* **48**, 629–654.

Gao, C., van der Vaart, A. W. and Zhou, H. H. (2020). A general framework for Bayes structured linear models. *The Annals of Statistics* **48**, 2848–2878.

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

Guntuboyina, A., Lieu, D., Chatterjee, S. and Sen, B. (2020). Adaptive risk bounds in univariate total variation denoising and trend filtering. *The Annals of Statistics* **48**, 205–229.

Guntuboyina, A. and Sen, B. (2018). Nonparametric shape-restricted regression. *Statistical Science* **33**, 568–594.

Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The Elements of Statistical Learning*. Springer Science & Business Media.

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

Hutter, M. (2007). Exact Bayesian regression of piecewise constant functions. *Bayesian Analysis* **2**, 635–664.

Jeong, S. and Ghosal, S. (2021). Posterior contraction in sparse generalized linear models. *Biometrika* **108**, 367–379.

Jiang, W. and Zhang, C.-H. (2009). General maximum likelihood empirical Bayes estimation of normal means. *The Annals of Statistics* **37**, 1647–1684.

Johnstone, I. M. and Silverman, B. W. (2004). Needles and straw in haystacks: Empirical Bayes estimates of possibly sparse sequences. *The Annals of Statistics* **32**, 1594–1649.

Kim, S.-J., Koh, K., Boyd, S. and Gorinevsky, D. (2009). ℓ_1 trend filtering. *SIAM Review* **51**, 339–360.

Kyung, M., Gill, J., Ghosh, M. and Casella, G. (2010). Penalized regression, standard errors, and Bayesian Lassos. *Bayesian Analysis* **5**, 369–411.

Lee, K., Lee, J. and Lin, L. (2019). Minimax posterior convergence rates and model selection consistency in high-dimensional dag models based on sparse Cholesky factors. *The Annals of Statistics* **47**, 3413–3437.

Lin, K., Sharpnack, J. L., Rinaldo, A. and Tibshirani, R. J. (2017). A sharp error analysis for the fused Lasso, with application to approximate changepoint screening. In *Advances in Neural Information Processing Systems*, 6884–6893.

Liu, C., Yang, Y., Bondell, H. and Martin, R. (2021). Bayesian inference in high-dimensional linear models using an empirical correlation-adaptive prior. *Statistica Sinica* **31**, 2051–2072.

Mammen, E. and van de Geer, S. (1997). Locally adaptive regression splines. *The Annals of Statistics* **25**, 387–413.

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

Martin, R. and Ning, B. (2020). Empirical priors and coverage of posterior credible sets in a sparse normal mean model. *Sankhya A* **87**, 477–498.

Martin, R. and Tang, Y. (2020). Empirical priors for prediction in sparse high-dimensional linear regression. *Journal of Machine Learning Research* **21**, 1–30.

Martin, R. and Walker, S. G. (2014). Asymptotically minimax empirical Bayes estimation of a sparse normal mean vector. *Electronic Journal of Statistics* **8**, 2188–2206.

Martin, R. and Walker, S. G. (2019). Data-driven priors and their posterior concentration rates. *Electronic Journal of Statistics* **13**, 3049–3081.

Miller, J. W. and Dunson, D. B. (2019). Robust Bayesian inference via coarsening. *Journal of the American Statistical Association* **114**, 1113–1125.

Pekkanen, J. and Lappi, O. (2017). A new and general approach to signal denoising and eye movement classification based on segmented linear regression. *Scientific Reports* **7**, 1–13.

Qian, J. and Jia, J. (2016). On stepwise pattern recovery of the fused Lasso. *Computational Statistics & Data Analysis* **94**, 221–237.

Rinaldo, A. (2009). Properties and refinements of the fused Lasso. *The Annals of Statistics* **37**, 2922–2952.

Roualdes, E. A. (2015). Bayesian trend filtering. *arXiv:1505.07710*.

Syring, N. and Martin, R. (2019). Calibrating general posterior credible regions. *Biometrika* **106**, 479–486.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J. and Knight, K. (2005). Sparsity and smoothness via the fused Lasso. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **67**, 91–108.

Tibshirani, R. J. (2014). Adaptive piecewise polynomial estimation via trend filtering. *The Annals of Statistics* **42**, 285–323.

Turner, R. (2015). *Iso: Functions to Perform Isotonic Regression*. R package version 0.0-17.

van der Pas, S. and Ročková, V. (2017). Bayesian dyadic trees and histograms for regression. In *Advances in Neural Information Processing Systems*, 2089–2099.

van der Pas, S., Szabó, B. and van der Vaart, A. (2017). Adaptive posterior contraction rates for the horseshoe. *Electronic Journal of Statistics* **11**, 3196–3225.

Vig, E., Dorr, M. and Cox, D. (2012). Space-variant descriptor sampling for action recognition based on saliency and eye movements. In *European Conference on Computer Vision*, 84–97. Springer.

Walker, S. and Hjort, N. L. (2001). On Bayesian consistency. *Journal of the Royal Statistical Society Series B. Statistical Methodology* **63**, 811–821.

(Received October 2022; accepted April 2023)