2020-01-01

Spatially Adaptive Estimation Of Spectrum

Yi None Xie
University of Texas at El Paso

SPATIALLY ADAPTIVE ESTIMATION OF SPECTRUM

YI XIE

Master's Program in Computational Science

APPROVED:

Ori Rosen, Ph.D., Chair

Elizabeth J. Walsh, Ph.D.

Suneel Babu Chatla, Ph.D.

Stephen L. Crites, Jr., Ph.D.
Dean of the Graduate School

to my

MOTHER and FATHER

with love

SPATIALLY ADAPTIVE ESTIMATION OF SPECTRUM

by

YI XIE

THESIS

Presented to the Faculty of the Graduate School of

The University of Texas at El Paso

in Partial Fulfillment

of the Requirements

for the Degree of

MASTER OF SCIENCE

Computational Science Program

THE UNIVERSITY OF TEXAS AT EL PASO

December 2020

Abstract

When analyzing a stationary time series, one of the questions we are often interested in is how to estimate its spectrum. Many approaches have been proposed to this end. Most are focused on smoothing the periodogram using a single smoothing parameter across all Fourier frequencies. In this paper, we smooth the log periodogram by placing a spatially adaptive prior called the dynamic shrinkage prior, so that varying degrees of smoothing may be applied to different intervals of Fourier frequencies, resulting in less biased estimates of the spectrum. Further research will extend this approach to spectral estimation for nonstationary time series.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

The following definitions are taken from Shumway and Stoffer (2017).

Definition 1 A discrete time series is a sequence of data points being recorded at specific times. Usually these time points are equally spaced, in which case the time series is denoted by $\{x_t; t = 0, \pm 1, \pm 2, \ldots\}$.

Definition 2 The **mean function** of time series $\{x_t; t = 0, \pm 1, \pm 2, \ldots\}$ is defined as

$$\mu_{xt} = E(x_t),$$

where E denotes the usual expectation operator. When no confusion exists about which time series we are referring to, we will drop a subscript and write μ_{xt} as μ_t.

Definition 3 The **auto-covariance function** of a time series $\{x_t; t = 0, \pm 1, \pm 2, \ldots\}$ is defined as

$$\gamma_x(s,t) = \mathsf{cov}(x_s, x_t) = E((x_s - \mu_s)(x_t - \mu_t))$$

for all s and t. When no possible confusion exists about which time series we are referring to, we will drop the subscript and write $\gamma_x(s,t)$ as $\gamma(s,t)$.

Definition 4 A **weakly stationary** time series is a finite variance process where

1. the mean value function, μ_t, is constant and does not depend on time t, and
2. the auto-covariance function, $\gamma(s,t)$, depends on s and t only through their difference $|s - t|$.

Since the mean function, $E(x_t) = \mu_t$, of a stationary time series is independent of time t, we will write $\mu_t = \mu$. Also, because the auto-covariance function, $\gamma(s,t)$, of a stationary time series, x_t, depends on s and t only through their difference $|s-t|$, we may simplify the notation. Let $s = t+h$, where h represents the time shift or lag. Then

$$\gamma_x(t+h,t) = \mathsf{cov}(x_{t+h}, x_t) = \mathsf{cov}(x_t, x_0) = \gamma(h,0)$$

because the time difference between times $t+h$ and t is the same as the time difference between times h and 0. Thus, the auto-covariance function of a stationary time series does not depend on the time argument t. Henceforth, for convenience, we will drop the second argument of $\gamma(h,0)$.

Definition 5 The **auto-covariance function of a stationary time series** will be written as

$$\gamma(h) = \mathsf{cov}(x_{t+h}, x_t) = E((x_{t+h}-\mu)(x_t-\mu)).$$

Definition 6 A **strictly stationary** time series is one for which the probabilistic behavior of every collection of values and shifted values

$$\{x_{t_1}, x_{t_2}, \ldots, x_{t_k}\} \quad \textit{and} \quad \{x_{t_1+h}, x_{t_2+h}, \ldots, x_{t_k+h}\}$$

are identical, for all $k = 1, 2, \ldots$, all time points $t_1, t_2, \ldots, t_k$, and all time shifts $h = 0, \pm 1, \pm 2, \ldots$.

Definition 7 A time series $\{x_t; t = 0, \pm 1, \pm 2, \ldots\}$ is $\mathbf{ARMA}(p,q)$ if it is stationary and

$$x_t = \phi_1 x_{t-1} + \cdots + \phi_p x_{t-p} + w_t + \theta_1 w_{t-1} + \cdots + \theta_q w_{t-q}$$

with $\phi_p \neq 0$, $\theta_q \neq 0$, and $\sigma_w^2 > 0$. The parameters p and q are called the autoregressive and the moving average orders, respectively. We assume that w_t is a Gaussian white noise series with mean zero and variance σ_w^2.

Definition 8 An **autoregressive model** of order p, abbreviated **AR(p)**, is of the form

$$x_t = \phi_1 x_{t-1} + \cdots + \phi_p x_{t-p} + w_t,$$

where x_t is stationary, and $\phi_1, \phi_2, \ldots, \phi_p$ are constants $\phi_p \neq 0$. We assume that w_t is a Gaussian white noise series with mean zero and variance σ_w^2.

Example. $x_t = x_{t-1} - 0.9x_{t-2} + w_t$ is an AR(2) model, where w_t is white Gaussian noise with σ_w^2.

Definition 9 The **moving average model** of order q, or **MA(q)**, is defined to be

$$x_t = w_t + \theta_1 w_{t-1} + \cdots + \theta_q w_{t-q},$$

where x_t is stationary, and $\phi_1, \phi_2, \ldots, \phi_p$ are constants such that $\phi_p \neq 0$. We assume that w_t is a Gaussian white noise series with mean zero and variance σ_w^2.

Example. $x_t = w_t + \theta w_{t-1}$ is an MA(1) model, where w_t is white Gaussian noise with σ_w^2, $\theta \neq 0$.

Definition 10 If the auto-covariance function, $\gamma(h)$, of a stationary process satisfies

$$\sum_{h=-\infty}^{\infty} |\gamma(h)| < \infty,$$

then it has the representation

$$\gamma(h) = \int_{-\frac{1}{2}}^{\frac{1}{2}} e^{2\pi i \omega h} f(\omega) d\omega \quad h = 0, \pm 1, \pm 2, \ldots,$$

where $f(\omega)$ is the **spectral density**. The latter has the representation

$$f(\omega) = \sum_{h=-\infty}^{\infty} \gamma(h) e^{-2\pi i \omega h} \quad -\frac{1}{2} \leq \omega \leq \frac{1}{2}.$$

Properties of the spectral density function:

1. $f(\omega) \geq 0$ for all ω.

2. $f(-\omega) = f(\omega)$, it is an even function.

3. It is a periodic function, $f(\omega + 1) = f(\omega)$.

In addition, putting $h = 0$ in

$$\gamma(h) = \int_{-\frac{1}{2}}^{\frac{1}{2}} e^{2\pi i\omega h} f(\omega) d\omega \quad h = 0, \pm 1, \pm 2, \ldots$$

yields

$$\gamma(0) = Var(x_t) = \int_{-\frac{1}{2}}^{\frac{1}{2}} f(\omega) d\omega,$$

which expresses the total variance as the integrated spectral density over all of the frequencies.

Definition 11 Given data $x_1, \ldots, x_n$, we define the **discrete Fourier transform (DFT)** to be

$$d(\omega_j) = n^{-\frac{1}{2}} \sum_{t=1}^{n} x_t e^{-2\pi i\omega_j t}$$

for $j = 0, 1, \ldots, M$, where the frequencies $w_j = \frac{j}{n}$ are called the **Fourier** or **fundamental frequencies**, $M = [\frac{n-1}{2}]$ (the largest positive integer no greater than $\frac{n-1}{2}$).

Definition 12 Given data $x_1, \ldots, x_n$, we define the **periodogram** to be

$$I(\omega_j) = |d(\omega_j)|^2$$

for $j = 0, 1, \ldots, M$, $M = [\frac{n-1}{2}]$.

When the sample size n is large, $I(\omega_j) \overset{ind}{\sim} \text{Exponential}(f(\omega_j))$, approximately.

Definition 13 Let $z = z_1 + iz_2$, where $i = \sqrt{-1}$ and $z_1, z_1 \overset{iid}{\sim} N(0, \frac{\sigma^2}{2})$. Then

$$f(z_1, z_2) \propto \frac{1}{\sigma} \exp\Big(-\frac{z_1^2}{\sigma^2}\Big) \times \frac{1}{\sigma} \exp\Big(-\frac{z_2^2}{\sigma^2}\Big) = \frac{1}{\sigma^2} \exp\Big(-\frac{z_1^2 + z_2^2}{\sigma^2}\Big).$$

We say z has a **complex normal distribution** with mean 0 and variance σ^2 and denote it as $z \sim CN(0, \sigma^2)$. The pdf of z is given by

$$f(z) \propto \frac{1}{\sigma^2} \exp(-\frac{z^* z}{\sigma^2}),$$

where z^* is the complex conjugate.

Definition 14 A random variable X has a **Z distribution** with parameters a, b, μ, σ, denoted $X \sim Z(a, b, \mu, \sigma)$, if its pdf

$$f(x) = \frac{1}{\sigma * Beta(a, b)} \exp\{\frac{(x-\mu)}{\sigma}\}^a [1 + \exp\{\frac{(x-\mu)}{\sigma}\}]^{-(a+b)}$$

where the $Beta(a, b)$ mean beta distribution with parameters a and b.

Definition 15 A random variable X has a **Pólya-Gamma distribution** with parameters $b > 0$ and $c \in R$, denoted $X \sim PG(b, c)$, if

$$X = \frac{1}{2\pi^2} \sum_{k=1}^{\infty} \frac{g_k}{(k - \frac{1}{2})^2 + \frac{c^2}{4\pi^2}}$$

where the $g_k \sim Ga(b, 1)$ are independent gamma random variables.

Chapter 2

Estimation of spectral densities

Several approaches have been taken to estimating the spectral density nonparametrically.

When the sample size n is large, $d(\omega_j) \overset{ind}{\sim} CN(0, f(\omega_j))$, approximately. This implies

$$g(d(\omega_j)) \propto \frac{1}{f(\omega_j)} \exp\Big\{-\frac{|d(\omega_j)^2|}{f(\omega_j)}\Big\} = \frac{1}{f(\omega_j)} \exp\Big\{-\frac{I(\omega_j)}{f(\omega_j)}\Big\}, \tag{2.1}$$

where $g(d(\omega_j))$ is the pdf of $d(\omega_j)$ and $I(\omega_j)$ is the peridogram. From (2.1) we see that $I(\omega_j) \sim \text{Expon}\Big(f(\omega_j)\Big)$ approximately, where $\text{Expon}\Big(f(\omega_j)\Big)$ denotes the exponential distribution with mean $f(\omega_j)$. Let $\epsilon_j = \frac{I(\omega_j)}{f(\omega_j)}$, then $\epsilon_j \sim \text{Expon}(1)$. It follows that $I(\omega_j) = \epsilon_j f(\omega_j)$. Taking logs of both sides leads to the log-linear model

$$\log(I(\omega_j)) = \log(f(\omega_j)) + \eta_j, \quad \text{for} \quad j = 1, \ldots, M, \tag{2.2}$$

where $\eta_j = \log \epsilon_j \sim \log(\text{Expon}(1)) = \log(\frac{1}{2}\chi^2_2)$. Model (2.2) was used by Wahba (1980) to estimate the spectral density by smoothing splines.

Whittle likelihood

Since $I(\omega_j) \sim \text{Expon}\Big\{f(\omega_j)\Big\}$ approximately, we can write the likelihood as follows:

$$\begin{aligned} L(I|f) &\propto \textstyle\prod_{m=1}^{M} \frac{1}{f(\omega_m)} \exp\Big\{-\frac{I(\omega_m)}{f(\omega_m)}\Big\} \\ &= \exp\Big\{-\textstyle\sum_{m=1}^{M}[\log f(\omega_m) + \exp\{\log I(\omega_m) - \log f(\omega_m)\}]\Big\}, \end{aligned} \tag{2.3}$$

which is called the Whittle likelihood (Whittle, 1962).

Pawitan and O'Sullivan (1994), used the penalized Whittle likelihood to estimate the spectral density of a stationary time series.

Carter and Kohn (1997), used a Bayesian approach where the error term η_j was approximated by a mixture of normal distributions with fixed values.

Some traditional methods for estimating the spectral density (such as averaged periodogram) are mentioned in Shumway and Stoffer (2017).

2.1 Nonparametric regression

If in a regression analysis we assume there is a predetermined relation between independent variables and a dependent variable then this regression analysis is called parametric, otherwise it is called a nonparametric. In nonparametric regression we need to estimate the form of the relationship between the independent variables and the dependent variable based on observed data. To this end, we need a set of basis functions, whose linear combination will capture the interesting features of the data. B-splines are an example of a possible type of basis functions.

B-splines

To define a family of B-spline functions of order $p+1$ uniquely, two things are needed:

1. A polynomial of degree p (the order of a B-spline function equals the polynomial degree p plus 1).

2. A non-decreasing sequence of knots, $t_1, \ldots, , t_q$.

Then the ith member of of a family of B-splines of order 1 is defined as

$$B_{i,1}(x) := \begin{cases} 1 & \text{if} \quad t_i \leq x < t_{i+1} \\ 0 & \text{otherwise.} \end{cases}$$

B-splines of higher order k are defined recursively as follows,

$$B_{i,k}(x) := \delta_{i,k} B_{i,k-1}(x) + (1 - \delta_{i+1,k}) B_{i+1,k-1}(x),$$

where

$$\delta_{i,k} := \begin{cases} \frac{x - t_i}{t_{i+k-1} - t_i} & \text{if} \quad t_i \neq t_{i+k-1} \\ 0 & \text{otherwise.} \end{cases}$$

Here are some general properties of a B-spline of order $p + 1$, see Eilers and Marx (1996).

1. It consists of $p + 1$ polynomial pieces, each of degree p.
2. The polynomial pieces join at p inner knots.
3. At the joining points, derivatives up to order $p - 1$ are continuous.
4. The B-spline is positive on a domain spanned by $p + 2$ knots; everywhere else it is zero.
5. Except at the boundaries, it overlaps with $2p$ polynomial pieces of its neighbours.
6. At a given $x, p + 1$ B-splines are non-zero.

With these good properties, B-splines are ideal basis functions for nonparametric modeling.

P-splines

Marx and Eilers (1999) proposed a generalized linear regression model for curve fitting, in which the idea of P-splines is proposed. P-splines consist of a combination of B-splines and a second-order difference penalty placed on the coefficients of these B-splines (to control the smoothness of the fitted curve).

Lang and Brezger (2004) developed a Bayesian version of P-splines and put a random-walk prior (up to a second-order) on the B-spline coefficients. In this paper, the first order random-walk prior is defined as

$$\beta_\rho = \beta_{\rho-1} + u_\rho, \tag{2.4}$$

and the second order random-walk prior is defined as

$$\beta_\rho = 2\beta_{\rho-1} - \beta_{\rho-2} + u_\rho,$$

where the β_ρs are B-spline coefficients, $u_\rho \sim N(0, \tau^2)$, and diffuse priors are placed on β_1 (for a first-order random-walk) or β_1 and β_2 (for a second order random-walk prior). The diffuse (improper) prior is $p(\beta_i) \propto 1$, $i = 1, 2$. ($p(\beta_i)$ means the prior on β_i). The amount of smoothness is controlled by the smoothing parameter τ^2, which is a global smoothing parameter. In other words, the same amount of smoothing is applied to different covariate values (frequency in our case).

Bayesian variable selection

P-splines are one way to prevent over-fitting. Another approach is to start with a relatively large number of basis functions and to allow some of the coefficients to be close to zero. This approach is common in Bayesian variable selection for regression models (George and McCulloch, 1997).

Spike and slab prior

One of the first priors on the regression coefficients used in Bayesian variable selection was the spike and slab prior (George and McCulloch, 1997). It is often written as a two-component mixture of Gaussians

$$\beta_i|\rho_i, c \sim \rho_i N(0, c^2) + (1 - \rho_i)N(0, \epsilon^2), \quad \rho_i \sim Ber(\pi), \tag{2.5}$$

where $\rho_i \sim Ber(\pi)$ means ρ_i has a Bernoulli distribution with probability π that $\rho_i = 1$. The parameter c is called the slab width.

In (2.5), the first term on the right-hand side is called slab. The variance c^2 is relatively large so $N(0, c^2)$ has its support over a wide range of plausible values of β_i. The second component is the spike with $\epsilon^2 << c^2$. If we set $\epsilon = 0$, then the spike is called a Dirac's delta at δ_0.

The horseshoe prior

The setting of the horseshoe prior is as follows.

$$\beta_i|\lambda_i, \tau \sim N(0, \lambda_i^2\tau^2), \quad \lambda_i|\sigma \sim C^+(0, \sigma), \quad \tau|\eta \sim C^+(0, \eta). \tag{2.6}$$

In (2.6), $C^+(0, a)$ means the half-Cauchy distribution with scale parameter a. We can see that the level of shrinkage of β_i is controlled by two parameters, λ_i (the local smoothing parameter) and τ (the global smoothing parameter). Thus, the horseshoe prior has the freedom to shrink globally (via τ) and yet act locally (via λ_i). The global parameter τ pulls all the weights globally towards zero, while the thick half-Cauchy tails for the local scales λ_j allow some of the weights to escape the shrinkage, see Carvalho et al. (2010).

The density function of the horseshoe prior (Figure 2.1) has an infinitely tall spike at the origin and flat, Cauchy-like tails. These two features allow β_is with large values to remain large and force small β_is to shrink to values close to zero. So it can accommodate unknown levels of sparsity by changing the value of τ.

In (2.6), set $\tau = \sigma = 1$ and let $\kappa_i = \frac{1}{1+\lambda_i^2}$, then we obtain

$$E(\beta_i|y_i, \lambda_i^2) = \Big(\frac{\lambda_i^2}{1+\lambda_i^2}\Big)y_i + \Big(\frac{1}{1+\lambda_i^2}\Big)0 = (1-\kappa_i)y_i, \tag{2.7}$$

where y_i is the observed data.

Under the setting $\tau = \sigma = 1$, $\lambda_i \sim C^+(0, 1)$, $\kappa_i = \frac{1}{1+\lambda_i^2} \sim Beta(\frac{1}{2}, \frac{1}{2})$. Figure 2.2 shows the density curve of κ_i, which looks like a horseshoe. Most of the mass is concentrated at $\kappa_i = 0$ and $\kappa_i = 1$.

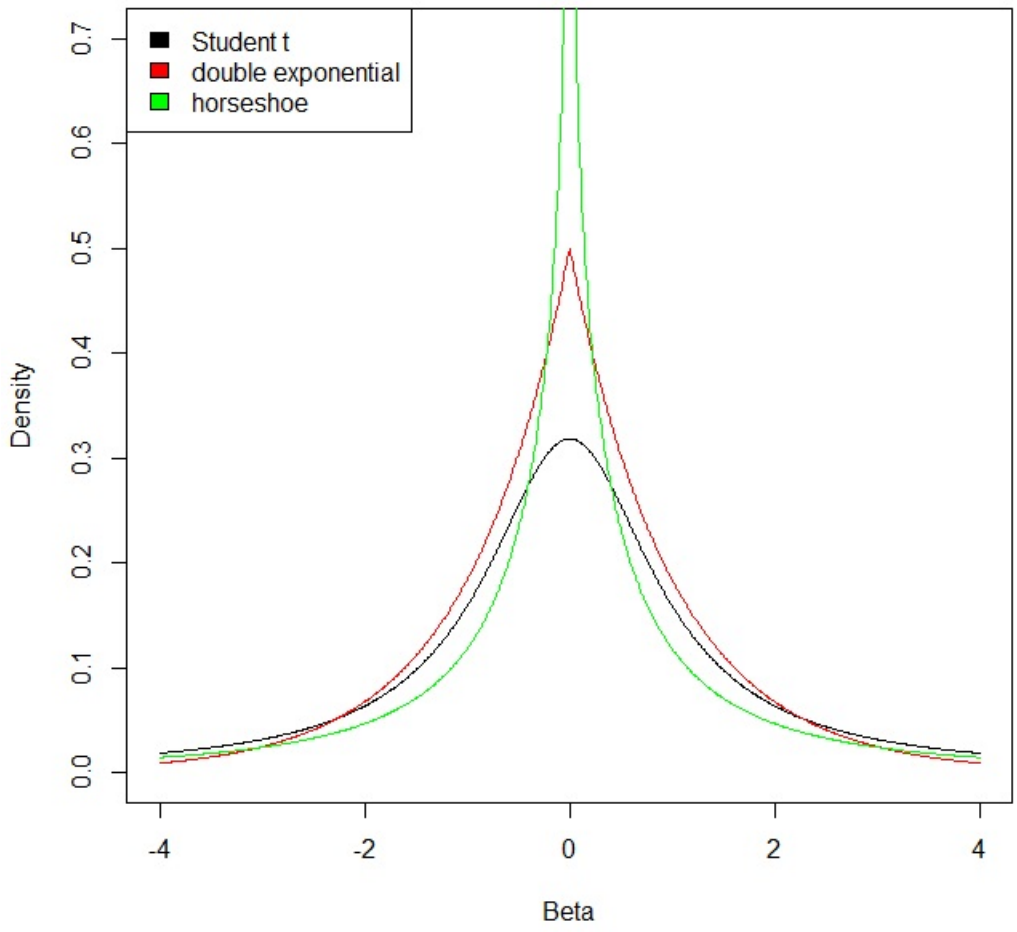

Figure 2.1: Different prior on β_i

In (2.7), if $\kappa_i = 0$, then $E(\beta_i|y_i, \lambda_i^2) = y_i$, which means there is no shrinkage. If $\kappa_i = 1$, then $E(\beta_i|y_i, \lambda_i^2) = 0$, which means total shrinkage, see Piironen and Vehtari (2017).

In Bayesian linear regression, we usually assume that regression coefficients β_is are independently normally distributed. In this case, the spike and slab prior can be rewritten as

$$\beta_i \sim \rho N(0, c^2) + (1 - \rho)\delta_0(\beta_i),$$

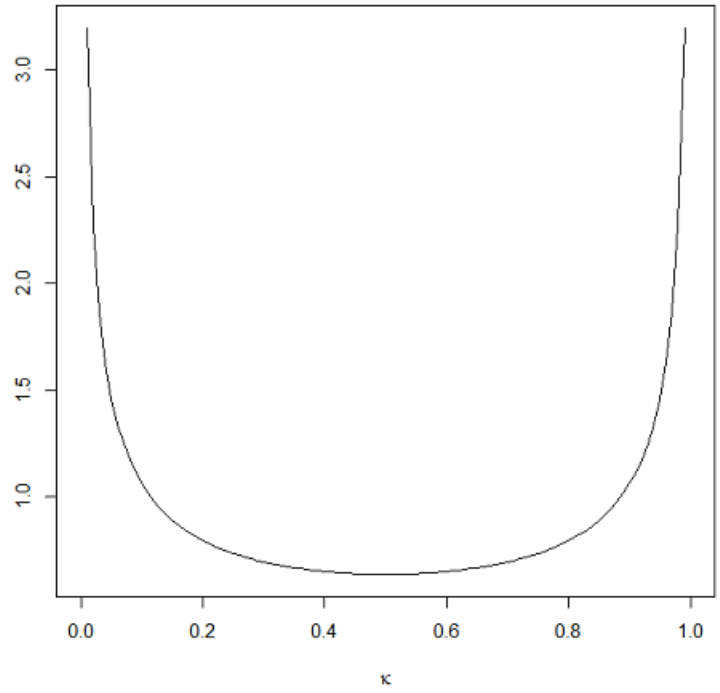

Figure 2.2: Density of the κ_i

where

$$\delta_0(\beta_i) := \begin{cases} 1 & \text{if} \quad \beta_i = 0 \\ 0 & \text{otherwise,} \end{cases}$$

which concentrates all its mass at zero and makes those β_i corresponding to unimportant covariates shrink to zero.

If the value of β_i is not close to zero, then by the slab component, $\beta_i \sim N(0, c^2)$, and

$$E(\beta_i|y_i) = \frac{c^2}{1+c^2} y_i = \Big(1 - \frac{1}{1+c^2}\Big) y_i,$$

so the shrinkage factor is $\kappa_i = \frac{1}{1+c^2}$, which has the same form as $\kappa_i = \frac{1}{1+\lambda_i^2}$. If we set $\lambda_i = c$, then both priors will be the same, which means under this setting, both the horseshoe prior and the spike and slab prior will assign the same amount of shrinkage to the nonzero β_is. If the value of β_i is close to zero, then by the spike component, it will be shrunk to zero,

which means total shrinkage ($\kappa_i = 1$). Thus, the horseshoe prior can closely mimic the spike and slab prior.

The performance of the spike and slab prior mainly depends on the choice of $g(\cdot)$ and ρ. A spike and slab prior is often considered as the 'gold standard' for variable selection. The horseshoe prior often performs better than the spike and slab prior in terms of the mixing of the MCMC (Markov chain Monte Carlo) algorithm. For more details about the Spike and Slab prior and the Horseshoe prior, see Piironen and Vehtari (2017), who also proposed the regularized horseshoe prior as an improvement of the horseshoe prior.

The regularized horseshoe prior

The setting of the regularized horseshoe prior is as follows.

$$\begin{aligned} &\beta_i|\lambda_i, \tau, c \sim N(0, \tilde{\lambda_i}^2\tau^2), \quad \lambda_i \sim C^+(0,1), \quad \tilde{\lambda_i}^2 = \frac{c^2\lambda_i^2}{c^2+\tau^2\lambda_i^2}, \\ &\tau \sim C^+(0, \tau_0), \quad \tau_0 = \frac{p_0}{L-p_0}\frac{\iota}{\sqrt{M}}. \end{aligned} \tag{2.8}$$

In (2.8), p_0 is the number of relevant covariates expected, and L is the total number of coefficients of basis functions. In linear regression, ι is the standard deviation of the error term. In the context of spectral estimation, it is the standard deviation of η_j in (2.1), which is equal to $\frac{\pi}{\sqrt{6}}$.

Compared with the horseshoe prior with $\sigma = 1$, we can see that if β_i is close to 0, then its corresponding local shrinkage parameter λ_i will be small, thus $\tau^2\lambda_i^2 << c^2$. In this case $\tilde{\lambda_i}^2 = \frac{c^2\lambda_i^2}{c^2+\tau^2\lambda_i^2} \approx \lambda_i^2$, which leads to $\beta_i|\lambda_i, \tau, c \sim N(0, \lambda_i^2\tau^2)$, the same as the horseshoe prior.

When β_i has a large value, then its corresponding local shrinkage parameter λ_i will be large, thus $\tau^2\lambda_i^2 >> c^2$. In this case, $\tilde{\lambda_i}^2 = \frac{c^2\lambda_i^2}{c^2+\tau^2\lambda_i^2} \approx \frac{c^2}{\tau^2}$, which leads to $\beta_i|\lambda_i, \tau, c \sim N(0, \frac{c^2}{\tau^2}\tau^2) = N(0, c^2)$. This is identical to the slab term in the spike and slab prior.

Now we can see that, both the horseshoe prior and the regularized horseshoe prior will shrink β_is that are close to 0 in a similar fashion. However, large β_is will be regularized by the regularized horseshoe prior, but the horseshoe prior will not do any regularization.

Chapter 3

Baysian framework

The Posterior Distribution

Combining the likelihood with the prior distributions yields the posterior distribution needed for Bayesian inference, i.e.

$$\text{posterior} \propto \text{prior} \times \text{likelihood}. \tag{3.1}$$

In our case, the likelihood is the Whittle likelihood (2.3), and we let $y_m = \log I(\omega_m)$. Let $\boldsymbol{q}_m = (1, q_{m1}, \ldots, q_{mL})'$, where $q_{m1}, q_{m2}, \ldots, q_{mL}$ are basis functions evaluated at ω_m and let $\boldsymbol{\beta} = (\alpha_0, \beta_1, \ldots, \beta_L)'$ be a vector of unknown coefficients such that $\log f(\omega_m) = \mathbf{q}_\mathbf{m}'\boldsymbol{\beta}$. We then rewrite the Whittle likelihood as

$$\exp\Big\{-\sum_{m=1}^{M}[\boldsymbol{q}_m'\boldsymbol{\beta} + \exp\{y_m - \boldsymbol{q}_m'\boldsymbol{\beta}\}]\Big\}. \tag{3.2}$$

The prior we place on $\boldsymbol{\beta}$ depends on what model we use.

The priors for P-splines

In this case, the posterior distribution is given by

$$P(\beta, \tau|y) \propto \exp\Big\{-\sum_{m=1}^{M}[\boldsymbol{q}_m'\boldsymbol{\beta} + \exp\{y_m - \boldsymbol{q}_m'\boldsymbol{\beta}\}]\Big\} \times P(\boldsymbol{\beta}) \times P(\tau),$$

where

1. The prior on $\boldsymbol{\beta}$ satisfies $P(\boldsymbol{\beta}) = P(\alpha_0) \times P(\beta_1) \times P(\beta_2) \times \cdots \times P(\beta_L)$, where

$$\alpha_0 \sim N(0, 10^2), \beta_1 \sim N(0, \tau^2), \beta_\rho = \beta_{\rho-1} + u_\rho, u_\rho \sim N(0, \tau^2),$$

for $\rho = 2, 3, \ldots, L$.

2. The prior on τ is Half-$t_3(0, 10^3)$, where Half-$t_3(0, 10^3)$ means the half t distribution with degrees of freedom 3, location parameter 0, and scale parameter 10^3. If $\tau \sim$ Half-$t_\nu(\mu, \sigma)$, then its density function is

$$g(\tau) = \frac{2\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\nu\pi\sigma^2}}\Big(1 + \frac{1}{\nu}\frac{\tau^2}{\sigma^2}\Big)^{-\frac{\nu+1}{2}} \quad \text{for } \tau \geq 0.$$

The Horseshoe prior

In this case, the posterior distribution is given by

$$P(\beta, \tau, \lambda|y) \propto \exp\Big\{-\sum_{m=1}^{M}[\boldsymbol{q}_m'\boldsymbol{\beta} + \exp\{y_m - \boldsymbol{q}_m'\boldsymbol{\beta}\}]\Big\} \times P(\boldsymbol{\beta}) \times P(\tau) \times P(\lambda).$$

1. The prior on $\boldsymbol{\beta}$ satisfies $P(\boldsymbol{\beta}) = P(\alpha_0) \times P(\beta_1) \times P(\beta_2) \times \cdots \times P(\beta_L)$, where

$$\alpha_0 \sim N(0, 10^2), \beta_\rho \sim N(0, \lambda_\rho^2\tau^2),$$

for $\rho = 1, 2, \ldots, L$.

2. The prior on λ_ρ is $\lambda_\rho|\sigma \sim C^+(0, \sigma)$. In our case, we let $\sigma = 1$, so $P(\lambda_\rho|\sigma) = \frac{2}{\pi(1+\lambda_\rho^2)}$.

3. The prior on τ is $\tau|\eta \sim C^+(0, \eta)$. In our case, we let $\eta = 1$, so $P(\tau|\eta) = \frac{2}{\pi(1+\tau^2)}$.

The regularized Horseshoe prior

In this case, the posterior distribution is given by

$$P(\beta, \tau, \lambda, c^2, \iota|y) \propto \exp\Big\{-\sum_{m=1}^{M}[\boldsymbol{q}_m'\boldsymbol{\beta} + \exp\{y_m - \boldsymbol{q}_m'\boldsymbol{\beta}\}]\Big\} \times P(\boldsymbol{\beta}) \times P(\tau) \times P(\lambda) \times P(c^2).$$

1. The prior on $\boldsymbol{\beta}$ satisfies $P(\boldsymbol{\beta}) = P(\alpha_0) \times P(\beta_1) \times P(\beta_2) \times \cdots \times P(\beta_L)$, where

$$\alpha_0 \sim N(0, 10^2), \beta_\rho \sim N(0, \tilde{\lambda_\rho}^2 \tau^2),$$

   for $\rho = 1, 2, \ldots, L$.

2. The $\tilde{\lambda_i}^2 = \frac{c^2 \lambda_i^2}{c^2 + \tau^2 \lambda_i^2}$ is a transformed parameter. The prior on λ_ρ is $\lambda_\rho \sim C^+(0, 1)$. The prior on c^2 is Inv-Gamma(15,375), where Inv-Gamma(15,375) means the inverse Gamma distribution with shape parameter 15 and scale parameter 375. It allows c^2 to take larger values, so the amount of shrinkage of large β_is will be small

3. The prior on τ is $\tau \sim C^+(0, \tau_0) = C^+(0, \frac{p_0}{L - p_0} \frac{\iota}{\sqrt{M}})$, where $\iota = \frac{\pi}{\sqrt{6}}$.

Hamiltonian Monte Carlo

The random walk nature of the Metropolis algorithm makes it slow to explore the parameter space and to converge to the target distribution (Gelman et al., 2013).

The Hamiltonian Monte Carlo (HMC) was proposed by Duane et al. (1987), and was first used in statistics by Neal (2011). HMC is based on Hamiltonian dynamics borrowed from physics to reduce the local random walk behaviour of the Metropolis algorithm.

Hamiltonian dynamics use an object's location $\boldsymbol{\beta}$ and momentum $\boldsymbol{\zeta}$ at time t to describe its motion in the system. Each location of the object is associated with potential energy $U(\boldsymbol{\beta})$, and for each momentum there is an associated kinetic energy $K(\boldsymbol{\zeta})$. The sum of these two types of energy is given by $H(\boldsymbol{\beta}, \boldsymbol{\zeta})$.

$$H(\boldsymbol{\beta}, \boldsymbol{\zeta}) = U(\boldsymbol{\beta}) + K(\boldsymbol{\zeta}), \tag{3.3}$$

where the H($\boldsymbol{\beta}$, $\boldsymbol{\zeta}$) is the total energy.

Equation (3.3) leads to the Hamiltonian equations

$$\frac{d\,\boldsymbol{\beta}}{d\,t} = \frac{d\,H}{d\,\boldsymbol{\zeta}} = \frac{d\,K(\boldsymbol{\zeta})}{d\,\boldsymbol{\zeta}},$$

$$\frac{d\,\boldsymbol{\zeta}}{d\,t} = -\frac{d\,H}{d\,\boldsymbol{\beta}} = -\frac{d\,U(\boldsymbol{\beta})}{d\,\boldsymbol{\beta}}.$$

Given $\frac{d\,K(\boldsymbol{\zeta})}{d\,\boldsymbol{\zeta}}$ and $\frac{d\,U(\boldsymbol{\beta})}{d\,\boldsymbol{\beta}}$ and a set of initial values of $\boldsymbol{\zeta}$ and $\boldsymbol{\beta}$, we can use the Hamiltonian equations to predict the location $\boldsymbol{\beta}$ and momentum $\boldsymbol{\zeta}$.

In HMC, we use the vector of unknown coefficients $\boldsymbol{\beta}$ as the location, and for each β_i in $\boldsymbol{\beta}$ we assign a corresponding momentum variable ζ_i. The potential energy $U(\boldsymbol{\beta})$ is the log-posterior density of $\boldsymbol{\beta}$, i.e., $U(\boldsymbol{\beta}) = \log P(\boldsymbol{\beta}|y)$. As for $\boldsymbol{\zeta}$, we assume it has a multivariate normal distribution with independent components, so its variance covariance matrix M is diagonal, i.e., $\zeta_i \sim N(0, M_{i,i})$ where $M_{i,i}$ is the ith diagonal element of M. The kinetic energy is given by $K(\boldsymbol{\zeta}) = \frac{1}{2}\boldsymbol{\beta}^T M^{-1}\boldsymbol{\beta}$.

At the beginning of the HMC iterations, draw random values of $\boldsymbol{\beta}$ and denote them as $\boldsymbol{\beta}^0$, where $\boldsymbol{\beta}^i$ is the value of $\boldsymbol{\beta}$ after the ith HMC iteration.

Then in the ith HMC iteration ($i = 1, 2, \ldots$)

1. Get current values for this iteration. Let $\boldsymbol{\beta} = \boldsymbol{\beta}^{i-1}$, draw a random sample of $\boldsymbol{\zeta}$ from its posterior distribution, $\boldsymbol{\zeta} \sim N(0, M)$ and denote it as $\boldsymbol{\zeta}^0$, let $\boldsymbol{\zeta} = \boldsymbol{\zeta}^0$.

2. Propose a new candidate for the next position. Update $\boldsymbol{\beta}$ and $\boldsymbol{\zeta}$ by R 'leapfrog steps', each scaled by a factor ϵ. In each leapfrog step, we do

   (a) $\boldsymbol{\zeta} = \boldsymbol{\zeta} - \frac{1}{2}\epsilon\frac{d\,\log P(\boldsymbol{\beta}|y)}{d\,\boldsymbol{\beta}}$. Use $\frac{d\,\log P(\boldsymbol{\beta}|y)}{d\,\boldsymbol{\beta}}$ to update $\boldsymbol{\zeta}$ for half a step.

   (b) $\boldsymbol{\beta} = \boldsymbol{\beta} + \epsilon M^{-1}\boldsymbol{\zeta}$. Use ζ and M^{-1} to update $\boldsymbol{\beta}$ for a whole step.

   (c) $\boldsymbol{\zeta} = \boldsymbol{\zeta} - \frac{1}{2}\epsilon\frac{d\,\log P(\boldsymbol{\beta}|y)}{d\,\boldsymbol{\beta}}$. Use $\frac{d\,\log P(\boldsymbol{\beta}|y)}{d\,\boldsymbol{\beta}}$ to update $\boldsymbol{\zeta}$ for another half step.

   After (a), (b) and (c), we have updated both β and ζ by a whole step. Steps (a), (b) and (c) together are called a leapfrog step. At the end of R leapfrog steps, the values of $\boldsymbol{\beta}$ and $\boldsymbol{\zeta}$ are denoted by $\boldsymbol{\beta}^*$ and $\boldsymbol{\zeta}^*$.

3. Compute the acceptance ratio r.

$$r = \frac{P(\boldsymbol{\beta}^*|y)P(\boldsymbol{\zeta}^*)}{P(\boldsymbol{\beta}^{i-1}|y)P(\boldsymbol{\zeta}^0)}.$$

4. Update

$$\boldsymbol{\beta}_i := \begin{cases} \boldsymbol{\beta}^* & \text{with probability min(r,1)} \\ \boldsymbol{\beta}^{i-1} & \text{otherwise.} \end{cases}$$

The main difference between the Metropolis-Hastings method and Hamiltonian Monte Carlo is how they propose the candidate for the next iteration. In the Metropolis-Hastings algorithm, the proposal distribution only depends on $\boldsymbol{\beta}$, but in the Hamiltonian Monte Carlo, the proposal distribution is the joint distribution $P(\boldsymbol{\beta}|y)P(\boldsymbol{\zeta})$, which depends also on $\boldsymbol{\zeta}$. Beside that, the Hamiltonian Monte Carlo also uses the gradient of $\log P(\boldsymbol{\beta}|y)$, so compared with the Metropolis-Hastings algorithm, each single iteration of Hamiltonian Monte Carlo will be more costly but with a higher acceptance rate which allows Hamiltonian Monte Carlo to move faster and reach convergence earlier.

Chapter 4

Spatially Adaptive Estimation of Time Series in the Time Domain Using the Dynamic Shrinkage Prior

To motivate the dynamic shrinkage prior (DSP), proposed by Kowal et al. (2019), consider figures 4.1 and 4.2 which display on their left panels true spatially inhomogeneous signals. These were used by Donoho and Johnstone (1994) as examples of spatially inhomogeneous signals. The right panels display data generated by adding noise to these signals, along with two types of fitted curves. For each signal, there are 128 equally spaced sample points, in the time interval $[0, 1]$. The red lines are based on smoothing splines while the green ones were fit using the single-component nonparametric regression with the DSP, implemented in the dsp R package (Kowal (2020)). The DSP is explained later in this chapter. As evident from these plots, the smoothing splines which are not spatially adaptive, miss some important features of the signals. The single-component nonparametric regression with the DSP does a good job due to its spatial adaptivity. This method allows for different amounts of smoothing in different intervals. This property is advantageous when fitting data from spatially inhomogeneous signals like the ones shown in these plots. In Chapter 5 we propose using the DSP in the *frequency domain* for estimating spectra of stationary time series.

The goal of trend filtering is to smooth out a time series by filtering out the noise. Consider the single-component nonparametric regression model

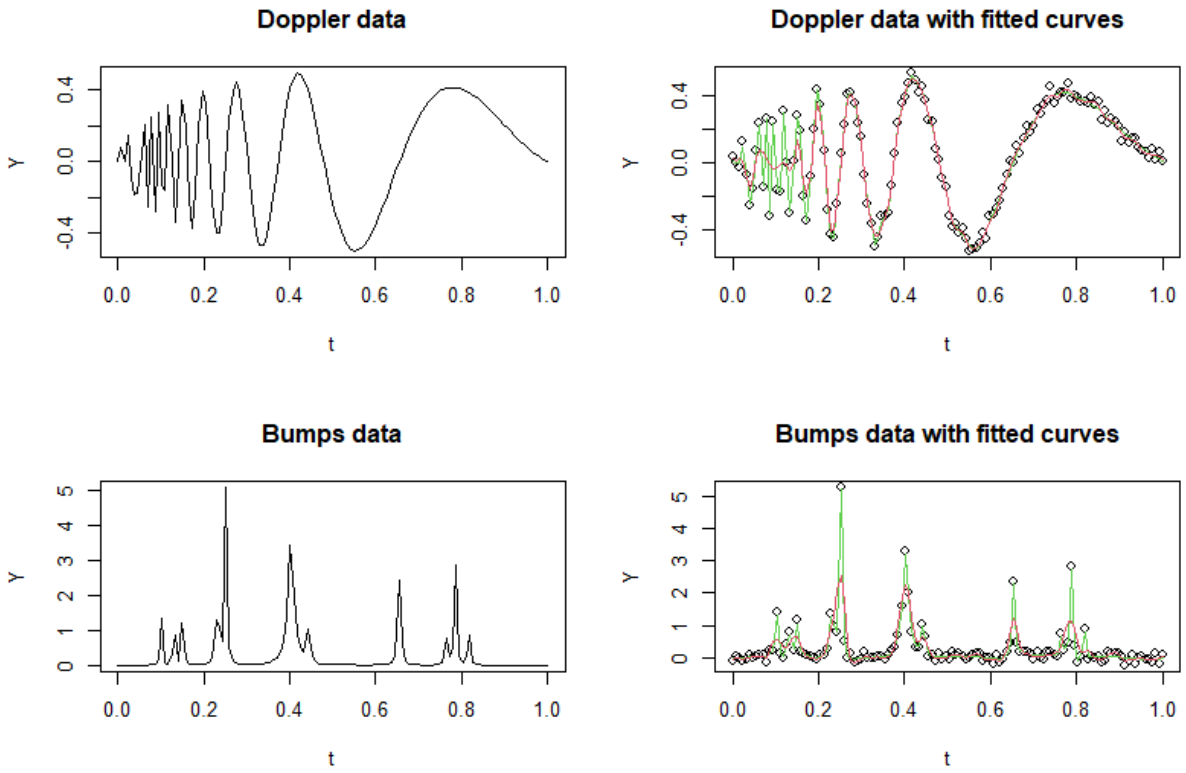

Figure 4.1: Left: Doppler and Bumps signals. Right: Data along with fitted curves. The red lines are based on smoothing splines while the green ones were fit using the single-component nonparametric regression with the DSP.

$$y_i = \beta_i + \epsilon_i, \quad \epsilon \mid \sigma_\epsilon \overset{ind}{\sim} \mathrm{N}(0, \sigma_\epsilon^2), \quad \text{for } 1 \leq i \leq M.$$

Given an observed time series $\{y_i\}$, the goal is to filter out the noise and estimate the smooth filtered time series.

L_1 trend filtering model

Kim et al. (2009) find the βs that minimize the following objective function

$$\tfrac{1}{2}\textstyle\sum_{i=1}^{M}(y_i - \beta_i)^2 + \lambda \sum_{i=2}^{M-1} |\beta_{i-1} - 2\beta_i + \beta_{i+1}|. \tag{4.1}$$

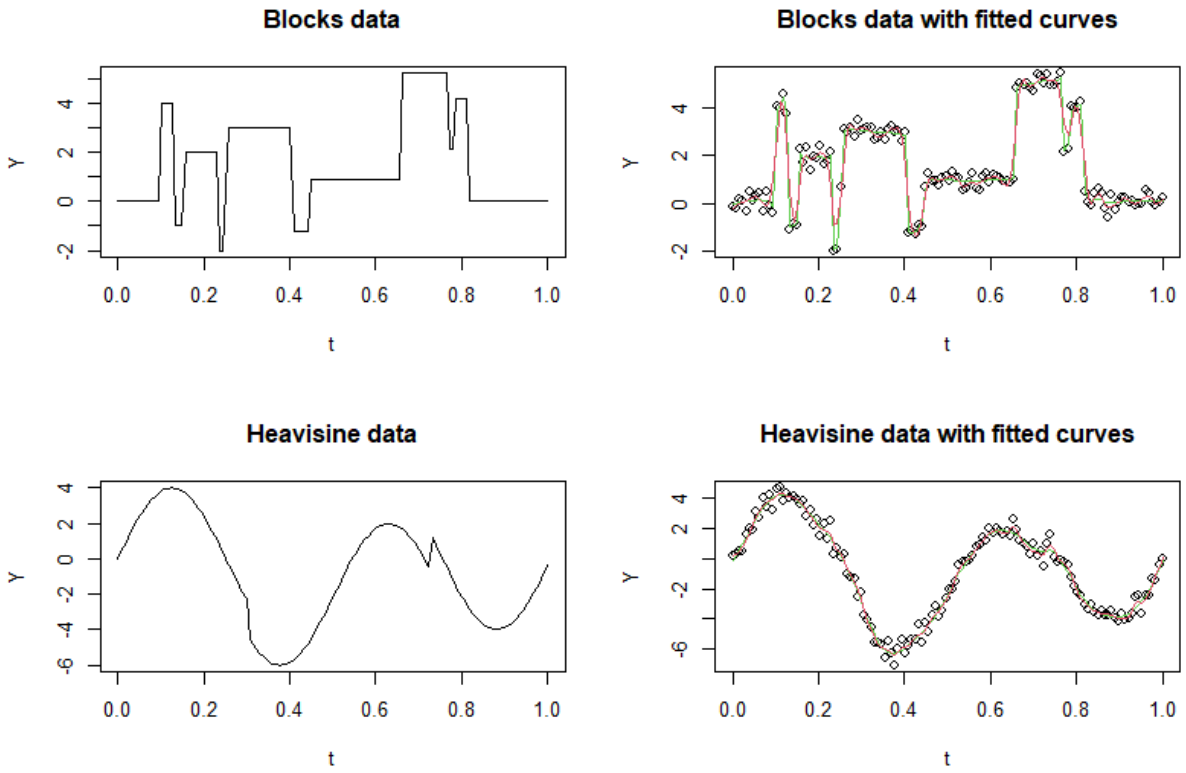

Figure 4.2: Left: Blocks and Heavisine signals. Right: Data along with fitted curves. The red lines are based on smoothing splines while the green ones were fit using the single-component nonparametric regression with the DSP.

In (4.1), $\sum_{i=1}^{M}(y_i - \beta_i)^2$ is the residual sum of squares, which measures goodness of fit. The expression $\sum_{i=2}^{n-1}|\beta_{i-1} - 2\beta_i + \beta_{i+1}|$ is the sum of the absolute values of the second-order differences, often used to impose smoothness on the values of β_i. First-order or higher orders may be used instead. The parameter λ is a global smoothing parameter, which controls the trade-off between the two objectives of minimizing the residuals and maximizing smoothness. Since λ is a constant that does not depend on i, the same amount of smoothing is applied everywhere.

The Baysian trend filter model

Faulkner and Minin (2018), proposed the following model

$$\begin{aligned} &y_i = \beta_i + \epsilon_i, \;\; \epsilon_i \mid \sigma_\epsilon \overset{ind}{\sim} \mathrm{N}(0, \sigma_\epsilon^2), \;\; \text{for } 1 \leq i \leq M, \\ &\Delta^2 \beta_{i+1} = w_i, \;\; w_i \mid \tau, \lambda_i \overset{ind}{\sim} \mathrm{N}(0, \tau^2 \lambda_i^2), \;\; \text{for } 1 \leq i \leq M. \end{aligned} \tag{4.2}$$

Model (4.2) is a Bayesian adaptation of the L_1 trend filtering model (4.1). The prior placed on the second-order difference of $\{\beta_i\}$, $\Delta^2\beta_{i+1}$, is the horseshoe prior, proposed by Carvalho et al. (2010), which is a global-local shrinkage prior. The shrinkage imposed by this prior induces a locally adaptive smoothing of the trend, where the local scale parameters $\{\lambda_i\}$ are assumed to be iid. Kowal et al. (2019) propose the dynamic shrinkage prior (DSP) which does not assume independence of the $\{\lambda_i\}$.

The dynamic shrinkage prior (DSP)

.

Let $h_i = \log(\tau^2 \lambda_i^2)$ and define

$$\begin{aligned} &h_{i+1} = \mu + \phi(h_i - \mu) + \eta_i, \\ &\eta_i \overset{iid}{\sim} Z(\tfrac{1}{2}, \tfrac{1}{2}, 0, 1), \end{aligned} \tag{4.3}$$

where $\mu = \log(\tau^2)$ and $\phi(h_{i-1} - \mu) + \eta_{i-1} = \log(\lambda_i^2)$. This formulation induces dependence between the λ_i and λ_{i+1}. If $\phi = 0$, then $h_i = \mu + \eta_i$, and since $\eta_i \overset{iid}{\sim} Z(\alpha, \beta, 0, 1)$, the h_i are i.i.d, which is the standard global-local prior, i.e., there is no extra spatial adaptivity. When $\phi > 0$, the first equation of (4.3) shows that the $\{h_i\}$ follow an AR(1) model, and the dependence between λ_i and λ_{i+1} is controlled by the AR(1) coefficient ϕ. Because ϕ is positive, the correlation between λ_i and λ_{i+1} is positive. The larger ϕ, the stronger the relation between λ_i and λ_{i+1}. This means that the value of λ_{i+1} will be more likely to be close to λ_i so the degree of smoothing in adjacent intervals will not change a lot.

As for the parameter τ, we set $\tau \sim C^+(0, \gamma)$, where $\gamma = \frac{\sigma_\epsilon}{\sqrt{M}}$ and C^+ is the half-Cauchy distribution with pdf $P(\tau) = \frac{2}{\pi \cdot \gamma} \cdot \frac{1}{1 + (\frac{\tau}{\gamma})^2}$. In the Horseshoe prior $\beta_i \sim \mathrm{N}(0, \tau^2 \lambda_i^2)$, so the

conditional posterior distribution of β_i is conditional on the value of τ or τ^2 directly. But in the DSP, the evolution equation (4.3) is used to get h_{i+1} for $i = 1, 2, 3, \ldots, M-1$. Thus when implementing the DSP, we work with $\mu = \log(\tau^2)$ rather than with τ^2. For this reason, we now derive the distribution of μ. Since $\tau \sim C^+(0,1)$, it follows that the density function of $A = \tau^2$ is

$$P(A) = \frac{2}{\pi \cdot \gamma} \cdot \frac{1}{1 + \frac{A}{\gamma^2}} \cdot \frac{1}{2} \frac{1}{\sqrt{A}}$$

$$= \frac{1}{\pi\gamma} \cdot \frac{\gamma^2}{\gamma^2 + A} \cdot \frac{1}{\sqrt{A}}$$

$$= \frac{\gamma}{\pi\sqrt{A}(A^2 + \gamma^2)}.$$

Now, letting $\mu = \log(A) = \log(\tau^2) \implies A = \exp(\mu)$, we obtain

$$P(\mu) = \frac{1}{\pi} \frac{e^{\frac{1}{2}\mu - \log(\gamma)}}{1 + e^{\mu - 2\log(\gamma)}} = \frac{1}{\pi} \cdot \frac{e^{\frac{1}{2}(\mu - 2\log(\gamma))}}{1 + e^{\mu - 2\log(\gamma)}}.$$

By Theorem 1 of Polson et al. (2013), let $p(\xi_\mu)$ denote the pdf of the Pólya-Gamma random variable $\xi_\mu \sim PG(b; 0),\ b > 0$. Then the following integral identity holds for all $a \in \mathbb{R}$.

$$\frac{(e^\psi)^a}{(1 + e^\psi)^b} = 2^{-b} e^{\zeta\psi} \int_0^\infty e^{-\xi_\mu \frac{\psi^2}{2}} p(\xi_\mu) d\xi_\mu,$$

where $\zeta = a - \frac{b}{2}$. Letting $\psi = \mu - 2\log(\gamma)$, we see that

$$P(\psi) = \frac{1}{\pi} \cdot \frac{e^{\frac{1}{2}(\mu - 2\log(\gamma))}}{1 + e^{\mu - 2\log(\gamma)}} = \frac{1}{\pi} \cdot \frac{(e^\psi)^{\frac{1}{2}}}{(1 + e^\psi)^1}$$

$$= \frac{1}{2\pi} e^{(\frac{1}{2} - \frac{1}{2})\psi} \int_0^\infty e^{-\xi_\mu \frac{\psi^2}{2}} p(\xi_\mu) d\xi_\mu = \frac{1}{2\pi} \int_0^\infty e^{-\xi_\mu \frac{\psi^2}{2}} p(\xi_\mu) d\xi_\mu.$$

Thus, given $\xi_\mu \sim PG(b, 0)$, $\psi \mid \xi_\mu \sim \mathrm{N}(0, \xi_\mu^{-1})$. Also, the conditional distribution $\xi_\mu | \psi \sim PG(b, \psi)$.

In our case, $\psi = \mu - 2\ln(\gamma)$, $a = \frac{1}{2}$ and $b = 1$, then $\zeta = \frac{1}{2} - \frac{1}{2} = 0$. So $P(\mu) = \frac{1}{2\pi}\int_0^\infty e^{-\frac{(\mu-2\log(\gamma))^2\xi_\mu}{2}} p(\xi_\mu)d\xi_\mu$. Thus $(\mu \mid \xi_\mu, \sigma_\epsilon) \sim \mathrm{N}(2\log(\gamma), \xi_\mu^{-1}) = \mathrm{N}(\log(\frac{\sigma_\epsilon^2}{M}), \xi_\mu^{-1})$ and $\xi_\mu \sim PG(1, 0)$.

Since $w_i = \Delta^2\beta_{i+1}$ and $w_i \sim \mathrm{N}(0, \tau^2\lambda_i^2) = \mathrm{N}(0, \exp(h_i))$, we see that $\frac{w_i}{\exp(\frac{h_i}{2})} \sim \mathrm{N}(0, 1)$ and $\frac{w_i^2}{\exp(h_i)} \sim \chi_1^2$. Taking the log, we obtain $\log(w_i^2) - h_i \sim \log(\chi_1^2) \iff \log(w_i^2) = h_i + \log(\chi_1^2)$. Kastner and Frühwirth-Schnatter (2014) use this expression to write the joint distribution of the h_i. The distribution of $\log(\epsilon_i^2)$ can be approximated by the 10-component mixture of normal distributions proposed in Omori et al. (2007). Conditional on the mixture component indicators s_i, $\log(\epsilon_i^2)|s_i \sim \mathrm{N}(m_{s_i}, \nu_{s_i})$, where m_j, p_j and $\nu_j, j = 1, \ldots, 10$, are the pre-specified means, weights and variances of the 10-component Gaussian mixture provided in Omori et al. (2007). Thus $\log(w_i^2) = h_i + \log(\epsilon_i^2)|s_i$ implies $\log(w_i^2) \sim \mathrm{N}(h_i + m_{s_i}, \nu_{s_i})$. In practice, to avoid numerical issues when w_i^2 is too small, we add a small offset $c = 10^{-4}$ to w_i^2, resulting in $\log(w_i^2 + c) \sim \mathrm{N}(h_i + m_{s_i}, \nu_{s_i})$.

A random variable from $\eta_i \sim Z(\frac{1}{2}, \frac{1}{2}, 0, 1)$, can be generated by drawing from $\eta_i \mid \xi_i \sim \mathrm{N}(0, \xi_i^{-1})$ and $\xi_i \sim PG(1, 0)$.

As for ϕ, we let $\frac{\phi+1}{2} \sim \mathrm{Beta}(10, 2)$, which places most of the mass of the density of ϕ on (0, 1), so ϕ has a prior mean of 2/3 and a prior mode of 4/5.

For σ_ϵ, we use Jeffreys' prior, i.e $p(\sigma_\epsilon) \propto \frac{1}{\sigma_\epsilon}$.

The setting of the DSP is summarized as follows.

$$
\begin{aligned}
&\Delta^2\beta_{i+1} = w_i, \;\; w_i|\tau, \lambda_i \overset{ind}{\sim} \mathrm{N}(0, \tau^2\lambda_i^2), \;\; \text{for } 1 \le i \le M \\
&\tau \sim C^+(0, \tfrac{\sigma_\epsilon}{\sqrt{M}}), \mu = \log(\tau^2) \implies (\mu|\sigma_\epsilon, \xi_\mu^{-1}) \sim \mathrm{N}(\log(\tfrac{\sigma_\epsilon^2}{M}), \xi_\mu^{-1}), \;\; \xi_\mu \sim PG(1, 0), \\
&\eta_i|\xi_i \sim \mathrm{N}(0, \xi_i^{-1}), \;\; \xi_i \overset{iid}{\sim} PG(1, 0), \;\; i = 1, 2, \ldots, M \\
&\tfrac{\phi+1}{2} \sim \mathrm{Beta}(10, 2), \;\; p(\sigma_\epsilon) \propto \tfrac{1}{\sigma_\epsilon^2}.
\end{aligned}
\tag{4.4}
$$

As mentioned earlier, in the DSP, the local parameters λ_i depend on the AR(1) coeffecient ϕ, so the shrinkage paramter κ_t introduced in Chapet 2 also depends on ϕ. Figure 4.3 below displays simulation-based estimates of the stationary distribution of κ_t for various AR(1) coefficients ϕ. The blue line represents the density of the shrinkage paramenter κ_t

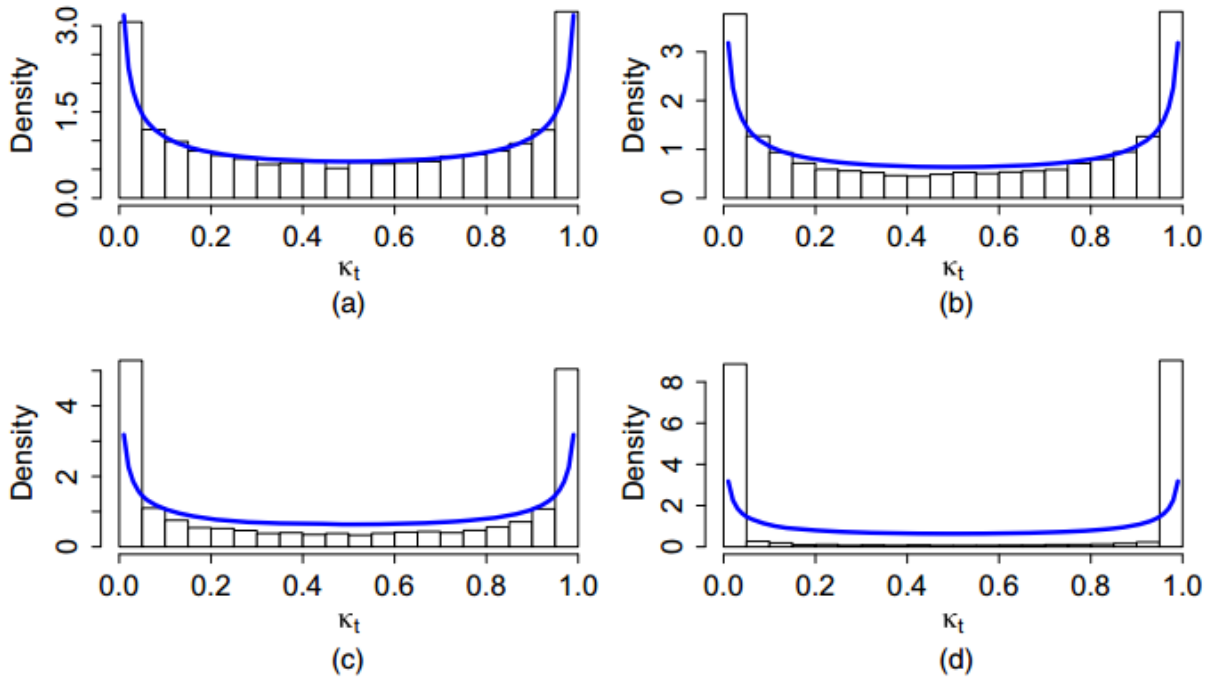

Figure 4.3: Blue line: density of κ_t of the horseshoe prior. Histograms: densities of κ_t for the DSP. (a) $\phi = 0.25$, (b) $\phi = 0.5$, (c) $\phi = 0.75$, (d) $\phi = 0.99$. The plot is from Kowal et al. (2019).

of the horseshoe prior, i.e., $\kappa_t \sim \text{Beta}(\frac{1}{2}, \frac{1}{2})$. The histograms show the densities of κ_t in the DSP for different values of ϕ: (a) $\phi = 0.25$, (b) $\phi = 0.5$, (c) $\phi = 0.75$, (d) $\phi = 0.99$. We see that when ϕ is close to 0, which means weak dependence between λ_i and λ_{i+1}, the DSP works very similar to the horseshoe prior, where there is no dependence between λ_i and λ_{i+1}. But when ϕ is close to 1, compared with horseshoe prior, the density of κ_t in the DSP gives more mass to values near 0 (no shrinkage) and 1 (maximum shrinkage).

Chapter 5

Spatially Adaptive Estimation in the Frequency Domain Using the Dynamic Shrinkage Prior

Kowal et al. (2019) proposed the dynamic shrinkage prior (DSP) but applied it only in the time domain. In this paper our goal is to apply the DSP to the frequency domain. By using the DSP, we aim at finding a spatially adaptive estimate of the spectrum of stationary time series.

5.1 The Posterior Distribution

Multiplying the likelihood by the prior distributions yields the posterior distribution needed for Bayesian inference. In our case, the likelihood is the Whittle likelihood (2.3). Given an observed time series $\{x_i\}$, we apply to it the discrete Fourier transform (DFT) introduced in Chapter 1 and denoted it by $d(\omega_i)$. Based on the $d(\omega_j)$ we then calculate the periodogram $I(\omega_i)$ as described in Chapter 1. In Chapter 2 we mentioned that when the sample size is large, we have approximately $I(\omega_i) \sim \text{Expon}\Big(f(\omega_i)\Big)$ from which the Whittle likelihood (2.3) follows. Letting $y_i = \log(I(\omega_i))$, the Whittle likelihood is given by

$$\exp\Big\{-\sum_{i=1}^{M}[\log f(\omega_i) + \exp\{y_i - \log f(\omega_i)\}]\Big\}.$$

5.2 Single-component nonparametric estimation

We start with the single-component nonparametric estimation, where $\log f(\omega_i)$ is modeled by β_i. Under this setting the Whittle likelihood becomes

$$\exp\Big\{-\sum_{i=1}^{M}[\beta_i + \exp\{y_i - \beta_i\}]\Big\}. \tag{5.1}$$

5.2.1 The conditional posterior distribution of $\boldsymbol{\beta}$

The prior on $\boldsymbol{\beta}$ is

$$P(\boldsymbol{D}_2\boldsymbol{\beta}|\Sigma_w) = (2\pi)^{-\frac{M}{2}}\det(\Sigma_w)^{-\frac{1}{2}}\exp\{-\frac{1}{2}(\boldsymbol{D}_2\boldsymbol{\beta})^T\Sigma_w^{-1}(\boldsymbol{D}_2\boldsymbol{\beta})\},$$

where $\boldsymbol{\beta} = \Big[\beta_1, \quad \beta_2, \quad \cdots, \quad \beta_M\Big]^T$ is an $M \times 1$ vector and

$$D_2 = \begin{bmatrix} 1 & 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 1 & 0 & 0 & \cdots & 0 & 0 \\ 1 & -2 & 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & -2 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & -2 & 1 \end{bmatrix},$$

which is an $M \times M$ matrix. $\Sigma_\omega = diag(\{\sigma^2_{\omega_i}\}_{i=1}^M)$ for $\sigma^2_{\omega_i} = \tau^2\lambda_i^2$.

Since it is the second-order difference, the priors on β_1 and β_2 are $\beta_i \sim \mathrm{N}(0, a_i^2), \;\; i = 1, 2$. This guarantees that the prior is proper pdf. The conditional posterior of $\boldsymbol{\beta}$ is given by

$$\exp\{-\sum_{i=1}^{M}[\beta_i + \exp\{y_i - \beta_i\}]\} \times \exp\{-\frac{1}{2}(\boldsymbol{D}_2\boldsymbol{\beta})^T\Sigma_w^{-1}(\boldsymbol{D}_2\boldsymbol{\beta})\} \times \exp\{-\frac{\beta_1^2}{2a_1^2}\} \times \exp\{-\frac{\beta_2^2}{2a_2^2}\}.$$

Taking the log yields

$$\log P = -\sum_{i=1}^{M}[\beta_i + \exp\{y_i - \beta_i\}] - \frac{1}{2}(\boldsymbol{D}_2\boldsymbol{\beta})^T\Sigma_w^{-1}(\boldsymbol{D}_2\boldsymbol{\beta}) - \frac{\beta_1^2}{2a_1^2} - \frac{\beta_2^2}{2a_2^2}.$$

We want to use the Metropolis-Hastings algorithm or the Hamiltonian Monte Carlo algorithm mentioned in Chapter 3 to draw samples of $\boldsymbol{\beta}$ from the conditional posterior distribution, for which the gradient and Hessian with respect to $\boldsymbol{\beta}$ are needed. Next we will derive these two.

$$\frac{\partial}{\partial \boldsymbol{\beta}} \log P = -\frac{\partial}{\partial \boldsymbol{\beta}} \{\sum_{i=1}^{M} [\beta_i + \exp\{y_i - \beta_i\}]\} - \frac{1}{2} \frac{\partial}{\partial \boldsymbol{\beta}} [(\boldsymbol{D}_2 \boldsymbol{\beta})^T \Sigma_w^{-1} (\boldsymbol{D}_2 \boldsymbol{\beta})] + \frac{\partial}{\partial \boldsymbol{\beta}} \{-\frac{\beta_1^2}{2a_1^2} - \frac{\beta_2^2}{2a_2^2}\}$$

$$= \begin{bmatrix} \exp(y_1 - \beta_1) - 1 \\ \exp(y_2 - \beta_2) - 1 \\ \vdots \\ \exp(y_M - \beta_M) - 1 \end{bmatrix} - \frac{1}{2} \frac{\partial}{\partial \boldsymbol{\beta}} [\boldsymbol{\beta}^T D_2^T \Sigma_w^{-1} D_2 \boldsymbol{\beta}] + \begin{bmatrix} -\frac{\beta_1}{a_1^2} \\ -\frac{\beta_1}{a_1^2} \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

$$= \begin{bmatrix} \exp(y_1 - \beta_1) - 1 - \frac{\beta_1}{a_1^2} \\ \exp(y_2 - \beta_2) - 1 - \frac{\beta_2}{a_2^2} \\ \vdots \\ \exp(y_M - \beta_M) - 1 \end{bmatrix} - \frac{1}{2} \frac{\partial}{\partial \boldsymbol{\beta}} [\boldsymbol{\beta}^T D_2^T \Sigma_w^{-1} D_2 \boldsymbol{\beta}].$$

Denote $D_2^T \Sigma_w^{-1} D_2$ by C, then since Σ_w^{-1} is diagonal, it is symmetric and $C^T = C$. It follows that $\frac{1}{2} \frac{\partial}{\partial \boldsymbol{\beta}} [\boldsymbol{\beta}^T C \boldsymbol{\beta}] = C\boldsymbol{\beta}$ and

$$\frac{\partial}{\partial \boldsymbol{\beta}} \log P = \begin{bmatrix} \exp(y_1 - \beta_1) - 1 - \frac{\beta_1}{a_1^2} \\ \exp(y_2 - \beta_2) - 1 - \frac{\beta_2}{a_2^2} \\ \vdots \\ \exp(y_M - \beta_M) - 1 \end{bmatrix} - C\boldsymbol{\beta}.$$

The Hessian is given by

$$\frac{\partial^2}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} \log P = \frac{\partial}{\partial \boldsymbol{\beta}} [\begin{bmatrix} \exp(y_1 - \beta_1) - 1 - \frac{\beta_1}{a_1^2} \\ \exp(y_2 - \beta_2) - 1 - \frac{\beta_2}{a_2^2} \\ \vdots \\ \exp(y_M - \beta_M) - 1 \end{bmatrix} - C\boldsymbol{\beta}]$$

$$= \begin{bmatrix} -\exp(y_1 - \beta_1) - \frac{1}{a_1^2} & 0 & \cdots & 0 & 0 \\ 0 & -\exp(y_2 - \beta_2) - \frac{1}{a_1^2} & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & -\exp(y_{M-1} - \beta_{M-1}) & 0 \\ 0 & 0 & \cdots & 0 & -\exp(y_M - \beta_M) \end{bmatrix}$$

$$-C,$$

which is an $M \times M$ matrix.

The derivation of the conditional posterior distributions for other parameters are similar to the one in Kowal et al. (2019). Details are provided in Section A.2

Chapter 6

Future Work

6.1 Future Work

6.1.1 Spectral estimation using the DSP in the single-component model

I will continue working on programming the sampling scheme for estimating the spectrum of a stationary time series using the DSP in the single-component model. As described in Chapter 5, in this case, the estimation is based on Whittle likelihood, given in Equation (5.1), where the log spectrum is modeled using the vector $\boldsymbol{\beta}$ and the DSP. I will use the Metropolis-Hastings algorithm or the Hamiltonian Monte Carlo algorithm mentioned in Chapter 3 to draw samples of $\boldsymbol{\beta}$ from the conditional posterior distribution. After finishing the coding, I will evaluate the model performance through simulations, as well as fitting the model to real date.

The Metropolis-Hastings algorithm we want to use is described below:

- Find the conditional posterior distribution of $\boldsymbol{\beta}$, as done in Chapter 5, and denote it by $P(\boldsymbol{\beta} \mid \ldots)$, where $\ldots$ represents parameters like Σ_w, a_1^2 and a_2^2 on which the conditional posterior distribution of $\boldsymbol{\beta}$ is conditioned.
- Calculate the gradient and the Hessian of $\log P(\boldsymbol{\beta} \mid \ldots)$, $\frac{\partial}{\partial \boldsymbol{\beta}} \log P(\boldsymbol{\beta} \mid \ldots)$ and $\frac{\partial^2}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} \log P(\boldsymbol{\beta} \mid \ldots)$.
- Find $\hat{\boldsymbol{\beta}}$ which maximizes $\log P(\boldsymbol{\beta} \mid \ldots)$.
- Propose from the multivariate normal distribution $\mathrm{N}(\hat{\boldsymbol{\beta}}, \Sigma_{\hat{\boldsymbol{\beta}}})$, where

$\Sigma_{\hat{\boldsymbol{\beta}}} = (-\frac{\partial^2}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T}|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}} \log P(\boldsymbol{\beta} \mid \ldots))^{-1}$. We denote the pdf of this multivariate normal distribution by $q(\cdot)$ and a proposed value by $\hat{\boldsymbol{\beta}}^p$. Let the current value be $\hat{\boldsymbol{\beta}}^c$.

- Calculate the acceptance probability as $a = \min\{1, \frac{P(\hat{\boldsymbol{\beta}}^p)}{P(\hat{\boldsymbol{\beta}}^c)} \frac{q(\hat{\boldsymbol{\beta}}^c)}{q(\hat{\boldsymbol{\beta}}^p)}\}$, where $P(\hat{\boldsymbol{\beta}}^p)$ and $P(\hat{\boldsymbol{\beta}}^c)$ are the values of the conditional posterior distribution $P(\boldsymbol{\beta} \mid \ldots)$ at $\hat{\boldsymbol{\beta}}^p$ and $\hat{\boldsymbol{\beta}}^c$, respectively. Similarly, $q(\hat{\boldsymbol{\beta}}^p)$ and $q(\hat{\boldsymbol{\beta}}^c)$ are the values of the proposal pdf at $\hat{\boldsymbol{\beta}}^p$ and $\hat{\boldsymbol{\beta}}^c$, respectively.

6.1.2 Spectral estimation using the DSP in the multi-component model

We will extend the single-component model to the multi-component model (Kowal et al. (2019)), using B-spline basis functions to model $\log f(\omega_i)$ by $\sum_{j=1}^{L} \beta_{i,j} x_{i,j}$, where $x_{i,j}$ is the j-th B-spline basis function evaluated at the Fourier frequency ω_i, and let $\boldsymbol{x}_i = (x_{i,1}, \ldots, x_{i,L})^T$ and $\boldsymbol{\beta}_i = (\beta_{i,1}, \ldots, \beta_{i,L})^T$ such that $\log f(\omega_i) = \boldsymbol{x}_i^T \boldsymbol{\beta}_i$. We then rewrite the Whittle likelihood as

$$\exp\Big\{-\sum_{i=1}^{M}[\boldsymbol{x}_i^T\boldsymbol{\beta}_i + \exp\{y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}_i\}]\Big\}. \tag{6.1}$$

Model (6.1) is a varying-coefficient model (Hastie and Tibshirani (1993)), which means that the coefficients are allowed to vary as smooth functions. The DSP is then applied to $\boldsymbol{\beta}_i$. The setting of the DSP for the multi-component model is described below.

$$\begin{aligned} &\Delta^2\boldsymbol{\beta}_{i+1} = \boldsymbol{\beta}_{i+1} - 2\boldsymbol{\beta}_i + \boldsymbol{\beta}_{i-1} = \boldsymbol{w}_i, \;\; (w_{i,j} \mid \tau_0, \tau_i, \lambda_{i,j}) \overset{ind}{\sim} \mathrm{N}(0, \tau_0^2\tau_i^2\lambda_{i,j}^2), \\ &h_{i,j} = \log(\tau_0^2\tau_i^2\lambda_{i,j}^2), \;\; \boldsymbol{h}_{i+1} = \boldsymbol{\mu} + \Phi(\boldsymbol{h}_i - \boldsymbol{\mu}) + \boldsymbol{\eta}_i, \end{aligned} \tag{6.2}$$

where $\boldsymbol{w}_i = (w_{i,1}, \ldots, w_{i,L})^T$, $\boldsymbol{h}_i = (h_{i,1}, \ldots, h_{i,L})^T$, $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_L)^T$, $\boldsymbol{\eta}_i = (\eta_{i,1}, \ldots, \eta_{i,L})^T$ and $\Phi = \mathrm{diag}(\phi_1, \ldots, \phi_M)$. As before, $(\eta_{i,j} \mid \xi_{i,j}) \sim \mathrm{N}(0, \xi_{i,j}^{-1})$ and $\xi_{i,j} \overset{ind}{\sim} PG(1, 0)$. The $\tau_i \sim C^+(0, 1)$ and $\tau_0 \sim C^+(0, \frac{\sigma_\epsilon}{\sqrt{ML}})$.

I will derive the conditional posterior distributions required for implementing MCMC methods for this model and write code for implementing it.

6.1.3 Other tasks

In the proposed future research we will extend spectral estimation to a single *nonstationary* time series (Rosen et al. (2012)), as well as to multiple time series with covariates.

6.2 Time Schedule of Future Research

Table 6.1: Time schedule

Tasks to complete	Approximate time
Finish coding the single-component model for spectrum estimation.	Jan - Feb 2021
Deriving the conditional posterior distributions for the multi-component model.	Feb - Mar 2021
Coding the multi-component model.	Mar 2021
Working on spectrum estimation of nonstationary time series.	Apr - May 2020
Deriving the algorithm for estimation for nonstationary time series.	Jun -July 2021
Implementing the estimation for nonstationary time series.	Aug - Sep 2021
Dissertation writing: Literature review	Oct 2021
Dissertation writing: Methodology	Nov 2022
Dissertation writing: Experiment	Dec 2022
Dissertation writing: Results	Jan -Feb 2022
Dissertation writing: Writing and revising	March 2022
Defense	Apr 2022
Submission for publication.	May 2022

References

Carter, C. K. and Kohn, R. (1997), "Semiparametric Bayesian Inference for Time Series with Mixed Spectra," *Journal of the Royal Statistical Society. Series B (Methodological)*, 59, 255–268.

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2010), "The horseshoe estimator for sparse signals," *Biometrika*, 97, 465–480.

Donoho, D. L. and Johnstone, I. M. (1994), "Ideal spatial adaptation by wavelet shrinkage," *Biometrika*, 81, 425–455.

Duane, S., Kennedy, A., Pendleton, B. J., and Roweth, D. (1987), "Hybrid Monte Carlo," *Physics Letters B*, 195, 216 – 222.

Eilers, P. H. C. and Marx, B. D. (1996), "Flexible Smoothing with B-splines and Penalties," *Statistical Science*, 11, 89–102.

Faulkner, J. R. and Minin, V. N. (2018), "Locally adaptive smoothing with Markov random fields and shrinkage priors," *Bayesian Anal.*, 13, 225–252.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., and Vehtari, A. (2013), *Bayesian Data Analysis*, Taylor & Francis Ltd.

George, E. I. and McCulloch, R. E. (1997), "Approaches for Bayesian variable selection," *Statistica Sinica*, 7, 339–373.

Hastie, T. and Tibshirani, R. (1993), "Varying-Coefficient Models," *Journal of the Royal Statistical Society. Series B (Methodological)*, 55, 757–796.

Kastner, G. and Frühwirth-Schnatter, S. (2014), "Ancillarity-sufficiency interweaving strategy (ASIS) for boosting MCMC estimation of stochastic volatility models," *Comput. Statist. Data Anal.*, 76, 408–423.

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

Kowal, D. R. (2020), *dsp: Dynamic Shrinkage Processes*, r package version 0.1.0 — For new features, see the 'Changelog' file (in the package source).

Kowal, D. R., Matteson, D. S., and Ruppert, D. (2019), "Dynamic shrinkage processes," *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 81, 781–804.

Lang, S. and Brezger, A. (2004), "Bayesian P-Splines," *Journal of Computational and Graphical Statistics*, 13, 183–212.

Marx, B. D. and Eilers, P. H. C. (1999), "Generalized Linear Regression on Sampled Signals and Curves: A P-Spline Approach," *Technometrics*, 41, 1–13.

Neal, R. M. (2011), "MCMC using Hamiltonian dynamics," in *Handbook of Markov chain Monte Carlo*, CRC Press, Boca Raton, FL, Chapman & Hall/CRC Handb. Mod. Stat. Methods, pp. 113–162.

Omori, Y., Chib, S., Shephard, N., and Nakajima, J. (2007), "Stochastic volatility with leverage: fast and efficient likelihood inference," *J. Econometrics*, 140, 425–449.

Pawitan, Y. and O'Sullivan, F. (1994), "Nonparametric spectral density estimation using penalized Whittle likelihood," *J. Amer. Statist. Assoc.*, 89, 600–610.

Piironen, J. and Vehtari, A. (2017), "Sparsity information and regularization in the horseshoe and other shrinkage priors," *Electron. J. Statist.*, 11, 5018–5051.

Polson, N. G., Scott, J. G., and Windle, J. (2013), "Bayesian inference for logistic models using Pólya-Gamma latent variables," *J. Amer. Statist. Assoc.*, 108, 1339–1349.

Rosen, O., Wood, S., and Stoffer, D. S. (2012), "AdaptSPEC: adaptive spectral estimation for nonstationary time series," *J. Amer. Statist. Assoc.*, 107, 1575–1589.

Shumway, R. H. and Stoffer, D. S. (2017), *Time series analysis and its applications*, Springer Texts in Statistics, Springer, Cham, 4th ed., with R examples.

Wahba, G. (1980), "Automatic Smoothing of the Log Periodogram," *Journal of the American Statistical Association*, 75, 122–132.

Whittle, P. (1962), "Gaussian estimation in stationary time series," *Bull. Inst. Internat. Statist.*, 39, 105–129.

Appendix A

Appendix

A.1 Derive the distribution of the μ

In chapter 4 we have $\mu = \log(A) = \log(\tau^2) \implies A = \exp(\mu)$, then the density function of μ is $P(\mu) = \frac{\gamma}{\pi}\frac{1}{\gamma^2+e^\mu} \cdot \frac{1}{e^{\frac{1}{2}\mu}} \cdot e^\mu$, we use $\frac{1}{\gamma^2} = e^{-2\ln(\gamma)}$ to rewrite the equation above and get $P(\mu) = \frac{1}{\pi\gamma}\frac{1}{1+e^{\mu-2\ln(\gamma)}} \cdot e^{\frac{1}{2}\mu}$. Then use $\frac{1}{\gamma} = e^{-\log(\gamma)}$ to rewrite the equation above and get

$$P(\mu) = \frac{1}{\pi}\frac{e^{\frac{1}{2}\mu-\log(\gamma)}}{1+e^{\mu-2\log(\gamma)}} = \frac{1}{\pi} \cdot \frac{e^{\frac{1}{2}(\mu-2\log(\gamma))}}{1+e^{\mu-2\log(\gamma)}}.$$

A.2 Posterior distribution for parameters in single-component nonparametric estimation

For $\boldsymbol{h}$

We define$\boldsymbol{h} = \begin{bmatrix} h_1 \\ h_2 \\ \vdots \\ h_M \end{bmatrix}$ The evolution equation contains h_i, which is $h_{i+1} = \mu + \phi(h_i - \mu) + \eta_i$, $\eta_i \overset{iid}{\sim} Z(\frac{1}{2}, \frac{1}{2}, 0, 1)$. We can rewrite it as

$$\eta_i = (h_{i+1} - \mu) - \phi(h_i - \mu)$$

since $\eta_i \sim Z(\frac{1}{2}, \frac{1}{2}, 0, 1)$, we can draw samples of $\eta_i \sim Z(\frac{1}{2}, \frac{1}{2}, 0, 1)$ by $\eta_i|\xi_i \overset{ind}{\sim} \mathrm{N}(0, \xi_i^{-1})$ for $\xi_i \overset{iid}{\sim} PG(1, 0)$.

So if we let $\tilde{\boldsymbol{h}} = \begin{bmatrix} h_1 - \mu \\ h_2 - \mu \\ \vdots \\ h_M - \mu \end{bmatrix}$, $\boldsymbol{D_\phi} = \begin{bmatrix} 1 & 0 & 0 & 0 & \cdots & 0 & 0 \\ -\phi & 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & -\phi & 1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & -\phi & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & -\phi & 1 \end{bmatrix}$. $\boldsymbol{\eta} = \begin{bmatrix} \eta_1 \\ \eta_2 \\ \vdots \\ \eta_M \end{bmatrix}$. Then we have $\boldsymbol{\eta} = \boldsymbol{D_\phi}\tilde{\boldsymbol{h}} \sim \mathrm{N}(0, \Sigma_\xi)$, here $\Sigma_\xi = diag\{(\xi_i^{-1})_{i=1}^M\}$.

Another part which contains $\boldsymbol{h}$ is $\log(\omega_i^2 + c) \sim \mathrm{N}(h_i + m_{s_i}, \nu_{s_i})$, denote $\log(\omega_i^2 + c)$ by $\tilde{y_i}$, then from $\tilde{y_i} \overset{ind}{\sim} \mathrm{N}(h_i + m_{s_i}, \nu_{s_i})$ we have $\tilde{\boldsymbol{y}} \sim \mathrm{N}(\boldsymbol{m} + \tilde{\boldsymbol{h}} + \tilde{\boldsymbol{\mu}}, \Sigma_\nu)$. Here $\tilde{\boldsymbol{y}} = \begin{bmatrix} \tilde{y_1} \\ \tilde{y_2} \\ \vdots \\ \tilde{y_M} \end{bmatrix}$, $\boldsymbol{m} = \begin{bmatrix} m_{s_1} \\ m_{s_2} \\ \vdots \\ m_{s_M} \end{bmatrix}$, $\tilde{\boldsymbol{\mu}} = \begin{bmatrix} \mu \\ \mu \\ \mu \\ \vdots \\ \mu \end{bmatrix}$, $\tilde{\boldsymbol{h}}$ is the same as we defined above, $\Sigma_\nu = diag\{(\nu_{s_i})_{i=1}^n\}$.

From $\boldsymbol{\eta} = \boldsymbol{D_\phi}\tilde{\boldsymbol{h}} \sim \mathrm{N}(0, \Sigma_\xi)$ we have

$$P(\boldsymbol{D_\phi}|\Sigma_\xi) = (2\pi)^{-\frac{M}{2}} det(\Sigma_\xi) \exp\{-\frac{1}{2}(\boldsymbol{D_\phi}\tilde{\boldsymbol{h}})^T \Sigma_\xi^{-1} (\boldsymbol{D_\phi}\tilde{\boldsymbol{h}})\},$$

if we denote ${\boldsymbol{D_\phi}}^T \Sigma_\xi^{-1} \boldsymbol{D_\phi}$ by Σ_3^{-1}, then we can rewrite it as

$$P(\boldsymbol{D_\phi}|\Sigma_\xi) = (2\pi)^{-\frac{M}{2}} det(\Sigma_\xi) \exp\{-\frac{1}{2}\tilde{\boldsymbol{h}}^T \Sigma_3^{-1} \tilde{\boldsymbol{h}}\}.$$

From $\tilde{\boldsymbol{y}} \sim \mathrm{N}(\boldsymbol{m} + \tilde{\boldsymbol{h}} + \tilde{\boldsymbol{\mu}}, \Sigma_\nu)$, we have

$$P(\tilde{\boldsymbol{y}}|\Sigma_\nu) = (2\pi)^{-\frac{M}{2}} det(\Sigma_\nu) \exp\{-\frac{1}{2}(\tilde{\boldsymbol{y}} - \boldsymbol{m} - \tilde{\boldsymbol{h}} - \tilde{\boldsymbol{\mu}})^T \Sigma_\nu^{-1} (\tilde{\boldsymbol{y}} - \boldsymbol{m} - \tilde{\boldsymbol{h}} - \tilde{\boldsymbol{\mu}})\},$$

we denote $\tilde{\boldsymbol{y}} - \boldsymbol{m} - \tilde{\boldsymbol{\mu}}$ by $\tilde{\boldsymbol{k}}$, then we can rewrite it as

$$P(\tilde{\boldsymbol{y}}|\Sigma_\nu) = (2\pi)^{-\frac{M}{2}} det(\Sigma_\nu) \exp\{-\frac{1}{2}(-\tilde{\boldsymbol{h}} + \tilde{\boldsymbol{k}})^T \Sigma_\nu^{-1}(-\tilde{\boldsymbol{h}} + \tilde{\boldsymbol{k}})\}.$$

So

$$(2\pi)^{-\frac{M}{2}} det(\Sigma_\xi) \exp\{-\frac{1}{2}(\boldsymbol{D_\phi}\tilde{\boldsymbol{h}})^T \Sigma_\xi^{-1}(\boldsymbol{D_\phi}\tilde{\boldsymbol{h}})\}$$

$$\times (2\pi)^{-\frac{M}{2}} det(\Sigma_\nu) \exp\{-\frac{1}{2}(\tilde{\boldsymbol{y}} - \boldsymbol{m} - \tilde{\boldsymbol{h}} - \tilde{\boldsymbol{\mu}})^T \Sigma_\nu^{-1}(\tilde{\boldsymbol{y}} - \boldsymbol{m} - \tilde{\boldsymbol{h}} - \tilde{\boldsymbol{\mu}})\}$$

$$\propto \exp\{-\frac{1}{2}\tilde{\boldsymbol{h}}^T \Sigma_3^{-1}\tilde{\boldsymbol{h}}\} \times \exp\{-\frac{1}{2}(-\tilde{\boldsymbol{h}} + \tilde{\boldsymbol{k}})^T \Sigma_\nu^{-1}(-\tilde{\boldsymbol{h}} + \tilde{\boldsymbol{k}})\}$$

$$\propto \exp\{-\frac{1}{2}[\tilde{\boldsymbol{h}}^T \Sigma_3^{-1}\tilde{\boldsymbol{h}} + \tilde{\boldsymbol{h}}^T \Sigma_\nu^{-1}\tilde{\boldsymbol{h}} - \tilde{\boldsymbol{h}}^T \Sigma_\nu^{-1}\tilde{\boldsymbol{k}} - \tilde{\boldsymbol{k}}^T \Sigma_\nu^{-1}\tilde{\boldsymbol{h}}]\}$$

$$= \exp\{-\frac{1}{2}[\tilde{\boldsymbol{h}}^T (\Sigma_3^{-1} + \Sigma_\nu^{-1})\tilde{\boldsymbol{h}} - \tilde{\boldsymbol{h}}^T \Sigma_\nu^{-1}\tilde{\boldsymbol{k}} - \tilde{\boldsymbol{k}}^T \Sigma_\nu^{-1}\tilde{\boldsymbol{h}}]\}$$

$$\propto \exp\{-\frac{1}{2}[(\tilde{\boldsymbol{h}} - (\Sigma_3^{-1} + \Sigma_\nu^{-1})^{-1}\Sigma_\nu^{-1}\tilde{\boldsymbol{k}})^T (\Sigma_3^{-1} + \Sigma_\nu^{-1})(\tilde{\boldsymbol{h}} - (\Sigma_3^{-1} + \Sigma_\nu^{-1})\Sigma_\nu^{-1}\tilde{\boldsymbol{k}})]\}$$

so we can see that the conditional distribution of $\tilde{\boldsymbol{h}}$ is $\tilde{\boldsymbol{h}} \sim \mathrm{N}((\Sigma_3^{-1} + \Sigma_\nu^{-1})^{-1}\Sigma_\nu^{-1}\tilde{\boldsymbol{k}}, (\Sigma_3^{-1} + \Sigma_\nu^{-1})^{-1}) = \mathrm{N}((\boldsymbol{D_\phi}^T \Sigma_\xi^{-1} \boldsymbol{D_\phi} + \Sigma_\nu^{-1})^{-1}\Sigma_\nu^{-1}(\tilde{\boldsymbol{y}} - \boldsymbol{m} - \tilde{\boldsymbol{\mu}}), (\boldsymbol{D_\phi}^T \Sigma_\xi^{-1} \boldsymbol{D_\phi} + \Sigma_\nu^{-1})^{-1})$. If we let $Q_{\tilde{h}} = \boldsymbol{D_\phi}^T \Sigma_\xi^{-1} \boldsymbol{D_\phi} + \Sigma_\nu^{-1}$ and $l_{\tilde{h}} = \Sigma_\nu^{-1}(\tilde{\boldsymbol{y}} - \boldsymbol{m} - \tilde{\boldsymbol{\mu}})$, then we can rewrite the conditional distribution of $\tilde{\boldsymbol{h}}$ as $\tilde{h} \sim \mathrm{N}(Q_{\tilde{h}}^{-1} l_{\tilde{h}}, Q_{\tilde{h}}^{-1})$. We can sample $\tilde{\boldsymbol{h}}$ in this way:

For μ

From $\mu|\sigma_\epsilon, \xi_\mu^{-1} \sim \mathrm{N}(\log(\frac{\sigma_\epsilon^2}{M}), \xi_\mu^{-1})$, we have

$$P(\mu|\sigma_\epsilon, \xi_\mu^{-1}) = \frac{1}{\sqrt{2\pi\xi_\mu^{-1}}} \exp\{-\frac{1}{2}\frac{(\mu - \log(\frac{\sigma_\epsilon^2}{M}))^2}{\xi_\mu^{-1}}\} \propto \exp\{-\frac{1}{2}\frac{(\mu - \log(\frac{\sigma_\epsilon^2}{M}))^2}{\xi_\mu^{-1}}\}.$$

From the evolution equation $h_{i+1} = \mu + \phi(h_i - \mu) + \eta_i$, we have $\mu + \eta_0 = h_1 \sim \mathrm{N}(\mu, \xi_0^{-1}), \xi_0 \sim PG(1, 0)$ and

$$P(h_1|\mu,\xi_0) = \frac{1}{\sqrt{2\pi\xi_0^{-1}}}\exp\{-\frac{(h_1-\mu)^2}{2\xi_0^{-1}}\} \propto \exp\{-\frac{(h_1-\mu)^2}{2\xi_0^{-1}}\}.$$

For $i = 1, 2, \ldots, M-1$. $h_{i+1} = \mu + \phi(h_t - \mu) + \eta_t$, $\eta_i \overset{iid}{\sim} Z(\alpha, \beta, 0, 1)$ implies that $h_{i+1} = \phi h_i + (1-\phi)\mu + \eta_i$, so $h_{i+1} \sim \mathrm{N}(\phi h_i + (1-\phi)\mu, \xi_i^{-1}), \xi_t \sim PG(1,0)$. Then we can see that

$$p(h_{i+1}|\mu,\phi,\xi_t) = \frac{1}{\sqrt{2\pi\xi_i^{-1}}}\exp\{-\frac{(h_{i+1}-\phi h_i-(1-\phi)\mu)^2}{2\xi_i^{-1}}\}$$

$$\propto \exp\{-\frac{(h_{i+1}-\phi h_i-(1-\phi)\mu)^2}{2\xi_i^{-1}}\} \quad \text{for} i = 1, 2, \ldots, n-1$$

so the posterior distribution of μ is

$$\exp\{-\frac{1}{2}\frac{(\mu-\log(\frac{\sigma_\epsilon^2}{M}))^2}{\xi_\mu^{-1}}\} \times \exp\{-\frac{(h_1-\mu)^2}{2\xi_0^{-1}}\} \times \prod_{i=1}^{M-1}\exp\{-\frac{(h_{i+1}-\phi h_i-(1-\phi)\mu)^2}{2\xi_i^{-1}}\}$$

$$\propto \exp\{-\frac{1}{2}[(\xi_\mu+\xi_0+(1-\phi)^2\sum_{i=1}^{M-1}\xi_i)\mu^2 - 2(\log(\frac{\sigma_\epsilon^2}{M})\xi_\mu + h_1\xi_0 + (1-\phi)^2\sum_{i=1}^{M-1}(h_{i+1}-\phi h_i)\xi_i)]\}$$

Implies that μ has a normal distritbution with mean $\xi_\mu+\xi_0+(1-\phi)^2\sum_{i=1}^{M-1}\xi_i)^{-1}(\log(\frac{\sigma_\epsilon^2}{M})\xi_\mu + h_1\xi_0 + (1-\phi)^2\sum_{i=1}^{M-1}(h_{i+1}-\phi h_i)\xi_i$ and variance $\xi_\mu + \xi_0 + (1-\phi)^2\sum_{i=1}^{M-1}\xi_i)^{-1}$

For ϕ

The prior of ϕ is $\frac{1+\phi}{2} \sim Beta(10, 2)$, it means

$$p(\frac{1+\phi}{2}) = \frac{(\frac{1+\phi}{2})^9(1-\frac{1+\phi}{2})^1}{B(10,2)} = \frac{(\frac{1+\phi}{2})^9(1-\frac{1+\phi}{2})^1}{\frac{1}{110}} = 110(\frac{1+\phi}{2})^9(\frac{1-\phi}{2}),$$

then $\frac{d}{d\phi}(\frac{1+\phi}{2}) = \frac{1}{2}$, so we have $p(\phi) = \frac{110}{2}(\frac{1+\phi}{2})^9(\frac{1-\phi}{2}) \propto (1+\phi)^9(1-\phi)$.

The evolution equation contains ϕ, which is $h_{i+1} = \mu + \phi(h_i - \mu) + \eta_i$. We have $\mu + \eta_0 = h_1 \sim \mathrm{N}(\mu, \xi_0^{-1}), \xi_0 \sim PG(1,0)$, which implies

$$P(h_1|\mu,\xi_0) = \frac{1}{\sqrt{2\pi\xi_0^{-1}}}\exp\{-\frac{(h_1-\mu)^2}{2\xi_0^{-1}}\} \propto \exp\{-\frac{(h_1-\mu)^2}{2\xi_0^{-1}}\},$$

and for $i = 1, 2, \ldots, M-1$,

$$p(h_{i+1}|\mu,\phi,\xi_i) = \frac{1}{\sqrt{2\pi\xi_i^{-1}}}\exp\{-\frac{(h_{i+1}-\phi h_i-(1-\phi)\mu)^2}{2\xi_i^{-1}}\},$$

so we have

$$\prod_{i=1}^{M-1}\frac{1}{\sqrt{2\pi\xi_i^{-1}}}\exp\{-\frac{(h_{i+1}-\phi h_i-(1-\phi)\mu)^2}{2\xi_i^{-1}}\}$$

$$\propto \exp\{-\frac{1}{2}\sum_{i=1}^{M-1}[(h_{i+1}-\mu)-\phi(h_i-\mu)]^2\xi_i\}$$

$$\propto \exp\{-\frac{1}{2}[\phi^2\sum_{i=1}^{M-1}(h_i-\mu)^2\xi_i - 2\phi\sum_{i=1}^{M-1}(h_{i+1}-\mu)(h_i-\mu)\xi_i]\}$$

then

$$p(\phi)\times\prod_{i=1}^{M-1}p(h_{i+1}|\mu,\phi,\xi_i)$$

$$\propto (1+\phi)^9(1-\phi)\times\exp\{-\frac{1}{2}[\phi^2\sum_{i=1}^{M-1}(h_i-\mu)^2\xi_i - 2\phi\sum_{i=1}^{M-1}(h_{i+1}-\mu)(h_i-\mu)\xi_i]\}.$$

The second term in the formula above is a kernel of normal distribution, which is

$$\mathrm{N}(\frac{\sum_{i=1}^{M-1}(h_{i+1}-\mu)(h_i-\mu)\xi_i}{\sum_{i=1}^{M-1}(h_i-\mu)^2\xi_i}, \frac{1}{\sum_{i=1}^{M-1}(h_i-\mu)^2\xi_i}).$$

For mixture component indicators $\{s_i\}$

In Omori et al. (2007) , the $\mathrm{P}(s_i = k)$, m_k and ν_k for $k = 1, 2, 3, \ldots, 10$ are show in the table below.

From $\log(\omega_i^2 + c) = \tilde{y}_i \sim \mathrm{N}(h_i + m_{s_i}, \nu_{s_i})$, we have

Table A.1: Table of the 10-component Gaussian mixture

k	$\mathrm{P}(s_i = k)$	m_k	ν_k
1	0.00609	1.92677	0.11265
2	0.04775	1.34744	0.17788
3	0.13057	0.73504	0.26768
4	0.20674	0.02266	0.40611
5	0.22715	-0.85173	0.62699
6	0.18842	-1.97278	0.98583
7	0.12047	-3.46788	1.57469
8	0.05591	-5.55246	2.54498
9	0.01575	-8.68384	4.16591
10	0.00115	-14.65000	7.33342

$$P(\tilde{y}_i) = \frac{1}{\sqrt{2\pi\nu_{s_i}}} \exp\{-\frac{(\tilde{y}_i - h_i - m_{s_i})^2}{2\nu_{s_i}}\} \propto \frac{1}{\sqrt{\nu_{s_i}}} \exp\{-\frac{(\tilde{y}_i - h_i - m_{s_i})^2}{2\nu_{s_i}}\},$$

so the conditional distribution of s_i is

$$P(s_i = k|\cdot) = P(s_i = k) * \frac{1}{\sqrt{\nu_k}} \exp\{-\frac{(\tilde{y}_i - h_i - m_k)^2}{2\nu_k}\}.$$

For ξ_i

Since $\eta_i|xi_i \sim \mathrm{N}(0, \xi_{-1})$ and $\xi_i \sim PG(1, 0)$, then based on theorem 1 of **?**, we have the conditional distribution of $\xi_i|\eta_i \sim PG(1, \eta_i)$.

For σ_ϵ

The prior we place on σ_ϵ is $\sigma_\epsilon \propto \frac{1}{\sigma_\epsilon}$. Then from $y_i = \beta_i + \epsilon_i, \epsilon_i \sim \mathrm{N}(0, \sigma_\epsilon^2)$ for $i = 1, 2, \ldots, M$ and $\tau \sim C^+(0, \frac{\epsilon}{\sqrt{M}})$, we have the conditional distribution of σ_ϵ is

$$P(\sigma_\epsilon|\cdot) = \frac{1}{\sigma_\epsilon} \cdot \frac{1}{\sigma_\epsilon^M} \cdot \exp\{-\frac{1}{2\sigma_\epsilon^2}\sum_{i=1}^{M}(y_i - \beta_i)^2\} \cdot \frac{2}{\pi}\frac{\frac{\sqrt{M}}{\sigma_\epsilon}}{1 + (\frac{\tau}{\frac{\sigma_\epsilon}{\sqrt{M}}})^2}$$

$$\propto \sigma_\epsilon^{-(M+1)} \cdot \exp\{-\frac{1}{2\sigma_\epsilon^2}\sum_{i=1}^{M}(y_i - \beta_i)^2\} \cdot \frac{\sqrt{M}}{\sigma_\epsilon(1 + \frac{\tau^2 M}{\sigma_\epsilon^2})}$$

$$= \sigma_\epsilon^{-(M+2)} \cdot \exp\{-\frac{1}{2\sigma_\epsilon^2}\sum_{i=1}^{M}(y_i - \beta_i)^2\} \cdot \frac{\sqrt{M}}{1 + \frac{\tau^2 M}{\sigma_\epsilon^2}}$$

Curriculum Vitae

Yi Xie was born on July 18, 1988, as the first and only son of Heping Xie and Qi Gao. In 2006, he went to the Central South University in China and four years later received a bachelor degree of science (Mathematics and Applied Mathematics). After that he furthered his study at Changsha University of Science and Technology and received a masters degree of science (stochastic processes, Markov processes) in 2013. Then he went back to his hometown, a small city in Hunan province, and became a math teacher at a university. In 2016, he decided to resign his job and went to The University of Texas at El Paso. While pursuing a masters degree in Statistics, he worked as a Teaching and Research Assistant. Two years later in the summer of 2018, he started his Ph.D progam in the computational sicence at UTEP.

Contact Information: yxie3@miners.utep.edu