Sobre o uso de misturas de distribuições gaussianas através da escala em modelos semiparamétricos de regressão e séries temporais

Marcelo Magalhães Taddeo

TESE APRESENTADA
AO
INSTITUTO DE MATEMÁTICA E ESTATÍSTICA
DA
UNIVERSIDADE DE SÃO PAULO
PARA
OBTENÇÃO DO TÍTULO
DE
DOUTOR EM CIÊNCIAS

Área de Concentração: Estatística
Orientador: Prof. Dr. Pedro Alberto Morettin

Durante o desenvolvimento deste trabalho o autor recebeu auxílio financeiro do CNPq

São Paulo, julho de 2008

Sobre o uso de misturas de distribuições gaussianas através da escala em modelos semiparamétricos de regressão e séries temporais

Este exemplar corresponde à redação
final da tese devidamente corrigida
e defendida por Marcelo Magalhães Taddeo
e aprovada pela Comissão Julgadora.

Banca Examinadora:

- Prof. Dr. Pedro Alberto Morettin (orientador) - IME-USP.
- Prof. Dr. José Carlos Simon de Miranda - IME-USP.
- Prof. Dr. Pedro Luiz Valls Pereira - FGV/SP.
- Prof. Dr. Luiz Koodi Hotta - IMECC-UNICAMP.
- Profa. Dra. Beatriz Vaz de Melo Mendes - UFRJ

Agradecimentos

Em primeiro lugar, gostaria muito de agradecer ao Prof. Dr. Pedro Alberto Morettin pela orientação e pela liberdade de escolhas e de trabalho, os quais me fizeram crescer pessoalmente e me desenvolver como pesquisador. Posso dizer tranqüilamente que, ao longo deste período no qual trabalhamos juntos, aprendi verdadeiramente a admirar sua pessoa e seu trabalho, motivos pelos quais me orgulho muito desta orientação.

Embora não tenham relação direta com a tese, gostaria de agradecer, pelo apoio e companhia ao longo destes anos, minha família, especialmente minha mãe e meu pai (sem os quais, definitivamente nunca teria chegado até aqui), à Patricia Lima, minha namorada e grande companheira (e assessora) ao longo destes e outros anos e ao meu amigo Claudinei de Paula.

Gostaria também de agradecer aos membros da banca examinadora, Prof. Dr. José Carlos Alberto Simon, Prof. Dr. Pedro Luiz Valls Pereira, Prof. Dr. Luiz Koodi Hotta e Profa. Dra. Beatriz Vaz de Melo Mendes pelos comentários, elogios e críticas, todos muito benvindos e proveitosos.

Finalmente, gostaria de agradecer ao apoio e auxílio financeiro do CNPq.

Resumo

Este trabalho trata da estimação de modelos semiparamétricos quando assumimos que o ruído segue uma mistura de distribuições gaussianas na escala. Distribuições desta natureza são interessantes, pois, através dela podemos representar qualquer distribuição simétrica contínua como, por exemplo, distribuições com caudas pesadas, tais como a t de Student, as da família de distribuições estáveis e a exponencial dupla. Estas distribuições são especialmente úteis em problemas com altas taxas de valores extremos e *outliers*. Combinando a natureza de tais distribuições com o algoritmo EM, mostramos que, se usarmos bases de funções convenientemente escolhidas, tais como B-splines ou ondaletas, obtemos uma ferramenta poderosa para a estimação de curvas, tanto em problemas de regressão quanto de séries temporais, mesmo sob a presença de valores extremos e *outliers* e sem a necessidade de uma especificação a priori da forma funcional da mesma. Mais ainda, podemos associar outros métodos robustos a algumas destas distribuições e assim replicá-los para a estimação de curvas. Assim sendo, pudemos unificar diversos métodos de estimação em um único procedimento, no qual devemos apenas alterar a ponderação dos ruídos para passarmos de um método a outro. Aplicamos esta técnicas a modelos univariados onde a função média, ou sinal, era desconhecida, a modelos parcialmente lineares e a modelos mistos. Além disso, (i) demonstramos que quando aplicamos este método a séries temporais lineares autorregressivas, os estimadores obtidos são consistentes, e (ii) mostramos como combinar bases de funções a misturas de distribuições gaussianas dentro de um enfoque bayesiano para estimação de curvas, tanto em modelos univariados quanto em modelos parcialmente lineares.

Abstract

This work is about the estimation of semiparametric models when assuming that the noise follows a Gaussian scale mixture. Such distributions are particularly interesting since any continuous symmetric distribution may be represented as a Gaussian scale mixture. Examples are the Student t, double exponential and stable distributions. They are particularly useful in the presence of extreme values and outliers since these are heavy tailed distributions. Together with the EM algorithm, we show that, by using suitably chosen function basis, such as B-splines or wavelets, we get a powerful tool for curve estimation in regression and time series problems, both under the presence of extreme values and without the need of any parametric assumption about the functional form of the target function. One can relate it to other robust estimation methods and switch between them by just adjusting the residuals weighting and hence unify several methodologies in one computational technique. Such techniques were applied to univariate models with unknown mean function, to partially linear models and mixed models. Moreover, (i) we show that, when applied to linear autorregressive time series, this method results in consistent estimates, and (ii) we show how to combine functions basis to Gaussian scale mixtures inside a bayesian framework in order to get curve estimates to univariate regression models with unknown mean function and partially linear models.

Sumário

Sumário

Introdução

Este trabalho é sobre a estimação de curvas (função média) em modelos de regressão e de séries temporais na presença de uma classe específica de ruído distribuido de acordo com uma mistura na escala de distribuições gaussianas. Tal classe tem as propriedades de englobar todas as distribuições contínuas simétricas e "robustificar" o modelo (dado que muitas destas distribuições têm caudas pesadas). Dado que a estrutura funcional de tais curvas não é conhecida a priori, necessitamos de meios para aproximá-las. Embora muito do que se dirá mais adiante seja aplicável a diversos meios de aproximação, tais como ondaletas ou séries de Fourier, daremos especial atenção ao uso de *splines*. Logo, a principal contribuição desta tese está em uma série de algoritmos para estimar curvas em modelos semiparamétricos de regressão e séries temporais (incluindo modelos parcialmente lineares e modelos mistos), sem, portanto, a necessidade de uma estruturação funcional a priori da função alvo. Além disso, demonstramos consistência para a classe de algoritmos considerada quando restritos a séries temporais lineares autorregressivas e como adaptar tais algoritmos dentro de um enfoque bayesiano, especialmente no caso de modelos semiparamétricos univariados e parcialmente lineares.

Os dois primeiros capítulos são uma introdução às misturas de normais através da escala e das principais ferramentas utilizadas neste trabalho, como splines e o algoritmo EM. No capítulo 3 falamos sobre o uso destas distribuições em modelos lineares de regressão e de séries temporais e apresentamos resultados de consistência no caso de séries temporais e estendemos a teoria no caso de regressão para modelos penalizados. No capítulo 4, desenvolvemos o algoritmo sugerido para modelos de regressão semiparamétrico e como calcular intervalos e bandas de confiança usando *bootstrap* e estendemos a teoria para modelos parcialmente lineares. O capítulo 5 é equivalente ao anterior, porém trata de dados de séries temporais. O capítulo 6 concentra-se na análise bayesiana destes modelos e utiliza basicamente o algoritmo de Gibbs para inferência dos mesmos. Portanto, apresentamos neste capítulo a derivação deste algoritmo aplicado a este modelo e, embora utilizemos principalmente splines, mostramos como poderíamos aplicar mistura de normais através da escala e ondaletas sob a perspectiva bayesiana. Em todos estes capítulos (4,5 e 6) realizamos diversos estudos de simulação para demonstrar a efetividade da técnica. O capítulo seguinte é uma breve nota sobre a utilização de modelos mistos dado que o ruído e/ou a componente aleatória pertence à classe de distribuições considerada aqui. Finalmente, concluimos com uma breve aplicação utilizando dados reais.

Capítulo 1

Mistura na Escala de Distribuições Gaussianas

1.1 Introdução

Dizemos que a variável aleatória X segue uma distribuição definida pela *mistura de gaussianas através do parâmetro de escala* se ela puder ser escrita na forma

$$X = Z/\sqrt{\sigma}$$

onde $Z \sim \mathcal{N}(0,1)$ e σ é uma variável aleatória qualquer (contínua ou discreta) assumindo valores em $(0,\infty)$. No caso em que σ segue uma distribuição contínua, sua função densidade de probabilidade h, a qual pode ser expressa analiticamente ou não, é chamada de *densidade de mistura*. Distribuições deste tipo são de interesse prático, pois, englobam quaisquer distribuições contínuas, unimodais e simétricas, veja [36]. Além disso, sabe-se que a forma geral das funções densidades de tais distribuições é dada por

$$p(x) = \int_0^\infty \sigma^{1/2}\phi(\sigma^{1/2}x)h(\sigma)d\sigma \tag{1.1}$$

para $x \in \mathbb{R}$. Ou, de modo mais geral,

$$p(x) = \int_0^\infty \sigma^{1/2}\phi(\sigma^{1/2}x)dH(\sigma). \tag{1.2}$$

Em alguns casos, pode ser mais conveniente considerar a seguinte generalização de (1.1)

$$p(x) = \int_0^\infty \frac{1}{\delta\psi(\sigma)}\phi\left(\frac{x-c}{\delta\psi(\sigma)}\right)dH(\sigma)$$

onde $-\infty < x < \infty$, ψ é uma função positiva e definida no intervalo $(0,\infty)$, c é uma constante e ϕ é a densidade associada à distribuição normal padrão. De fato, essa notação permite que incorporemos parâmetros associados exclusivamente a h em ψ de modo a tornar o processo de estimação dos parâmetros de interesse

mais simples. As constantes c e δ são, respectivamente, os parâmetros de localização e de escala do modelo. Note que para (1.1), $\psi(\sigma) = \sigma^{-1/2}$. É fácil ver (lema 1.1.2) que p é a densidade da variável aleatória $X = \psi(S)Z$, onde $Z \sim \mathcal{N}(c, \delta^2)$ e $S > 0$, com função distribuição H, são variáveis aleatórias independentes. Em particular, se a distribuição H admitir uma densidade h, então, p pode ser escrita na forma

$$p(x) = \int_0^\infty \frac{1}{\delta\psi(\sigma)}\phi\left(\frac{x-c}{\delta\psi(\sigma)}\right)h(\sigma)d\sigma. \tag{1.3}$$

Indicaremos que a densidade de X é dada por (1.3) através da notação $X \sim SM_h(c, \delta; \psi)$, onde δ é o vetor de parâmetros associados a h. No caso particular em que $\psi(\sigma) = \sigma^{-1/2}$, denotaremos simplesmente por $X \sim SM_h(c, \delta)$.

O lema 1.1.1 generaliza um resultado em [5] que mostra como determinar a densidade h em (1.3) quando $c = 0$.

Lema 1.1.1. *Suponha que a função ψ seja invertível e defina*

$$\zeta(s) = \frac{1}{\sqrt{\pi}s\psi'\left[\psi^{-1}\left(\frac{1}{\sqrt{2s}}\right)\right]}h\left(\psi^{-1}\left(\frac{1}{\sqrt{2s}}\right)\right).$$

Então, se

- $\lim_{\sigma\to 0}\psi(\sigma) = 0$;
- $\lim_{\sigma\to\infty}\psi(\sigma) = \infty$;

a função $g(x) \equiv p(\sqrt{x})$ é a transformada de Laplace de ζ. Por outro lado, se

- $\lim_{\sigma\to 0}\psi(\sigma) = \infty$;
- $\lim_{\sigma\to\infty}\psi(\sigma) = 0$;

então, $g(x)$ é transformada de Laplace de $-\zeta$.

Demonstração. Segue imediatamente da mudança de variáveis $s = \frac{1}{2\psi(\sigma)^2}$ na expressão (1.3). □

Consequentemente, para determinar h basta aplicar a transformada inversa de Laplace em g. No caso particular em que $\psi(\sigma) = \sigma^{-1}$, $g(x)$ é a transformada de Laplace de

$$\zeta(s) = \frac{2}{\sqrt{\pi}}h(\sqrt{2s}).$$

O lema abaixo indica um modo de se obter variáveis aleatórias com densidades da forma (1.3). É interessante também notar que o lema abaixo permite estender a teoria para outras misturas na escala além da distribuição gaussiana.

Lema 1.1.2. *Se $X = \psi(S)Z$, onde S e Z são variáveis aleatórias independentes com $S > 0$, $S \sim h$ e ψ uma função positiva, então a densidade de X pode ser escrita na forma*

$$p(x) = \int_0^\infty \frac{1}{\delta\psi(\sigma)} f\left(\frac{x}{\delta\psi(\sigma)}\right) h(\sigma)d\sigma, \tag{1.4}$$

onde a densidade de Z é tal que

$$\varphi(z) = \frac{1}{\delta} f\left(\frac{z}{\delta}\right).$$

Por outro lado, se a densidade de X é dada por (1.4) e se h é uma função densidade de probabilidade, então, X é igual, em distribuição, a $\psi(S)Z$, onde $S \sim h$ e $Z \sim f$ são independentes.

Demonstração. Seja $x \in \mathbb{R}$, então,

$$\begin{aligned} P[X < x] &= P[\psi(S)Z < x] \\ &= \int_0^\infty \int_{-\infty}^{x/\psi(\sigma)} p(z,\sigma)dzd\sigma \\ &= \int_0^\infty \int_{-\infty}^{x/\psi(\sigma)} \varphi(z)h(\sigma)dzd\sigma \\ &= \int_0^\infty \int_{-\infty}^{x/\psi(\sigma)} \frac{1}{\delta} f\left(\frac{z}{\delta}\right) h(s)dzd\sigma. \end{aligned}$$

Logo, fazendo a mudança de variáveis $z' = z\psi(\sigma)$, temos que

$$P[X < x] = \int_0^\infty \int_{-\infty}^{x} \frac{1}{\delta\psi(\sigma)} f\left(\frac{z}{\delta\psi(\sigma)}\right) h(\sigma)dzd\sigma.$$

□

Em particular, se $Z \sim \mathcal{N}(0, \delta^2)$, e se $X = \psi(S)Z$, então,

$$p(x) = \int_0^\infty \frac{1}{\delta\psi(\sigma)} \phi\left(\frac{x}{\delta\psi(\sigma)}\right) h(\sigma)d\sigma,$$

onde ϕ é a densidade da normal padrão.

Para modelar e simular variáveis seguindo uma mistura na escala, ie, com função densidade de probabilidade dada por (1.3), podemos usar o fato de que, se

$$\begin{aligned} X|\sigma &\sim \mathcal{N}\left(0, \delta^2\psi(\sigma)^2\right), \\ \sigma &\sim h. \end{aligned} \tag{1.5}$$

então, a densidade marginal de X é exatamente aquela dada por (1.3). No caso particular em que $\psi(\sigma) = \sigma^{-1/2}$, ao invés da variância, a variável aleatória σ representa a precisão de $X|\sigma$.

Abaixo listamos alguns exemplos de distribuições que podem ser representadas como misturas na escala de gaussianas.

Exemplo 1.1.1 (Distribuição t de Student). *A distribuição t com ν graus de liberdade e parâmetro de escala δ, cuja densidade é dada por*

$$p(x|\nu,\delta) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\Gamma\left(\frac{1}{2}\right)}\frac{1}{\sqrt{\nu\delta^2}}\left[1+\frac{1}{\nu}\left(\frac{x}{\delta}\right)^2\right]^{-\frac{n+1}{2}},$$

pode ser representada na forma (1.3)*, pois, se $X \sim t_\nu(0,\delta)$, então, vale a igualdade (em distribuição) $X = V^{-1/2}Z$, onde $Z \sim \mathcal{N}(0,\delta^2)$ e $V \sim \Gamma\left(\frac{\nu}{2},\frac{\nu}{2}\right)$. Ou seja, a densidade de X pode ser representada por* (1.3)*, onde h é densidade associada à distribuição gama, $\Gamma\left(\frac{\nu}{2},\frac{\nu}{2}\right)$. Neste caso, $\psi(\sigma) = \frac{1}{\sqrt{\sigma}}$.*

Um caso particular de interesse, por ser uma distribuição com variância infinita, é a distribuição de Cauchy, a qual obtemos tomando $\nu = 1$. A distribuição de Cauchy também é um caso particular das distribuições estáveis descritas no próximo exemplo. Neste caso, h é a densidade associada a $\Gamma\left(\frac{1}{2},\frac{1}{2}\right)$. ◊

Exemplo 1.1.2 (Distribuições estáveis). *No caso das distribuições estáveis com índice de estabilidade satisfazendo $1 < \alpha < 2$, a representação* (1.3) *vale tomando $h(\sigma) = g(\sigma^{-2})$, onde g é a densidade de uma variável aleatória estável com índice de estabilidade $\alpha/2$. Este fato é consequência direta do teorema 1.1.3, cuja demonstração pode ser vista em [32].*

Teorema 1.1.3. *Sejam $Z \sim S_{\alpha'}(\sigma,0,0)$ com $0 < \alpha < \alpha' \leq 2$ e A uma variável aleatória estável com índice de estabilidade α/α' totalmente assimétrica a direita (ie, com $\beta = 1$) e com transformada de Laplace*

$$Ee^{-\gamma A} = e^{-\gamma^{\alpha/\alpha'}}, \quad \gamma > 0,$$

isto é, $A \sim S_{\alpha/\alpha'}\left((\cos\frac{\pi\alpha}{2\alpha'})^{\alpha'/\alpha},1,0\right)$. Assuma também que Z e A são independentes. Então,

$$X \equiv A^{1/\alpha'}Z \sim S_\alpha(\sigma,0,0).$$

De fato, dado $X \sim S_\alpha(\lambda,0,0)$ e tomando $\alpha' = 2$ e $\sigma = \lambda$, existem variáveis aleatórias $\tilde{Z} \sim \mathcal{N}(0,2\lambda^2)$ (ie, $\tilde{Z} \sim S_2(\lambda,0,0)$) e $A \sim S_{\alpha/2}\left((\cos\frac{\pi\alpha}{4})^{\alpha'/2},1,0\right)$ positiva tais que

$$X = A^{1/2}\tilde{Z}.$$

Ora, então pelo lema 1.1.2, a densidade de X pode ser escrita na forma

$$p(x) = \int_0^\infty \sigma^{-1/2}\phi(\sigma^{-1/2}x)h(\sigma)d\sigma$$

onde h é a densidade de A.

Convém lembrar que, em geral, a forma funcional das densidades associadas às distribuições estáveis não têm uma representação analítica fechada. No entanto, para gerar amostras pseudo-aleatórias a partir de uma distribuição estável $X \sim S_\alpha(1, \beta, 0)$*, podemos usar o método de Chambers-Mallows-Stuck (CMS), [10], de acordo com o qual, se*

$$V \sim U\left(-\frac{\pi}{2}, \frac{\pi}{2}\right) \text{ e } W \sim Exp(1),$$

então,

$$X = S_{\alpha,\beta} \times \frac{\operatorname{sen}(\alpha(V + B_{\alpha,\beta}))}{\cos(V)^{1/\alpha}} \times \frac{cos(V - \alpha(V + B_{\alpha,\beta}))^{(1-\alpha)/\alpha}}{W}$$

onde

$$B_{\alpha,\beta} = \frac{\operatorname{arctg}\left(\beta \operatorname{tg} \frac{\alpha\pi}{2}\right)}{\alpha} \text{ e } S_{\alpha,\beta} = \left(1 + \gamma^2 \operatorname{tg}^2 \frac{\alpha\pi}{2}\right)^{1/(2\alpha)},$$

com γ *definido de tal modo que*

$$\frac{\pi\gamma}{2} = \arg\left(1 - i\beta \operatorname{tg} \frac{\alpha\pi}{2}\right),$$

segue a distribuição desejada. ◊

Exemplo 1.1.3 (Distribuição exponencial dupla). *As distribuições da família potência-exponencial*[1] *são aquelas cujas funções densidades podem ser escritas na forma*

$$p(x) = k \exp(-|x|^b), \quad -\infty < x < \infty, \quad 1 \leq b \leq 2.$$

Elas têm como caso particular, para $b = 1$*, a distribuição de Laplace (ou exponencial dupla) que nos é de interesse, pois, utilizá-la equivale, como veremos, a estimar os parâmetros dos modelos estudados através da minimização da norma* L^1*. A densidade de mistura é, então, dada por*

$$h(\sigma) = \frac{1}{2}\sigma^2 \exp\left(-\frac{1}{2\sigma}\right).$$

◊

Exemplo 1.1.4 (Distribuição logística). *A densidade da distribuição logística é dada por*

$$\frac{\exp(-x)}{[1 + \exp(-x)]^2}$$

e, neste caso, a densidade de mistura é igual a

$$h(\sigma) = \sum_{k=1}^{\infty} (-1)^{k-1} k \exp(-k\sigma^{1/2}).$$

◊

[1] *Exponential Power Family*

Exemplo 1.1.5 (Distribuição normal contaminada). *A distribuição normal contaminada é definida por*

$$p(x) = (1-\xi)\mathcal{N}(0,1) + \xi\mathcal{N}\left(0, \frac{1}{\lambda^2}\right)$$

e a densidade de mistura é dada por

$$h(\sigma) = \begin{cases} 1-\xi, & se\ \sigma = 1, \\ \xi, & se\ \sigma = \lambda^2, \\ 0, & caso\ contrário \end{cases}.$$

Note que, diferentemente dos demais exemplos, a distribuição de mistura é agora discreta. Como veremos mais adiante, a distribuição acima pode ser generalizada de modo a aceitar mais de uma fonte de contaminação. ◊

1.2 Propriedades das Distribuições Definidas pela Mistura na Escala de Gaussianas

Nas propriedades que se seguem,

$$p(x) = \int_0^\infty \frac{1}{\delta\psi(\sigma)}\phi\left(\frac{x}{\delta\psi(\sigma)}\right) dH(\sigma), \tag{1.6}$$

onde H é a função distribuição de probabilidade associada à σ (podendo ou não ser absolutamente contínua).

A proposição abaixo garante que a densidade p é estritamente quase-convexa[2] e uma característica interessante de funções estritamente quase convexas é que, assim como para funções estritamente convexas, caso estas funções admitam um ponto de mínimo, este ponto é único.

Proposição 1.2.1. *A densidade p é estritamente quase-côncava. Ou seja, para todo $\lambda \in (0,1)$, $p(\lambda x + (1-\lambda)y) > \min\{p(x), p(y)\}$.*

Demonstração. De fato, como a função $\exp\left\{-\frac{x^2}{2\psi(\sigma)^2}\right\}$ é estritamente quase-côncava, então,

$$\exp\left\{-\frac{(\lambda x + (1-\lambda)y)^2}{2\delta^2\psi(\sigma)^2}\right\} > \min\left\{\exp\left\{-\frac{x^2}{2\delta^2\psi(\sigma)^2}\right\}, \exp\left\{-\frac{y^2}{2\delta^2\psi(\sigma)^2}\right\}\right\}$$

[2]Uma função $f(x)$ é quase convexa se, para todo $\lambda \in (0,1)$ e para todos $x, y \in \mathbb{R}$, $f(\lambda x + (1-\lambda)y) \geq \min\{f(x), f(y)\}$. Se $f(\lambda x + (1-\lambda)y) > \min\{f(x), f(y)\}$, para todo $\lambda \in (0,1)$ e para todos $x, y \in \mathbb{R}$, então, dizemos que f é estritamente quase-convexa.

e pela continuidade da função exponencial e pelo fato dela ser positiva, segue que

$$\begin{aligned} p(\lambda x + (1-\lambda)y) &= \int_0^\infty \frac{1}{\delta\psi(\sigma)}\phi\left(\frac{\lambda x + (1-\lambda)y}{\delta\psi(\sigma)}\right) dH(\sigma) \\ &> \min\left\{\int_0^\infty \frac{1}{\delta\psi(\sigma)}\phi\left(\frac{x}{\delta\psi(\sigma)}\right) dH(\sigma), \int_0^\infty \frac{1}{\delta\psi(\sigma)}\phi\left(\frac{y}{\delta\psi(\sigma)}\right) dH(\sigma)\right\} \\ &= \min\{p(x), p(y)\}. \end{aligned}$$

Em particular, segue do fato acima que $-\log p(x)$ é estritamente quase-convexa. □

Propriedade 1.2.1. *A densidade $p(x)$ é finita em $x = 0$ se, e somente se, $E\frac{1}{\psi(\sigma)} < \infty$.*

Demonstração. Segue imediatamente do fato que $p(0) = \frac{\phi(0)}{\delta} E\frac{1}{\psi(\sigma)}$. □

Propriedade 1.2.2. *Se $X \sim p$, então,*

i. para $0 < |X| < \infty$, a distribuição condicional de σ dado X existe;

ii. para $X = 0$, a distribuição condicional de σ dado X existe se, e somente se, $E\frac{1}{\psi(\sigma)} < \infty$.

Demonstração. O item [i] segue imediatamente da expressão da distribuição conjunta de σ e X,

$$H(\sigma_0|X) = \frac{\int_0^{\sigma_0} \frac{1}{\psi(\sigma)}\phi\left(\frac{x}{\psi(\sigma)}\right) dH(\sigma)}{\int_0^{\infty} \frac{1}{\psi(\sigma)}\phi\left(\frac{x}{\psi(\sigma)}\right) dH(\sigma)}$$

Já o item [ii] segue imediatamente da combinação da expressão acima com a propriedade anterior. □

1.3 O Uso de Misturas Através da Escala na Prática

Nesta seção pretendemos justificar o uso de misturas de distribuições gaussianas através do parâmetro de escala em modelos de regressão e de séries temporais através de exemplos práticos.

1.3.1 Robustez

O uso de misturas na escala justifica-se principalmente como um meio de tornar os modelos estatísticos mais robustos, ou seja, menos sensíveis a *outliers* e a valores extremos do que ocorre ao se assumir explícita ou implicitamente (mínimos quadrados) um ruído gaussiano. De fato, pode-se observar esta propriedade a partir da figura 1.1, onde ilustramos algumas funções critérios, $\rho(x) = -\log p(x)$, obtidas ao assumir o ruído distribuido de acordo com uma mistura na escala. Note como algumas funções critério normalmente utilizadas em modelos robustos tais como o critério L^1 (distribuição de Laplace) ou função de Huber (distribuição logística) surgem naturalmente. Como veremos, o enfoque na verossimilhança aqui adotado permite o uso de técnicas próprias a estimadores de máxima verossimilhança, como o algoritmo EM, proporcionando um modo "universal" ou unificado de se estimar os parâmetros do modelo sob diferentes perspectivas (funções critério).

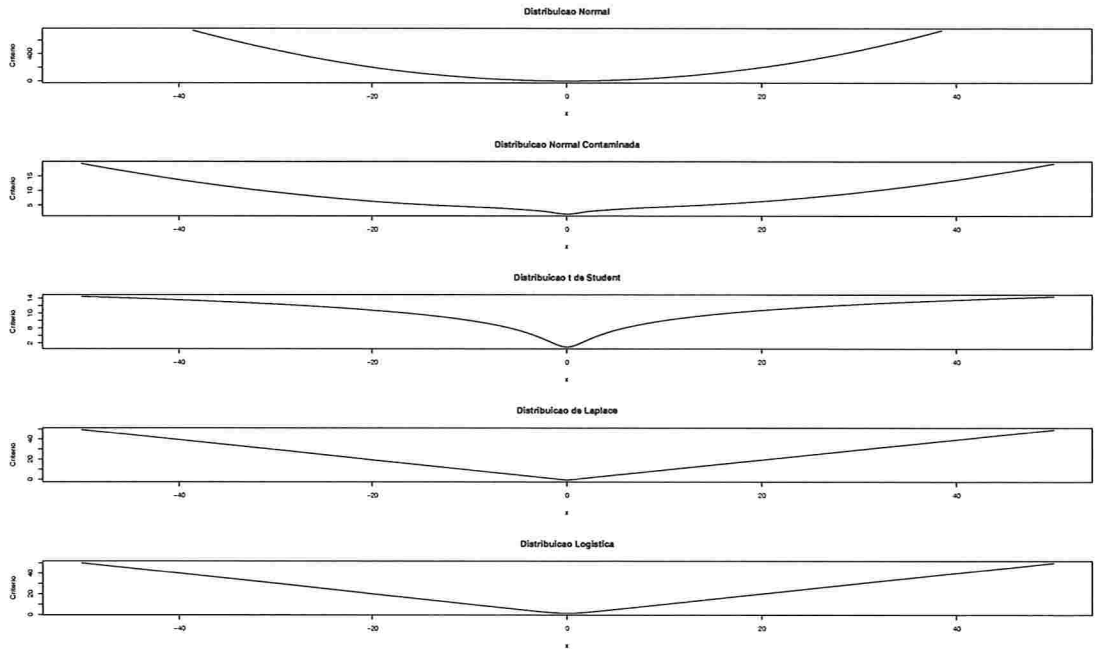


Figura 1.1: Funções critério associadas a algumas misturas na escala.

1.3.2 Aplicações

Para aplicações em finanças de distribuições pertencentes à classe das misturas de normais através da escala, veja [26], [27] para o uso da distribuição t de Student em modelos de volatilidade estocástica e [33] e [34], ou mais recentemente, o artigo [11] que trata da inferência via simulação de modelos generalizados de volatilidade estocástica definida por uma t de Student. Para o uso da distribuição t de Student em séries temporais do tipo GARCH, veja [6]. Veja também [29] para exemplos onde o uso da distribuição t de Student resulta em um ajuste melhor do que o uso da distribuição gaussiana.

Além de finanças, tem havido muito interesse em distribuições com caudas pesadas, especialmente aquelas pertencentes à classe das distribuições estáveis, na área de processamento de sinais. Para aplicações neste sentido, veja [31] e para aplicações no campo de processamento de sinais de audio e fala (com aplicações, por exemplo, em comunicação sem fio e sonares), para os quais existem justificativas físicas para a ocorrência de distribuições com caudas pesadas, veja [22], [20], [19], [21], nos quais utiliza-se principalmente o algoritmo de Gibbs para tratar sinais com ruído distribuido de acordo com uma distribuição estável.

Capítulo 2

Preliminares

2.1 Splines

2.1.1 Introdução

Seja $(\tau_i)_{i=1}^{l+1}$ uma seqüência estritamente crescente de pontos e seja $k \in \mathbb{Z}_+$. Uma função f é um *polinômio por partes de ordem* k se existir uma seqüência P_1,...,P_l de polinômios de ordem k tal que

$$f(x) = P_i(x), \text{ se } \tau_i < x < \tau_{i+1},$$

para $i = 1, ..., l$. Os pontos τ_i são chamados de *pontos de quebra* ou *nós* de f. Evidentemente, podemos sempre assumir que f está definida em toda a reta real, $\mathbb{R}$, definindo

$$f(x) = \begin{cases} P_1(x), & \text{se } x \leq \tau_1 \\ P_l(x), & \text{se } x \geq \tau_{l+1} \end{cases}$$

Dizemos que duas funções polinomiais *concordam entre si* se elas são compostas das mesmas componentes polinomiais e dos mesmos pontos de quebra.

O espaço de todas funções polinômiais de ordem k e com pontos de quebra $\boldsymbol{\tau} \equiv (\tau_i)_{i=1}^{l+1}$ é denotado por $\mathbb{P}_{k,\boldsymbol{\tau}}$. Note que $\mathbb{P}_{k,\boldsymbol{\tau}}$ é um espaço vetorial de dimensão kl. Note também que o espaço $\mathbb{P}_{k,\boldsymbol{\tau}}$ contém funções descontínuas em $\boldsymbol{\tau}$ já que não se impôs nenhuma condição sobre a continuidade de seus elementos nos pontos de quebra. Uma conseqüência disso é a possibilidade de se obter interpolações, ou estimativas, descontínuas. Além da continuidade, pode ser desejável que algumas das primeiras derivadas da função interpolante, ou estimada, sejam contínuas. Para tanto, consideramos um subconjunto de $\mathbb{P}_{k,\boldsymbol{\tau}}$ denotado por $\mathbb{P}_{k,\boldsymbol{\tau},\boldsymbol{\nu}}$, onde $\boldsymbol{\nu} \equiv (\nu_2, ..., \nu_l)$ representa o número de derivadas contínuas que f deve ter nos pontos de quebra $\tau_2, ..., \tau_l$.

Como pode-se ver abaixo, $\mathbb{P}_{k,\boldsymbol{\tau},\boldsymbol{\nu}}$ é um subespaço de $\mathbb{P}_{k,\boldsymbol{\tau}}$ cuja dimensão é $kl - \sum_{i=2}^{l} \nu_i$. Com efeito, isto segue do fato que o conjunto de potências truncadas é uma base de $\mathbb{P}_{k,\boldsymbol{\tau},\boldsymbol{\nu}}$.

2.1.2 B-Splines

Em resumo, as propriedades gerais de um B-spline, B, de ordem q são dadas por:

- B consiste de $q+1$ componentes polinomiais, cada uma delas de grau q;
- as componentes polinomiais se unem em q nós internos;
- nos pontos de junção, as derivadas de ordem até $q-1$ são contínuas;
- o B-spline é positivo no intervalo entre os nós extremos que o define e zero fora deste intervalo;
- exceto nas fronteiras, B sobrepõe-se com $2q$ componentes polinomiais de seus B-splines vizinhos;
- em um dado ponto x, $q+1$ B-splines são não-nulos (positivos);

As propriedades acima justificam a popularidade dos *B-splines* na aproximação de funções. Características tais como a compacidade das componentes do *B-spline* ou o fato dos mesmos serem compostos de partes polinomiais simplificam em muito a sua análise e implementação.

A seguir, apresentamos, em mais detalhes, definições e propriedades relativas aos B-splines.

Definição 2.1.1. *Seja $\boldsymbol{\tau} \equiv (\tau_i)$ uma seqüência não-decrescente de nós. A* i-ésima B-spline normalizada de ordem k para a seqüência de nós τ, *denotada por $B_{i,k,\boldsymbol{\tau}}$, é definida por*

$$B_{i,k,\boldsymbol{\tau}}(x) \equiv (\tau_{i+k} - \tau_i)[\tau_i, ..., \tau_{i+k}](\tau - x)_+^{k+1}$$

para todo $x \in \mathbb{R}$.

Na definição acima, $[\tau_1, ..., \tau_m]f(\cdot)$ representa a *k-ésima diferença dividida aplicada na função f*, veja [12] para mais detalhes.

Em geral, quando k e $\boldsymbol{\tau}$ são facilmente inferidos do contexto, a B-spline $B_{i,k,\boldsymbol{\tau}}$ é simplesmente denotada por B_i. Note que o operador de diferença dividida de ordem k na definição de B_i opera sobre τ.

Uma característica fundamental das B-splines é o fato delas terem suporte compacto. De fato, se B_i é uma B-spline, então B_i satisfaz:

i. $B_i(x) = 0$, para $x \in [\tau_i, \tau_{i+k}]^C$;

ii. $B_i(x) > 0$, para $x \in (\tau_i, \tau_{i+k})$.

Definição 2.1.2. *Uma* função spline de ordem k com nós em $\boldsymbol{\tau}$ *é qualquer combinação linear de B-splines de ordem k para a seqüência de nós $\boldsymbol{\tau}$. A coleção de todas estas funções é denotada por $\mathcal{S}_{k,\boldsymbol{\tau}}$, ie,*

$$\mathcal{S}_{k,\boldsymbol{\tau}} \equiv \left\{ \sum_i \alpha_i B_{i,k,\boldsymbol{\tau}} : \alpha_i \in \mathbb{R} \right\}.$$

O teorema abaixo caracteriza o espaço de funções spline $\mathcal{S}_{k,\boldsymbol{\tau}}$ em termos do espaço $\mathbb{P}_{k,\boldsymbol{\tau},\boldsymbol{\nu}}$.

Teorema 2.1.1. *Dada uma seqüência estritamente crescente* $\boldsymbol{\xi} \equiv (\xi_i)_1^{l+1}$ *e uma seqüência de inteiros não negativos* $\boldsymbol{\nu} \equiv (\nu_i)_2^l$, *com* $\nu_i \leq k$, *para todo* i, *defina*

$$n \equiv k + \sum_{i=2}^{l}(k - \nu_i) = kl - \sum_{i=2}^{l}\nu_i = \dim \mathbb{P}_{k,\boldsymbol{\tau},\boldsymbol{\nu}}.$$

Seja $\boldsymbol{\tau} \equiv (\tau_i)_1^{n+k}$ *uma seqüência não decrescente satisfazendo*

i. para $i = 2, ..., l$, *o número* ξ_i *ocorre* $k - \nu_i$ *vezes em* $\boldsymbol{\tau}$;

ii. $\tau_1 \leq ... \leq \tau_k \leq \xi_1$ *e* $\xi_{l+1} \leq \tau_{n+1} \leq ... \leq \tau_{n+k}$.

Então, a seqüência $B_1, ..., B_n$ *de B-splines de ordem* k *para a seqüência de nós é uma base para o espaço de funções* $\mathbb{P}_{k,\boldsymbol{\tau},\boldsymbol{\nu}}$ *quando restritas ao intervalo* $[\tau_k, \tau_{n+1}]$. *Ou seja*

$$\mathcal{S}_{k,\boldsymbol{\tau}} = \mathbb{P}_{k,\boldsymbol{\tau},\boldsymbol{\nu}},$$

em $[\tau_k, \tau_{n+1}]$.

Note que o teorema deixa em aberto a escolha dos k primeiros e k últimos nós. Como colocado por de Boor em [12], uma escolha conveniente é

$$\tau_1 = ... = \tau_k = \xi_1 \text{ e } \tau_{n+1} = ... = \tau_{n+k} = \xi_{l+1}$$

o que permite incluir estes nós no mesmo padrão que os demais definindo-se $\nu_1 = \nu_{l+1} = 0$. Ou seja, nenhuma condição de continuidade é imposta para os pontos extremos ξ_1 e ξ_{l+1} do intervalo de interesse.

2.1.3 P-Splines

Na prática, os coeficientes da aproximação de uma determinada função f em termos dos elementos de uma base de *B-splines* são desconhecidos e devem ser determinados de algum modo. A primeira alternativa a se considerar é via mínimos quadrados ordinários, ou seja, dados os pontos amostrais (y_t, x_t), para $t = 1, ..., T$, consideramos os coeficientes $\hat{a}_1, ..., \hat{a}_T$ que minimizam a função objetiva

$$S \equiv \sum_{t=1}^{T}\left\{y_t - \sum_{j=1}^{n} a_j B_j(x_t)\right\}^2.$$

O problema da aproximação acima é que, para um número relativamente grande de nós, a curva estimada

$$\hat{f}(\cdot) \equiv \sum_{j=1}^{n} \hat{a}_j B_j(\cdot)$$

pode apresentar uma variância muito grande, a depender da disposição dos dados. Um meio de reduzir a variabilidade da estimativa é impor algum tipo de penalização, levando assim à função objetiva

$$S \equiv \sum_{t=1}^{T} \left\{ y_t - \sum_{j=1}^{n} a_j B_j(x_t) \right\}^2 + \lambda J(\mathbf{a}).$$

O mais usual é considerar, como veremos mais abaixo, $J = \int [f'']^2$, porém, no contexto atual, onde buscamos aproximar f através de funções base, o equivalente a esta penalização seria assumir

$$J(\mathbf{a}) = \int \left\{ \sum_{j=1}^{n} a_j B_j''(x) \right\}^2 dx$$

, de modo que

$$S \equiv \sum_{t=1}^{T} \left\{ y_t - \sum_{j=1}^{n} a_j B_j(x_t) \right\} + \lambda \int \left\{ \sum_{j=1}^{n} a_j B_j''(x) \right\}^2 dx.$$

Observamos que não há nada de especial na escolha pela segunda derivada, de modo que, qualquer outra ordem de derivação poderia ser igualmente usada. De modo geral, o uso da primeira derivada leva a equações simples e a estimativas lineares por partes, enquanto que derivadas de maior ordem levam a estimativas mais suaves, porém, matematicamente mais complexas. Para o cálculo da integral na expressão, observamos que, no caso em que os pontos $x_1, ..., x_T$ são igualmente espaçados com espaçamento h e em que as funções base $\{B_j\}$ são *B-splines*, as segundas derivadas podem ser obtidas através da expressão

$$h^2 \sum_j a_j B_j''(x;k) = \sum_j \Delta^2 a_j B_j(x;k-2)$$

onde $\Delta^2 a_j = \Delta\Delta a_j = a_j - 2a_{j-1} + a_{j-2}$, veja [12].

Uma alternativa às penalizações usando derivadas é considerar, como proposto em [16], as diferenças finitas dos coeficientes de *B-splines* adjacentes:

$$S \equiv \sum_{t=1}^{T} \left\{ y_t - \sum_{j=1}^{n} a_j B_j(x_t) \right\}^2 + \lambda \sum_{j=d+1}^{n} (\Delta^d a_j)^2. \tag{2.1}$$

A penalização via operadores de diferença é uma boa aproximação discreta da integral do quadrado da k-ésima derivada da aproximação da função alvo via *B-splines*. O uso conjugado de *B-splines* e penalizações como em (2.1) é conhecido como *P-splines*[1]. Uma vantagem do uso de *P-splines* é que, relativamente à penalização diretamente através uso das derivadas dos *B-splines*, é muito fácil construir um procedimento automático de estimação dos coeficientes $a_1, ..., a_T$ através da minimização da função objetiva S.

[1] aqui P é para indicar o termo *penalization*

2.1.4 Smoothing Splines

Diferentemente dos suavizadores sugeridos acima (*B-splines* e *P-splines*), este suavizador não é definido explicitamente em termos de funções base. Ao contrário, ele surge como o resultado de um problema de otimização: *entre todas as funções $f(x)$ pertencentes ao espaço $\mathcal{S}_2[a,b]$ de funções definidas no intervalo $[a,b]$, com as duas primeiras derivadas contínuas e com f'' integrável, ache aquela que minimiza a soma residual dos quadrados penalizada:*

$$\mathcal{L}(f) \equiv \sum_{i=1}^{T}(y_i - f(x_i))^2 + \lambda \int_a^b [f''(t)]^2 dt. \tag{2.2}$$

O parâmetro λ, conhecido por *parâmetro de suavização* é uma constante fixada. Quanto à solução do problema acima, pode-se mostrar que a solução existe, é única e é dada por um *spline* cúbico natural com nós nos valores únicos de x_i. Observamos ainda que a função de penalização $\lambda \int_a^b [f''(t)]^2 dt$ age diretamente sobre a suavidade da função estimada e penaliza funções com curvatura muito acentuada em diversos pontos, suavizando, deste modo, a estimativa final da função alvo f. Mais ainda, sabe-se que, no extremo, quando $\lambda \to \infty$, o termo de penalização domina forçando a que $f'' \equiv 0$ e, conseqüentemente, a que a função que resolve o problema de otimização sugerido seja a reta de mínimos quadrados. Por outro lado, quando $\lambda \to 0$, o termo de penalização perde importância e a solução tende a uma função duas vezes diferenciável e que interpole os dados. Finalmente, observamos que a solução de um problema de otimização análogo ao sugerido acima, mas com função objetiva dada por

$$\mathcal{L}(f) \equiv \sum_{i=1}^{T}\omega_t(y_i - f(x_i))^2 + \lambda \int_a^b [f''(t)]^2 dt, \tag{2.3}$$

onde $(\omega_1, ..., \omega_T)'$ é um vetor de ponderações, também é um *spline* cúbico natural.

Para referência futura, enunciamos as seguintes proposições nos quais o primeiro afirma que um conjunto arbitrário de pontos pode ser interpolado de modo único, enquanto que o segundo garante a unicidade dentro do espaço gerado por *splines* cúbicos,

Proposição 2.1.1. *Se $T \geq 2$, então, dados quaisquer $z_1, ..., z_T$, existe um único spline cúbico natural s com nós em $x_1 < ... < x_T$ tal que $s(x_t) = z_t$, para $t = 1, ..., T$.*

Proposição 2.1.2. *Suponha que $T \geq 2$ e que s é um spline cúbico interpolando os valores $z_1, ..., z_T$ nos pontos $x_1, ..., x_T$, os quais satisfazem $0 < x_1 < ... < x_T < b$. Seja g qualquer função em $\mathcal{S}_2[a,b]$ para a qual $g(x_t) = z_t$, para $t = 1, ..., T$. Então, $\int [g'']^2 \geq \int [s'']^2$, com igualdade apenas se $g \equiv s$.*

Em particular, segue das proposições acima o resultado que garante que a solução do problema de otimização proposto acima é um *spline* cúbico natural.

Teorema 2.1.2. *Sejam $T \geq 3$ e $x_1, ..., x_T$ pontos satisfazendo $a < x_1 < ... < x_T < b$. Então, dadas as observações $y_1, ..., y_T$ e um parâmetro de suavização λ estritamente positivo, para toda $g \in \mathcal{S}_2[a,b]$, existe um spline cúbico natural s tal que $\mathcal{L}(s) \leq \mathcal{L}(g)$, com igualdade apenas se $g \equiv s$.*

Para mais detalhes com relação a *smoothing-splines* e aos resultados acima, sugerimos o livro [24].

2.2 O Algoritmo EM

O algoritmo EM é um método iterativo de se obter estimadores de máxima verossimilhança em situações nas quais os métodos usuais de maximização são difíceis de se aplicar. Originalmente, o algoritmo EM foi concebido como um meio de se maximizar distribuições a posteriori, veja [13]. O algoritmo EM consiste em aumentar os dados efetivamente observados, $\mathbf{y}$ utilizando variáveis latentes (ou não observadas) $\mathbf{X}$ para estimar o vetor de parâmetros, θ, através dos passos:

1. **esperança — passo E**: cálculo da esperança da log-verossimilhança da variável latente dadas as variáveis observadas e os parâmetros do modelo, ie,

$$Q(\theta|\theta^{(p)}) \equiv E\{\log g_c(x|\theta)|y, \theta^{(p)}\}$$

onde g_c é a densidade do conjunto completo de observações e $\theta^{(p)}$ é o vetor de parâmetros estimado no p-ésimo estágio do algoritmo;

2. **maximização — passo M**: escolher $\theta^{(p+1)}$ de modo a maximizar $Q(\theta|\theta^{(p)})$.

Suponha que $x \in \mathcal{X}$ seja observado apenas parcialmente através da variável $y \in \mathcal{Y}$, ie, $y = y(x)$. Em outros termos, x corresponde aos dados completos, enquanto que y aos dados efetivamente observados, embora incompletos. Seja $g_c(x|\theta)$ a densidade de X associada ao parâmetro θ. Então, a densidade dos dados incompletos é dada, conseqüentemente, por

$$g(y|\theta) = \int_{\mathcal{X}} f(x|\theta)dx.$$

O algoritmo EM consiste em determinar iterativamente o parâmetro θ^* que maximiza a log-verossimilhança $L(\theta) = \log g(y|\theta)$ dado um vetor de observações y. Como descrito acima, o algoritmo EM divide-se basicamente em duas etapas: a etapa da esperança (E) e a etapa da maximização (M). A esperança no passo (E) é com relação à densidade condicional de X dado $Y = y$ e θ denotada por

$$k(x|y, \theta) \equiv \frac{g_c(x|\theta)}{g(y|\theta)}.$$

Em particular, note que a log-verossimilhança $L(\theta)$ pode ser escrita na forma

$$\begin{aligned} L(\theta) &= \log g_c(x|\theta) - \log k(x|y, \theta) \\ &= L_c(\theta) - \log k(x|y, \theta), \end{aligned} \tag{2.4}$$

onde $L_c(\theta) \equiv \log g_c(x|\theta)$, e que

$$Q(\theta|\theta') = L(\theta) + E\{\log k(x|y, \theta)|y, \theta'\} = L(\theta) + H(\theta|\theta'), \tag{2.5}$$

onde $H(\theta|\theta') \equiv E\{\log k(x|y, \theta)|y, \theta'\}$. Finalmente, os passos (E) e (M) do algoritmo são repetidos alternadamente até que a diferença $L(\theta^{(p+1)}) - L(\theta^{(p)})$ seja suficientemente pequena.

Monotonicidade do Algoritmo EM

As estimativas obtidas a cada ciclo completo do algoritmo EM satisfazem a desigualdade:

$$L(\theta^{(k+1)}) \geq L(\theta^{(k)}) \tag{2.6}$$

para todo $k = 0, 1, 2, ...$ Ou seja, se a log-verossimilhança $L(\theta)$ admitir um limitante superior, então, a seqüência $\{L(\theta^{(k)})\}$ é convergente.

Agora, note que, da identidade (2.5), vale a igualdade

$$\begin{aligned}\log L(\theta^{(k+1)}) - \log L(\theta^{(k)}) = &\{Q(\theta^{(k+1)}|\theta^{(k)}) - Q(\theta^{(k)}|\theta^{(k)})\} \\ &- \{H(\theta^{(k+1)}|\theta^{(k)}) - H(\theta^{(k)}|\theta^{(k)})\}\end{aligned}$$

e que, por definição, $Q(\theta^{(k+1)}|\theta^{(k)}) \geq Q(\theta^{(k)}|\theta^{(k)})$. Logo, para demonstrar a desigualdade (2.6), basta checar a desigualdade $H(\theta^{(k+1)}|\theta^{(k)}) \leq H(\theta^{(k)}|\theta^{(k)})$, o que não é muito difícil, pois, ela segue imediatemente da desigualdade de Jensen e da concavidade da função logarítmica.

2.2.1 O Algoritmo EM Generalizado

O algoritmo EM generalizado enfraquece a exigência da maximização de Q no sentido de que ele pede apenas que

$$Q(\theta^{(k+1)}|\theta^{(k)}) \geq Q(\theta^{(k)}|\theta^{(k)}).$$

Note que pela argumentação acima, esta desigualdade é suficiente para garantir a monotonicidade do algoritmo EM generalizado.

2.2.2 A Taxa de Convergência do Algoritmo EM

O algoritmo EM implicitamente define uma aplicação $\theta \rightarrow M(\theta)$ do espaço paramétrico sobre ele mesmo tal que $\theta^{(k+1)} = M(\theta^{(k)})$, para $k = 1, 2, ...$ Além disso, se a seqüência $\theta^{(k)}$ converge para θ^* e M é contínua, então, $M(\theta^*) = \theta^*$. Consequentemente, expandindo M em uma série de Taylor ao redor de θ^*, temos

$$\theta^{(k+1)} - \theta^* = DM(\theta^{(k)} - \theta^*) + O(\|\theta^{(k)} - \theta^*\|^2),$$

onde

$$DM = \left(\frac{\partial M_i(\theta)}{\partial \theta_j}\right)\Bigg|_{\theta=\theta^*}$$

é a matriz Jacobiana para $M(\theta) = (M_1(\theta), ..., M_d(\theta))'$ calculada em $\theta = \theta^*$. A taxa de convergência é então definida por

$$R = \lim_{k\rightarrow\infty} \frac{\|M(\theta^{(k)}) - \theta^*\|}{\|\theta^{(k)} - \theta^*\|},$$

de modo que a taxa de convergência do algoritmo EM é o maior autovalor de $DM(\theta^*)$.

2.2.3 O Algoritmo EM Penalizado — EMP

Em determinados casos, o estimador de máxima verossimilhança usual pode não ser o mais adequado. Por exemplo, caso tenhamos alguma informação a priori a respeito de $\boldsymbol{\theta}$, então o estimador de máxima verossimilhança a posteriori pode ser visto como um problema de máxima verossimilhança penalizada. Um outro caso é quando necessitamos impor algum ajuste que pode ser representado na forma de uma penalização sobre os parâmetros do modelo como, por exemplo, no caso de *smoothing-splines* ou de *ridge regression* quando a matriz $X'X$ é quase singular. O algoritmo EM penalizado que descreveremos agora estende o algoritmo EM usual e foi introduzido em [23].

O algoritmo EM penalizado difere do algoritmo EM, pois, ele consiste em resolver o problema de encontrar o vetor de parâmetros $\boldsymbol{\theta}$ que maximiza

$$L(\boldsymbol{\theta}) - \lambda J(\boldsymbol{\theta})$$

ao invés de pura e simplesmente $L(\boldsymbol{\theta})$. Na prática, apenas o passo M sofre alguma alteração, enquanto que o passo E permanece exatamente o mesmo:

passo E: calcular a esperança

$$Q(\theta|\theta^{(k)}) = E_{\theta^{(k)}}\{\log g_c(x|\theta)|y, \theta^{(k)}\},$$

enquanto que, o passo M fica dado por

passo M: maximizar

$$S(\theta|\theta^{(k)}) = Q(\theta|\theta^{(k)}) - \lambda J(\theta),$$

onde $J(\theta)$ é a função de penalização associada ao modelo e λ um parâmetro de penalização. Em problemas regulares, o passo M é realizado resolvendo-se a equação

$$\frac{\partial}{\partial \boldsymbol{\theta}} Q(\theta|\theta^{(k)}) - \lambda \frac{\partial}{\partial \boldsymbol{\theta}} J(\theta) = 0. \tag{2.7}$$

Assim como o algoritmo EM usual, o algoritmo EMP define uma aplicação do espaço paramétrico nele mesmo denotada por M tal que $\boldsymbol{\theta}^{(k+1)} = M(\boldsymbol{\theta}^{(k)})$ e, de novo, assim como para o algoritmo EM, a taxa de convergência do algoritmo EMP é obtida através da matriz $DM(\boldsymbol{\theta}^*)$, a qual é dada por

$$DM(\boldsymbol{\theta}^*) = (B + C - \lambda K)^{-1} C, \tag{2.8}$$

onde $C = -D^{20}H(\boldsymbol{\theta}^*|\boldsymbol{\theta}^*)$, $B = -C - D^{20}Q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^*)$ e $K = D^2 J(\boldsymbol{\theta}^*)$ e onde D^{20} representa a segunda derivada (Hessiana de H) com relação ao primeiro parâmetro e D^2 a segunda derivada de J.

O Método One-Step-Late (OSL)

Green ([23]) sugere o algoritmo *One-Step-Late* (OSL) como meio de estimação de $\boldsymbol{\theta}$. O método é uma simples variação do algoritmo EM, sendo que a única diferença ocorre no passo M do algoritmo e consiste, no passo $k+1$, em se obter a estimativa $\boldsymbol{\theta}^{(k+1)}$ que resolve a equação

$$\frac{\partial}{\partial \boldsymbol{\theta}} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)}) - \lambda \frac{\partial}{\partial \boldsymbol{\theta}} J(\boldsymbol{\theta}^{(k)}) = 0. \tag{2.9}$$

A única diferença entre (2.9) e (2.7) é que, em (2.9), as derivadas do termo de penalização são calculadas no valor corrente de θ. É importante observar que ambas as equações, (2.9) e (2.7), têm os mesmos pontos fixos, logo, se o método OSL converge, o limite é uma estimativa de máxima verossimilhança penalizada.

O método OSL também define uma aplicação do espaço paramétrico nele mesmo denotada por $N(\theta)$. Neste caso, a matriz que controla o comportamento do método é dada por

$$DN(\theta^*) = (B + C)^{-1}(C - \lambda K), \tag{2.10}$$

onde B e C são como acima.

2.2.4 Informação Inexistente

A Estatística Escore

As *estatísticas escore* para o conjunto de dados completo e incompleto são dadas por

$$S(\mathbf{y}; \theta) = \frac{\partial}{\partial \theta} \log L(\theta)$$

e

$$S_c(\mathbf{x}; \theta) = \frac{\partial}{\partial \theta} \log L_c(\theta)$$

respectivamente.

As estatísticas escore para dados incompletos e completos estão relacionadas entre si no sentido que a estatística escore para dados incompletos pode ser escrita como o valor esperado da estatística escore para dados completos:

$$S(\mathbf{y}; \theta) = E\{S_c(\mathbf{x}; \theta)|y, \theta\}.$$

Princípio da Informação Faltante

Considere a matriz de informação de Fisher dada por

$$I(\theta; \mathbf{y}) = -\frac{\partial^2}{\partial \theta \partial \theta'} \log L(\theta).$$

Sob certas condições de regularidade, a matriz de informação de Fisher esperada $\mathcal{I}(\theta)$ é dada por

$$\begin{aligned} \mathcal{I}(\theta) &= E\{S(\mathbf{y}; \theta) S(\mathbf{y}; \theta)'|\theta\} \\ &= -E\{I(\theta; \mathbf{y})|\theta\}. \end{aligned}$$

Analogamente, considere a matriz de informação de Fisher

$$I_c(\theta; \mathbf{x}) = -\frac{\partial^2}{\partial \theta \partial \theta'} \log L_c(\theta).$$

e a matriz de informação de Fisher esperada $\mathcal{I}(\theta)$

$$\mathcal{I}_c(\theta) = -E\{I_c(\theta; \mathbf{x})|\theta\}$$

para o conjunto completo de observações. Segue de (2.4) que a relação entre a matriz de informação de Fisher para dados incompletos e completos satisfazem a relação

$$I(\theta; \mathbf{y}) = I_c(\theta; \mathbf{x}) + \frac{\partial^2}{\partial\theta\partial\theta'} \log k(\mathbf{x}|\mathbf{y}; \theta).$$

Tomando a esperança em ambos os lados da expressão acima, temos

$$\mathcal{I}(\theta; \mathbf{y}) = \mathcal{I}_c(\theta; \mathbf{y}) - \mathcal{I}_m(\theta; \mathbf{y})$$

onde

$$\mathcal{I}_c(\theta; \mathbf{y}) = E\{I_c(\theta; \mathbf{x})|\mathbf{y}, \theta\}$$

e

$$\mathcal{I}_m(\theta; \mathbf{y}) = -E\left\{\frac{\partial^2}{\partial\theta\partial\theta'} \log k(\mathbf{x}|\mathbf{y}; \theta)|\mathbf{y}, \theta\right\}.$$

A matriz de informação $\mathcal{I}_m(\theta; \mathbf{y})$ é conhecida por *matriz de informação inexistente*[2]. Em outras palavras, a informação observada é igual a informação completa (esperada e condicianada no vetor de observações $\mathbf{y}$) menos a informação faltante ou não-observada. Finalmente, tomando esperanças com relação à distribuição de $\mathbf{Y}$, temos que

$$\mathcal{I}(\theta) = \mathcal{I}_c(\theta) - E\{\mathcal{I}_m(\theta; \mathbf{y})\}.$$

2.2.5 Monte Carlo EM — MCEM

Embora em muitos casos espera-se, no passo E do algoritmo, calcular analiticamente a esperança $Q(\theta|\theta')$, isto nem sempre é possível. Nestes casos, deve-se recorrer a métodos numéricos ou de Monte-Carlo. Nesta tese sugerimos o uso do segundo método que consiste basicamente em se obter uma amostra pseudo-aleatória $x_1, ..., x_{j'}$ da distribuição $k(x|y, \theta)$ e, então, maximizar a log-verossimilhança

$$\hat{Q}(\theta|\theta') \equiv \frac{1}{j'} \sum_{i=1}^{j'} \log \mathcal{L}(\theta|x, z).$$

Note que o tamanho da amostra sorteada pode variar de passo para passo dentro do algoritmo EM. Como é sabido, quando $j' \to \infty$, a quantidade $\hat{Q}(\theta|\theta')$ converge quase certamente para $Q(\theta|\theta')$ de modo que, no limite, o método MCEM coincide com o método EM.

[2] *Missing Information Matrix.*

2.3 Convergência do Algoritmo EM

Para qualquer valor L_0, seja

$$\mathcal{L}(L_0) \equiv \{\theta \in \Omega : L(\theta) = L_0\}.$$

Teorema 2.3.1. *Seja $\{\theta^{(k)}\}$ uma instância de um algoritmo EM generalizado satisfazendo a propriedade*

$$\frac{\partial}{\partial\theta} Q(\theta|\theta^{(k)})\Big|_{\theta=\theta^{(k+1)}} = 0.$$

Assuma que $\partial Q(\theta|\zeta)/\partial\theta$ seja contínua em θ e ζ. Então, $\theta^{(k)}$ converge para um ponto estacionário θ^* com $L(\theta^*) = L^*$, o limite de $L(\theta^{(k)})$, se uma das seguintes condições for satisfeita:

i. $\mathcal{L}(L^*) = \{\theta^*\}$;

ii. $\|\theta^{(k+1)} - \theta^{(k)}\| \to 0$, *quando* $k \to \infty$, *e* $\mathcal{L}(L^*)$ *é discreto.*

Corolário 2.3.2. *Suponha que $L(\theta)$ seja unimodal em Ω com θ^* sendo o único ponto estacionário e que $\partial Q(\theta|\zeta)/\partial\theta$ seja contínuo em θ e ζ. Então, qualquer seqüência EM $\{\theta^{(k)}\}$ converge para o único maximizador θ^* de $L(\theta)$. Ou seja, $\{\theta^{(k)}\}$ converge para o único estimador de máxima verossimilhança de θ.*

Capítulo 3

Modelos Lineares

3.1 O Modelo

Considere o modelo

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \delta\boldsymbol{\epsilon}, \tag{3.1}$$

onde $\boldsymbol{\beta}$ é um vetor de coeficientes, δ um parâmetro de escala e $\boldsymbol{\epsilon}$ é um vetor cujas componentes, ϵ_t, são distribuídas de acordo com $Z/\sigma^{1/2}$, onde Z é uma v.a. seguindo a distribuição normal padrão e $\sigma \sim h$. Este modelo é importante para nós, pois, servirá como base para os desenvolvimentos futuros em modelos mais gerais. No modelo (3.1), o conjunto de parâmetros é formado pelos vetor de coeficientes e pelo parâmetro de escala, ie, por $\boldsymbol{\theta} = (\boldsymbol{\beta}', \delta)'$.

3.1.1 Estimadores de Máxima Verossimilhança

Dado que $y_t \sim SM_h((\mathbf{X}\boldsymbol{\beta})_t, \delta; \psi)$, onde $h = h(\cdot|\boldsymbol{\zeta})$, então sua função densidade de probabilidade é dada por

$$p(y_t|\mathbf{X}_t) = \int_0^\infty \frac{\sigma_t^{1/2}}{\delta}\phi\left(\frac{\sigma_t^{1/2}}{\delta}(y_t - \mathbf{X}_t'\boldsymbol{\beta})\right) h(\sigma_t|\boldsymbol{\zeta})d\sigma_t \tag{3.2}$$

e, portanto, a log-verossimilhança associada às observações $(y_1, ..., y_T)'$ e $(x_1, ..., x_T)'$, $l(\boldsymbol{\beta}, \delta, \boldsymbol{\zeta})$, é igual a

$$\sum_{t=1}^{T} \log \int_0^\infty \frac{\sigma_t^{1/2}}{\delta}\phi\left(\frac{\sigma_t^{1/2}}{\delta}(y_t - \mathbf{X}_t'\boldsymbol{\beta})\right) h(\sigma_t|\boldsymbol{\zeta})d\sigma_t.$$

onde $\mathbf{X}_t$ é a t-ésima linha da matriz $\mathbf{X}$. Embora as densidades acima sejam condicionadas no vetores $\mathbf{X}_t$, para reduzir a notação, passaremos a representa-las simplesmente por $p(y_t)$, ao invés de $p(y_t|\mathbf{X}_t)$.

Denotando

$$\int_0^\infty \frac{\sigma_t^{1/2}}{\delta}\phi\left(\frac{\sigma_t^{1/2}}{\delta}(y_t - \mathbf{X}_t'\boldsymbol{\beta})\right) h(\sigma_t|\boldsymbol{\zeta})d\sigma_t$$

por $I_t(\boldsymbol{\beta}, \delta, \boldsymbol{\zeta})$, e assumindo que podemos passar o operador de diferenciação para dentro da integral, temos que

$$\begin{aligned}\frac{\partial}{\partial \mathbf{c}} l(\mathbf{c}, \delta, \boldsymbol{\zeta}) &= \sum_{t=1}^{T} \frac{1}{I_t(\mathbf{c}, \delta, \boldsymbol{\zeta})} \int_0^{\infty} \frac{\sigma_t^{3/2}}{\delta^3} \phi\left(\frac{\sigma_t^{1/2}}{\delta}(y_t - \mathbf{X}_t'\boldsymbol{\beta})\right) h(\sigma_t|\boldsymbol{\zeta}) d\sigma_t \\ &\quad (y_t - \mathbf{X}_t'\boldsymbol{\beta})\mathbf{X}_t \\ &= \frac{1}{\delta^2} \sum_{t=1}^{T} E_{\mathbf{c},\delta,\boldsymbol{\zeta}}\{\sigma_t\}(y_t - \mathbf{X}_t'\boldsymbol{\beta})\mathbf{X}_t\end{aligned}$$

onde $E_{\mathbf{c},\delta,\boldsymbol{\zeta}}$ é calculada com relação à densidade proporcional a

$$\frac{\sigma_t^{1/2}}{\delta} \phi\left(\frac{\sigma_t^{1/2}}{\delta}(y_t - \mathbf{X}_t'\boldsymbol{\beta})\right) h(\sigma_t|\boldsymbol{\zeta}).$$

Definindo a matriz

$$W = W(\mathbf{c}, \delta, \boldsymbol{\zeta}) = \text{diag}\left(E_{\mathbf{c},\delta,\boldsymbol{\zeta}}\{\sigma_1\}, ..., E_{\mathbf{c},\delta,\boldsymbol{\zeta}}\{\sigma_T\}\right),$$

então,

$$\frac{\partial}{\partial \mathbf{c}} l(\mathbf{c}, \boldsymbol{\zeta}) = \frac{1}{\delta^2} \mathbf{X}' W(\mathbf{c}, \delta, \boldsymbol{\zeta})(\mathbf{y} - \mathbf{X}\mathbf{c}). \tag{3.3}$$

Para o parâmetro de escala, note que

$$\begin{aligned}\frac{\partial l}{\partial \delta} &= \sum_{t=1}^{T} \frac{1}{I_t(\mathbf{c}, \delta, \boldsymbol{\zeta})} \Bigg\{ -\frac{1}{\delta} I_t(\mathbf{c}, \delta, \boldsymbol{\zeta}) + \\ &\quad \int_0^{\infty} \frac{\sigma_t^{1/2}}{\delta} \phi\left(\frac{\sigma_t^{1/2}}{\delta}(y_t - \mathbf{X}_t'\boldsymbol{\beta})\right) \frac{\sigma_t}{\delta^3}(y_t - \mathbf{X}_t'\boldsymbol{\beta})^2 d\sigma_t \Bigg\} d\sigma_t \\ &= -\frac{T}{\delta} + \sum_{t=1}^{T} E_{\mathbf{c},\delta,\boldsymbol{\zeta}}\{\sigma_t\} \frac{1}{\delta^3}(y_t - \mathbf{X}_t'\boldsymbol{\beta})^2,\end{aligned}$$

e, portanto,

$$\frac{\partial l}{\partial \delta} = -\frac{T}{\delta} + \frac{1}{\delta^3}(\mathbf{y} - \mathbf{X}\mathbf{c})' W(\mathbf{c}, \delta, \boldsymbol{\zeta})(\mathbf{y} - \mathbf{X}\mathbf{c}). \tag{3.4}$$

De modo análogo, o gradiente da log-verossimilhança com relação a $\boldsymbol{\zeta}$ é dado por

$$\sum_{t=1}^{T} \frac{1}{I_t(\mathbf{c}, \delta, \boldsymbol{\zeta})} \int_0^{\infty} \frac{\sigma_t^{1/2}}{\delta} \phi\left(\frac{\sigma_t^{1/2}}{\delta}(y_t - \mathbf{X}_t'\boldsymbol{\beta})\right) \frac{\partial}{\partial \boldsymbol{\zeta}} h(\sigma_t|\boldsymbol{\zeta}) d\sigma_t.$$

Logo, ao dividir e multiplicar o integrando da expressão acima por $h(\sigma_t|\boldsymbol{\zeta})$, teremos

$$\frac{\partial}{\partial \boldsymbol{\zeta}} l(\mathbf{c}, \delta, \boldsymbol{\zeta}) = \sum_{t=1}^{T} E_{\mathbf{c},\delta,\boldsymbol{\zeta}} \left\{ \frac{\partial}{\partial \boldsymbol{\zeta}} \log h(\sigma_t|\boldsymbol{\zeta}) \right\}. \tag{3.5}$$

Finalmente, igualando as expressões (3.3), (3.4) e (3.5) a zero, obtemos as estimativas de c, δ e ζ, respectivamente. Mais precisamente,

$$\widehat{\mathbf{c}} = (\mathbf{X}'W(\widehat{\mathbf{c}},\widehat{\delta},\widehat{\boldsymbol{\zeta}})\mathbf{X})^{-1}\mathbf{X}'W(\widehat{\mathbf{c}},\widehat{\delta},\widehat{\boldsymbol{\zeta}})\mathbf{y}$$
$$\widehat{\delta}^2 = \frac{1}{T}(\mathbf{y}-\mathbf{X}\widehat{\mathbf{c}})'W(\widehat{\mathbf{c}},\widehat{\delta},\widehat{\boldsymbol{\zeta}})(\mathbf{y}-\mathbf{X}\widehat{\mathbf{c}})$$
$$\frac{\partial}{\partial\boldsymbol{\zeta}}l(\widehat{\mathbf{c}},\widehat{\delta},\widehat{\boldsymbol{\zeta}}) = 0.$$

Obviamente, as equações resultantes deste procedimento são não-lineares e um procedimento iterativo deve ser adotado. O problema com este enfoque é que assumimos explicitamente a possibilidade de inverter a ordem dos operadores de diferenciação e de integração. Ocorre que, se isso sempre fosse possível dentro da classe das distribuições obtidas através da mistura de gaussianas na escala, todos os elementos desta classe seriam diferenciáveis com relação aos parâmetros c e δ, o que é falso.[1] Para superar esta dificuldade, sugerimos o uso do algoritmo EM como descrito na seção a seguir.

3.1.2 Aplicando o Algoritmo EM

Para aplicar o algoritmo EM, consideramos $\sigma_1,...,\sigma_T$ como variáveis não observáveis e definimos o vetor completo de variáveis por $\mathbf{y}_c = (\mathbf{y}',\boldsymbol{\sigma}')'$, onde $\boldsymbol{\sigma} = (\sigma_1,...,\sigma_T)'$. Dado que, condicionado em $\boldsymbol{\sigma}$, y segue uma distribuição normal,

$$\begin{aligned}\log p(y_t,\sigma_t|\boldsymbol{\theta}) &= \log p(y_t|\sigma_t,\boldsymbol{\theta}) + \log h(\sigma_t|\boldsymbol{\theta}) \\ &= \log\left[\frac{\sigma^{1/2}}{\delta}\phi\left(\frac{\sigma_t^{1/2}}{\delta}(y_t-\mathbf{X}_t'\boldsymbol{\beta})\middle|\boldsymbol{\theta}\right)\right] + \log h(\sigma_t|\boldsymbol{\theta})\end{aligned}$$

onde $\mathbf{X}_t$ é a t-ésima linha de $\mathbf{X}$. Agora,

$$\frac{\sigma_t^{1/2}}{\delta}\phi\left(\frac{\sigma_t^{1/2}}{\delta}(y_t-\mathbf{X}_t'\boldsymbol{\beta})\middle|\boldsymbol{\theta}\right) = \frac{1}{\sqrt{2\pi}\delta\psi(\sigma_t)}e^{-\frac{\sigma_t}{2\delta^2}(y_t-\mathbf{X}_t'\boldsymbol{\beta})^2}$$

de modo que

$$\log\left[\frac{\sigma_t^{1/2}}{\delta}\phi\left(\frac{\sigma_t^{1/2}}{\delta}(y_t-\mathbf{X}_t'\boldsymbol{\beta})\middle|\boldsymbol{\theta}\right)\right] = -\frac{1}{2}\log(2\pi)+\frac{1}{2}\log\sigma_t-\log\delta-\frac{\sigma_t}{2\delta^2}(y_t-\mathbf{X}_t'\boldsymbol{\beta})^2.$$

[1] Um exemplo que quebra a hipótese de diferenciabilidade é a distribuição exponencial dupla, a qual pode ser representada como uma mistura de gaussianas, mas não é diferenciável.

Logo, ignorando o termo constante $-\frac{1}{2}\log(2\pi)$, temos

$$\begin{aligned} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)}) = & \frac{1}{2}\sum_{t=1}^{T} E_{\boldsymbol{\theta}^{(k)}}\{\log\sigma_t|y_t\} - T\log\delta \\ & - \frac{1}{2\delta^2}\sum_{t=1}^{T} E_{\boldsymbol{\theta}^{(k)}}\{\sigma_t|y_t\}(y_t - \mathbf{X}_t'\boldsymbol{\beta})^2 \\ & + \sum_{t} E_{\boldsymbol{\theta}^{(k)}}\{\log h(\sigma_t|\boldsymbol{\theta})|y_t\}. \end{aligned} \tag{3.6}$$

A expressão acima pode ser simplificada, definindo

$$C(\boldsymbol{\zeta};\boldsymbol{\theta}^{(k)}) \equiv \sum_{t=1}^{T} E_{\boldsymbol{\theta}^{(k)}}\left\{\log(\sigma_t^{1/2}h(\sigma_t|\boldsymbol{\theta}))|y_t\right\}, \tag{3.7}$$

como a componente que contém apenas os parâmetros associados à distribuição do ruído e

$$W_T(\boldsymbol{\theta}^{(k)}) \equiv \text{diag}\left(E_{\boldsymbol{\theta}^{(k)}}\{\sigma_1|y_1\}, ..., E_{\boldsymbol{\theta}^{(k)}}\{\sigma_T|y_T\}\right) \tag{3.8}$$

de modo que

$$\frac{1}{2\delta^2}\sum_{t=1}^{T} E_{\boldsymbol{\theta}^{(k)}}\{\sigma_t|y_t\}(y_t - f_{\boldsymbol{\theta}}(x_t))^2 = \frac{1}{2\delta^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'W_T(\boldsymbol{\theta}^{(k)})(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

e

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)}) = C(\boldsymbol{\zeta};\boldsymbol{\theta}^{(k)}) - T\log\delta - \frac{1}{2\delta^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'W_T(\boldsymbol{\theta}^{(k)})(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}). \tag{3.9}$$

Com relação à distribuição de σ_t dado y_t, lembramos que

$$K_{\boldsymbol{\theta}}(\sigma_t|y_t) = \frac{\int_0^{\sigma_t}\sigma_t^{1/2}\phi\left(\sigma_t^{1/2}(y_t - \mathbf{X}_t'\boldsymbol{\beta})\right)dH(\sigma_t)}{\int_0^{\infty}\sigma_t^{1/2}\phi\left(\sigma_t^{1/2}(y_t - \mathbf{X}_t'\boldsymbol{\beta})\right)dH(\sigma_t)}$$

de modo que no caso contínuo, a densidade de σ_t dado y_t é dada por

$$k_{\boldsymbol{\theta}}(\sigma_t|y_t) = \frac{\sigma_t^{1/2}\phi\left(\sigma_t^{1/2}(y_t - \mathbf{X}_t'\boldsymbol{\beta})\right)h(\sigma_t|\boldsymbol{\zeta})}{\int_0^{\infty}\sigma_t^{1/2}\phi\left(\sigma_t^{1/2}(y_t - \mathbf{X}_t'\boldsymbol{\beta})\right)h(\sigma_t|\boldsymbol{\zeta})d\sigma_t}$$

Estimação dos Parâmetros no $(k+1)$-ésimo Passo

Tomando o gradiente de (3.9) com relação $\boldsymbol{\beta}$ e com relação a δ e igualando o resultado a zero, temos que no $(k+1)$-ésimo passo do algoritmo, a estimativa de $\boldsymbol{\beta}$ e δ. O resultado deste procedimento pode ser visto no algoritmo 3.1.1.

Algoritmo 3.1.1 Resumo do Algoritmo de Estimação via algoritmo EM para modelos lineares com ruídos distribuídos de acordo com uma mistura de normais.

Passo 1. obter uma estimativa inicial para β;

- um modo de se obter a estimativa inicial é através do método de estimação usual associado ao suavizador escolhido, ie, supondo o ruído i.i.d. e gaussiano.

Passo 2. obter uma estimativa inicial para o parâmetro de escala δ^2 e para ζ;

Passo 3. *loop* principal do algoritmo EM.

- Enquanto $|\text{EQM}_k - \text{EQM}_{k-1}| \geq \text{tol}$:

  Passo $(k+1)$**.1:** calcular

$$\beta^{(k+1)} = (\mathbf{X}'W(\theta^{(k)})\mathbf{X})^{-1}\mathbf{X}'W(\theta^{(k)})\mathbf{y};$$

  Passo $(k+1)$**.2:** calcular

$$(\delta^2)^{(k+1)} = \frac{1}{T}(\mathbf{y} - \mathbf{X}\beta^{(k+1)})'W(\theta^{(k+1)})(\mathbf{y} - \mathbf{X}\beta^{(k+1)});$$

  Passo $(k+1)$**.3:** no caso do vetor de parâmetros ζ, a estimativa $\zeta^{(k+1)}$ deve satisfazer a equação

$$\frac{\partial}{\partial\zeta}C(\zeta; \beta^{(k+1)}, \delta^{(k+1)}, \zeta^{(k)}) = 0.$$

Sobre a Matriz de Ponderação $W_T(\theta)$

Denotemos as diagonais de $W_T(\theta)$ por $\omega(y_t)$, ou seja, $\omega(y_t) = E_\theta[\sigma_t|y_t]$. O teorema a seguir lista algumas propriedades interessantes dos elementos da matriz de ponderação. Note que, de acordo com o teorema, as ponderações ω decrescem de acordo com o resíduo.

Teorema 3.1.1. *(Dempster et al., [14].) Suponha que Z seja distribuída de acordo com uma mistura de normais. Então, para $0 < |Z| < \infty$,*

I. $E[\sigma^k|Z] < \infty$ *para* $k > -\frac{1}{2}$;

II. $\omega'(Z) = -Z\text{Var}(\sigma|Z)$;

III. a função ω é simétrica, positiva e não-decrescente para $Z > 0$.

Para $Z = 0$,

i. $\omega(0) \geq \omega(Z)$, *para* $Z \neq 0$;

ii. $\omega(0) < \infty$ *se, e somente se,* $E\sigma^{3/2} < \infty$;

iii. $\omega'(0) < \infty$ *se, e somente se,* $E\sigma^{5/2} < \infty$.

Demonstração. Lembre que $p(z) = \int_0^\infty \sigma^{1/2}\phi(\sigma^{1/2}z)dH(\sigma)$, logo,

$$\omega(z) = E_{\boldsymbol{\theta}}[\sigma|z] = \frac{\int_0^\infty \sigma^{3/2}\phi(\sigma^{1/2}z)dH(\sigma)}{\int_0^\infty \sigma^{1/2}\phi(\sigma^{1/2}z)dH(\sigma)}$$

e, portanto,

$$\begin{aligned}\omega'(z) = & \frac{1}{\left[\int_0^\infty \sigma^{1/2}\phi(\sigma^{1/2}z)dH(\sigma)\right]^2} \times \\ & \times \left\{ -z\int_0^\infty \sigma^{5/2}\phi(\sigma^{1/2}z)dH(\sigma)\int_0^\infty \sigma^{1/2}\phi(\sigma^{1/2}z)dH(\sigma) \right. \\ & \left. -z\int_0^\infty \sigma^{3/2}\phi(\sigma^{1/2}z)dH(\sigma)\int_0^\infty \sigma^{3/2}\phi(\sigma^{1/2}z)dH(\sigma) \right\} \\ = & -z\{E[\sigma^2|z] - E[\sigma|z]^2\} = -z\text{Var}(\sigma|z) \leq 0\end{aligned}$$

demonstrando assim o item [II] e o fato que ω é uma função não-decrescente. Os demais pontos de [III] são triviais e seguem imediatamente da definição de p.

O restante das afirmações do teorema seguem do item [ii] da definição de ω e de argumentos análogos ao utilizado na propriedade 1.2.2 do capítulo 1. □

Proposição 3.1.1. *Defina* $M_i(\rho)$ *como o* $(i-1)$*-ésimo momento de* $1/\psi(\sigma)$*, isto é,*

$$M_i(\rho) \equiv \frac{1}{M_1(\rho)}\int_0^\infty \frac{1}{\psi(\sigma)^i}\phi\left(\frac{\rho}{\psi(\sigma)}\right)dH(\sigma),$$

para $i = 2, 3, \ldots$ *e*

$$M_1(\rho) \equiv \int_0^\infty \frac{1}{\psi(\sigma)}\phi\left(\frac{\rho}{\psi(\sigma)}\right)dH(\sigma).$$

Então, $M_2(\rho)$ *é decrescente em* ρ *se, e somente se,*

$$M_2(\rho) < \frac{M_4(\rho)}{M_3(\rho)}$$

3.1.3 Convergência do Algoritmo

Para garantir a convergência da seqüência $\boldsymbol{\theta}^{(k)}$ para um máximo local, é necessário determinar condições necessárias para que a seqüência $\delta^{(k)}$ seja limitada acima de zero, ie, que exista uma constante $\kappa > 0$ tal que $\delta^{(k)} > \kappa$, para todo k.

Proposição 3.1.2. *(Dempster et al., [14].) Suponha que*

i. $E\sigma^{1/2} < \infty$;

ii. no máximo $m < n$ das n equações $\mathbf{Y} = \mathbf{X}\beta$ podem ser simultaneamente satisfeitas para qualquer escolha de β;

iii. existe $a > n/(n-m)$ tal que

$$\lim_{|x|\to\infty} C_a(x) = C_a < \infty$$

onde $C_a(x) = p(x)|x|^a$.

Então, existe $\kappa > 0$ tal que a seqüência $\sigma^{(k)} > \kappa$, para todo k.

Demonstração. Suponha, por contradição, que a seqüência $(\sigma^{(k)})_k$ não seja limitada por baixo acima de zero. Então, existe uma subseqüência $(\sigma^{(k_j)})_j$ que converge para zero. Sabemos que a seqüência $(\beta^{(k)})_k$ é limitada, logo, ela admite uma subseqüência, $(\beta^{(k_j)})_j$, convergente, digamos, tal que $\beta^{(k_j)} \to \beta^*$, quando $j \to \infty$.

Agora, a função de verossimilhança associada às observações $\mathbf{y}$ e $\mathbf{x}$ calculada em $(\beta^{(k_j)}, \sigma^{(k_j)})$ é da forma

$$l(\beta^{(k_j)}, \sigma^{(k_j)}) = \frac{1}{(\sigma^{(k_j)})^T} \prod_{t=1}^{T} p\left(\frac{r_t^{(k_j)}}{\sigma^{(k_j)}}\right),$$

onde $r_t^{(k_j)} = y_t - \mathbf{X}_t\beta^{(k_j)}$.

Como $p(x)$ é limitada superiormente por $E\sigma^{1/2} = \frac{1}{\sqrt{2\pi}}p(0)$, podemos reordenar os resíduos $r_t^{(k_j)}$, para cada k_j, de modo que os m menores resíduos ocupem as m últimas posições do vetor de resíduos. Logo,

$$l(\beta^{(k_j)}, \sigma^{(k_j)}) \leq \frac{1}{(\sigma^{(k_j)})^T} \prod_{t=1}^{T-m} p\left(\frac{r_t^{(k_j)}}{\sigma^{(k_j)}}\right) [E\sigma^{1/2}]^m$$

sendo que a igualdade vale apenas se os m menores resíduos forem iguais a zero. Pela condição [ii], temos que $r_1^{(k_j)}, ..., r_{T-m}^{(k_j)}$ são positivos, logo,

$$\begin{aligned} l(\beta^{(k_j)}, \sigma^{(k_j)}) &\leq \frac{1}{(\sigma^{(k_j)})^T} \prod_{t=1}^{T-m} \frac{|\sigma^{(k_j)}|}{|r_t^{(k_j)}|} C_a\left(\frac{r_t^{(k_j)}}{\sigma^{(k_j)}}\right) [E\sigma^{1/2}]^m \\ &= \frac{1}{(\sigma^{(k_j)})^{T-a(T-m)}} [E\sigma^{1/2}]^m \prod_{t=1}^{T-m} \frac{1}{|r_t^{(k_j)}|} C_a\left(\frac{r_t^{(k_j)}}{\sigma^{(k_j)}}\right). \end{aligned}$$

Como $T - a(T-m) < 0$, segue que $(\sigma^{(k_j)})^{T-a(T-m)} \to 0$. Além disso, $r_t^{(k_j)} \to r^* \neq 0$ e, portanto, pela condição [iii],

$$\frac{1}{|r_t^{(k_j)}|} C_a\left(\frac{r_t^{(k_j)}}{\sigma^{(k_j)}}\right) \to C_a < \infty$$

pois, $\left|\frac{r_t^{(k_j)}}{\sigma^{(k_j)}}\right| \to \infty$. Logo, $l(\beta^{(k_j)}, \sigma^{(k_j)}) \to 0$, o que é uma contradição. □

3.1.4 Taxa de Convergência

Como visto na seção 2.2.2, o estudo da taxa de convergência para o algoritmo EM consiste basicamente no cálculo da matriz DM em $\boldsymbol{\theta}^*$. Sabe-se que

$$DM(\boldsymbol{\theta}^*) = D^{20}H(\boldsymbol{\theta}^*|\boldsymbol{\theta}^*)[D^{20}Q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^*)]^{-1}$$

onde $H(\boldsymbol{\theta}|\boldsymbol{\theta}^*) = E_{\boldsymbol{\theta}^*}\{\log k_{\boldsymbol{\theta}}(\boldsymbol{\sigma}|\mathbf{y})|\mathbf{y}\}$. Para expressar a matriz $DM(\boldsymbol{\theta}^*)$ de um modo mais compacto, defina $V(Z) = \text{Var}(\sigma|Z)$ e considere a seguinte notação

$$[\mathbf{Z}^4]_{\mathbf{V}} = \sum_{t=1}^{T} V_t Z_t^4$$

$$[\mathbf{Z}^3\mathbf{X}]_{\mathbf{V}} = \sum_{t=1}^{T} V_t Z_t^3 \mathbf{X}_t$$

$$[\mathbf{Z}^2\mathbf{X}'\mathbf{X}]_{\mathbf{V}} = \sum_{t=1}^{T} V_t Z_t^2 \mathbf{X}_t'\mathbf{X}_t.$$

Teorema 3.1.2. *(Dempster et al., [14].) Se $\boldsymbol{\theta}^{(k)}$ converge para $\boldsymbol{\theta}^*$ sob o algoritmo EM, então*

$$-D^{20}Q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^*) = \frac{1}{(\delta^2)^*}\begin{bmatrix} \mathbf{X}'\mathbf{W}^*\mathbf{X} & \mathbf{0} \\ \mathbf{0} & 2T \end{bmatrix}$$

$$-D^{20}H(\boldsymbol{\theta}^*|\boldsymbol{\theta}^*) = \frac{1}{(\delta^2)^*}\begin{bmatrix} (\mathbf{Z}^{*2}\mathbf{X}'\mathbf{X})_{\mathbf{V}^*} & (\mathbf{Z}^{*3}\mathbf{X})_{\mathbf{V}^*} \\ (\mathbf{Z}^{*3}\mathbf{X}')_{\mathbf{V}^*} & (\mathbf{Z}^{*4})_{\mathbf{V}^*} \end{bmatrix}$$

$$DM(\boldsymbol{\theta}^*) = \frac{1}{(\delta^2)^*}\begin{bmatrix} (\mathbf{Z}^{*2}\mathbf{X}'\mathbf{X})_{\mathbf{V}^*}(\mathbf{X}'\mathbf{W}^*\mathbf{X})^{-1} & \frac{1}{2T}(\mathbf{Z}^{*3}\mathbf{X})_{\mathbf{V}^*} \\ (\mathbf{Z}^{*3}\mathbf{X}')_{\mathbf{V}^*}(\mathbf{X}'\mathbf{W}^*\mathbf{X})^{-1} & \frac{1}{2T}(\mathbf{Z}^{*4})_{\mathbf{V}^*} \end{bmatrix}$$

e

$$D^2L(\boldsymbol{\theta}^*) = \frac{1}{(\delta^2)^*}\begin{bmatrix} \mathbf{X}'\mathbf{W}^*\mathbf{X} - (\mathbf{Z}^{*2}\mathbf{X}'\mathbf{X})_{\mathbf{V}^*} & -(\mathbf{Z}^{*3}\mathbf{X})_{\mathbf{V}^*} \\ -(\mathbf{Z}^{*3}\mathbf{X}')_{\mathbf{V}^*} & 2T - (\mathbf{Z}^{*4})_{\mathbf{V}^*} \end{bmatrix}$$

3.2 Estimação Linear aplicada ao Modelo Linear

Assuma agora que ao estimar os parâmetros do modelo (3.1), incorporemos um termo de penalização $-\lambda J(\boldsymbol{\theta})$. Ou seja, ao invés de buscarmos maximizar somente a log-verossimilhança, tomaremos como função objetiva

$$l(\boldsymbol{\theta}) - \lambda J(\boldsymbol{\theta}).$$

Assumindo que o erro segue uma mistura na escala de normais, temos, assim como na seção anterior, que

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)}) = C(\boldsymbol{\zeta};\boldsymbol{\theta}^{(k)}) - T\log\delta - \frac{1}{2\delta^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'W_T(\boldsymbol{\theta}^{(k)})(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}),$$

onde $C(\boldsymbol{\zeta};\boldsymbol{\theta}^{(k)})$ e $W_T(\boldsymbol{\theta}^{(k)})$ são como em (3.7) e (3.8). No entanto, como veremos abaixo, $\boldsymbol{\theta}^{(k+1)}$ não é mais o vetor de parâmetros que maximiza $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)})$, mas aquele que maximiza a versão penalizada desta função, $S(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)}) = Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)}) - \lambda J(\boldsymbol{\theta})$.

3.2.1 Estimação — Enfoque EM

As estimativas de $\boldsymbol{\beta}$, δ e $\boldsymbol{\zeta}$ no $(k+1)$-ésimo passo são, portanto, obtidas, de acordo com o enfoque usual do algoritmo EM, através da resolução das equações (possivelmente não-lineares):

$$\begin{aligned}
\frac{1}{\delta^2}\mathbf{X}'W_T(\boldsymbol{\theta}^{(k)})\mathbf{y} - \frac{1}{\delta^2}\mathbf{X}'W_T(\boldsymbol{\theta}^{(k)})\mathbf{X}\boldsymbol{\beta} - \lambda\frac{\partial}{\partial\boldsymbol{\beta}}J(\boldsymbol{\theta}) &= 0 \\
-\frac{T}{2\delta^2} + \frac{1}{2(\delta^2)^2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'W_T(\boldsymbol{\theta}^{(k)})(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) - \lambda\frac{\partial}{\partial\delta}J(\boldsymbol{\theta}) &= 0 \\
\frac{\partial}{\partial\boldsymbol{\zeta}}C(\boldsymbol{\zeta};\boldsymbol{\theta}^{(k)}) &= 0.
\end{aligned}$$

Considerando o caso particular em que J é dado por uma forma quadrática envolvendo apenas os coeficientes do modelo, temos o seguinte resultado

Proposição 3.2.1. *Assumindo $\boldsymbol{\zeta}$ conhecido e $J(\boldsymbol{\theta}) = \frac{1}{2}\boldsymbol{\beta}'A\boldsymbol{\beta}$, então, as estimativas de $\boldsymbol{\beta}$ e δ no $(k+1)$-ésimo passo do algoritmo EM satisfazem*

$$\boldsymbol{\beta}^{(k+1)} = (\mathbf{X}'W_T(\boldsymbol{\theta}^{(k)})\mathbf{X} + (\delta^2)^{(k+1)}\lambda A)^{-1}\mathbf{X}'W_T(\boldsymbol{\theta}^{(k)})\mathbf{y}$$

e

$$(\delta^2)^{(k+1)} = \frac{1}{T}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}^{(k+1)})'W_T(\boldsymbol{\theta}^{(k)})(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}^{(k+1)})$$

respectivamente.

Considerando-se o seguinte particionamento da matriz $K = D^2J(\boldsymbol{\theta}^*)$,

$$K \equiv \begin{bmatrix} K_{\boldsymbol{\beta}} & K_{\boldsymbol{\beta},\delta} \\ K'_{\boldsymbol{\beta},\delta} & K_\delta \end{bmatrix} = \begin{bmatrix} \frac{\partial^2}{\partial\boldsymbol{\beta}'\partial\boldsymbol{\beta}}J(\boldsymbol{\theta}^*) & \frac{\partial^2}{\partial\boldsymbol{\beta}'\partial\delta}J(\boldsymbol{\theta}^*) \\ \left[\frac{\partial^2}{\partial\boldsymbol{\beta}'\partial\delta}J(\boldsymbol{\theta}^*)\right]' & \frac{\partial^2}{\partial\delta^2}J(\boldsymbol{\theta}^*) \end{bmatrix}$$

temos o sequinte teorema

Teorema 3.2.1. *Se $\theta(k)$ converge para $\boldsymbol{\theta}^*$ sob o algoritmo EMP, então*

$$DM(\boldsymbol{\theta}^*) = \begin{bmatrix} \mathbf{X}'\mathbf{W}^*\mathbf{X} + \lambda(\delta^2)^*K_{\boldsymbol{\beta}} & \lambda(\delta^2)^*K'_{\boldsymbol{\beta},\delta} \\ \lambda(\delta^2)^*K'_{\boldsymbol{\beta},\delta} & 2T + \lambda(\delta^2)^*K_\delta \end{bmatrix}^{-1} \cdot \begin{bmatrix} (\mathbf{Z}^{*2}\mathbf{X}'\mathbf{X})_{\mathbf{V}^*} & (\mathbf{Z}^{*3}\mathbf{X})_{\mathbf{V}^*} \\ (\mathbf{Z}^{*3}\mathbf{X}')_{\mathbf{V}^*} & (\mathbf{Z}^{*4})_{\mathbf{V}^*} \end{bmatrix} \tag{3.10}$$

Demonstração. De acordo com o teorema 3.1.2, temos que

$$C = \frac{1}{(\delta^2)^*}\begin{bmatrix} (\mathbf{Z}^{*2}\mathbf{X}'\mathbf{X})_{\mathbf{V}^*} & (\mathbf{Z}^{*3}\mathbf{X})_{\mathbf{V}^*} \\ (\mathbf{Z}^{*3}\mathbf{X}')_{\mathbf{V}^*} & (\mathbf{Z}^{*4})_{\mathbf{V}^*} \end{bmatrix}$$

e que

$$B + C = \frac{1}{(\delta^2)^*}\begin{bmatrix} \mathbf{X}'\mathbf{W}^*\mathbf{X} & \mathbf{0} \\ \mathbf{0} & 2T \end{bmatrix}$$

de modo que, inserindo estes resultados na expressão (2.8), $DM(\boldsymbol{\theta}^*)$ é dado pela expressão (3.10). □

Corolário 3.2.2. *Se $J(\boldsymbol{\theta}) = \frac{1}{2}\boldsymbol{\beta}' A\boldsymbol{\beta}$ para alguma matriz positiva-definida A, então*

$$DM(\boldsymbol{\theta}^*) = \begin{bmatrix} (\mathbf{X}'\mathbf{W}^*\mathbf{X} + \lambda(\delta^2)^* A)^{-1}(\mathbf{Z}^{*2}\mathbf{X}'\mathbf{X})_{\mathbf{V}^*} & \frac{1}{2T}(\mathbf{Z}^{*3}\mathbf{X})_{\mathbf{V}^*} \\ (\mathbf{X}'\mathbf{W}^*\mathbf{X} + \lambda(\delta^2)^* A)^{-1}(\mathbf{Z}^{*3}\mathbf{X}')_{\mathbf{V}^*} & \frac{1}{2T}(\mathbf{Z}^{*4})_{\mathbf{V}^*} \end{bmatrix}. \tag{3.11}$$

3.2.2 Estimação — Enfoque OSL

De acordo com o método OSL, seção 2.2.3, a cada passo M dentro do algoritmo, devemos usar a estimativa obtida no passo anterior como argumento do termo de penalização J. Logo, o sistema de equações a serem resolvidas no $(k+1)$-ésimo passo é dado por

$$\begin{aligned} \frac{1}{\delta^2}\mathbf{X}'W_T(\boldsymbol{\theta}^{(k)})\mathbf{y} - \frac{1}{\delta^2}\mathbf{X}'W_T(\boldsymbol{\theta}^{(k)})\mathbf{X}\boldsymbol{\beta} - \lambda\frac{\partial}{\partial\boldsymbol{\beta}}J(\boldsymbol{\theta}^{(k)}) &= 0 \\ -\frac{T}{2\delta^2} + \frac{1}{2(\delta^2)^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'W_T(\boldsymbol{\theta}^{(k)})(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) - \lambda\frac{\partial}{\partial\delta}J(\boldsymbol{\theta}^{(k)}) &= 0 \\ \frac{\partial}{\partial\boldsymbol{\zeta}}C(\boldsymbol{\zeta};\boldsymbol{\theta}^{(k)}) &= 0. \end{aligned}$$

Considerando o caso particular em que J é dado por uma forma quadrática envolvendo apenas os coeficientes do modelo, temos o seguinte resultado

Proposição 3.2.2. *Assumindo $\boldsymbol{\zeta}$ conhecido e $J(\boldsymbol{\theta}) = \boldsymbol{\beta}' A\boldsymbol{\beta}$, então, as estimativas de $\boldsymbol{\beta}$ e δ no $(k+1)$-ésimo passo do algoritmo EM satisfazem*

$$\boldsymbol{\beta}^{(k+1)} = (\mathbf{X}'W_T(\boldsymbol{\theta}^{(k)})\mathbf{X})^{-1}(\mathbf{X}'W_T(\boldsymbol{\theta}^{(k)})\mathbf{y} - (\delta^2)^{(k+1)}\lambda A)$$

e

$$(\delta^2)^{(k+1)} = \frac{1}{T}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^{(k+1)})'W_T(\boldsymbol{\theta}^{(k)})(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^{(k+1)})$$

respectivamente.

Neste caso, a matriz que controla o comportamento assintótico do algoritmo difere daquela exibida no teorema 3.2.1.

Teorema 3.2.3. *Se $\boldsymbol{\theta}(k)$ converge para $\boldsymbol{\theta}^*$ sob o método OSL, então*

$$\begin{aligned} DN(\boldsymbol{\theta}^*) = &\begin{bmatrix} \mathbf{X}'\mathbf{W}^*\mathbf{X} & \mathbf{0} \\ \mathbf{0} & 2T \end{bmatrix}^{-1} \cdot \\ &\begin{bmatrix} (\mathbf{Z}^{*2}\mathbf{X}'\mathbf{X})_{\mathbf{V}^*} - \lambda(\delta^2)^* K_{\boldsymbol{\beta}} & (\mathbf{Z}^{*3}\mathbf{X})_{\mathbf{V}^*} - \lambda(\delta^2)^* K_{\boldsymbol{\beta},\delta} \\ (\mathbf{Z}^{*3}\mathbf{X}')_{\mathbf{V}^*} - \lambda(\delta^2)^* K'_{\boldsymbol{\beta},\delta} & (\mathbf{Z}^{*4})_{\mathbf{V}^*} - \lambda(\delta^2)^* K_\delta \end{bmatrix} \end{aligned} \tag{3.12}$$

3.3 Aplicação a Modelos Lineares de Séries Temporais

Nesta seção nos concentraremos em séries temporais autorregressivas, mais precisamente, nos modelos paramétricos lineares do tipo AR(p). Como antes, o objetivo é usar uma mistura na escala da distribuição

gaussiana para amortecer os efeitos de valores extremos. Além da autocorrelação, uma importante diferença entre estes modelos e o modelo de regressão linear estudados até agora é o fato de que, no caso de séries temporais, os efeitos dos valores extremos (explicados aqui por um ruído seguindo uma determinada distribuição com caudas pesadas) são sentidos tanto no "eixo" da variável resposta quanto no "eixo" da variável explicativa, pois, devido à estrutura do problema, o mesmo ruído afeta tanto a variável resposta quanto a variável explicativa em instantes diferentes e consecutivos. A figura 3.1 ilustra este fato ao comparar o resultado de duas simulações envolvendo a mesma função alvo, $\text{sen}(2x)$, mas com ruídos diferentes.

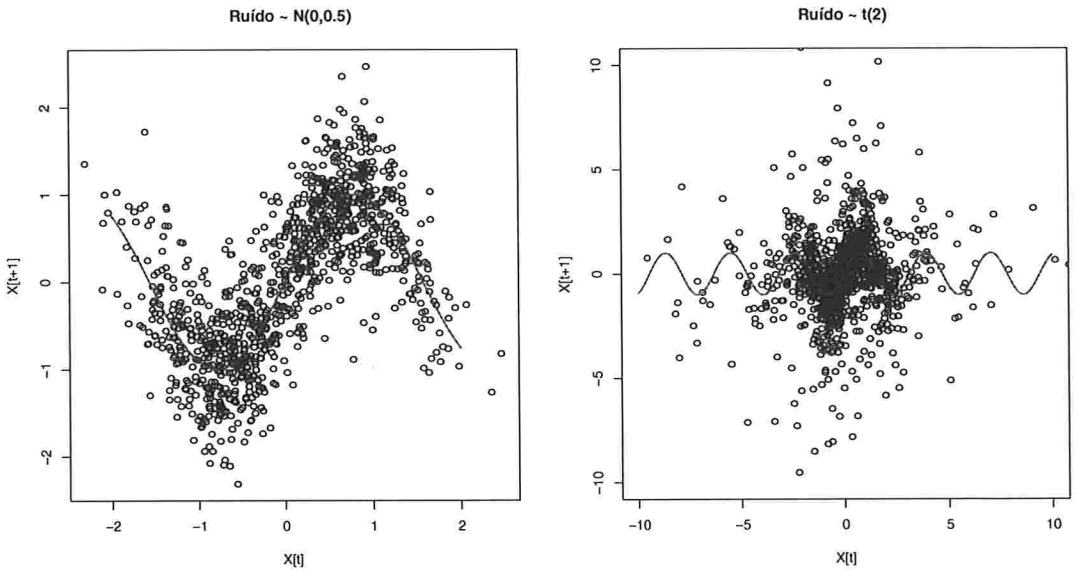


Figura 3.1: Comparativo do efeito da distribuição sobre os dados em modelos autorregressivos. Note que no caso em a distribuição do ruído tem caudas pesadas (imagem a direita) a dispersão dos dados ocorre em tanto com relação ao eixo vertical quanto com relação ao eixo horizontal.

3.3.1 Processos AR(1)

O objetivo desta seção é mostrar como estimar os parâmetros de um processo AR(1) quando os ruídos do modelo são distribuídos de acordo com uma mistura de gaussianas através do parâmetro de escala. Para tanto, seja

$$Y_t = c_0 + c_1 Y_{t-1} + \delta\epsilon_t$$

tal que a densidade de ϵ_t é dada por

$$p(\epsilon_t) = \int_0^\infty \frac{\sigma_t^{1/2}}{\delta}\phi\left(\frac{\sigma_t^{1/2}}{\delta}\epsilon_t\right) h(\sigma_t)d\sigma_t$$

e seja ζ o vetor de parâmetros associado à densidade p, ie, $h(\sigma_t) = h(\sigma_t|\zeta)$. Em particular, note que $Y_t|Y_{t-1} = y_{t-1} \sim SM(c_0 + c_1 y_{t-1}, \zeta)$. O objetivo principal aqui é estimar via máxima verossimilhança o vetor de parâmetros dado por

$$\boldsymbol{\theta} = (c_0, c_1, \zeta')'.$$

Como no caso do modelo de regressão, as variáveis σ_t's não são observadas e, portanto, devem ser consideradas como variáveis latentes do modelo. Além disso, embora seja possível tratar y_1 como a realização da variável aleatória Y_1 e determinar a densidade $p_{\boldsymbol{\theta}^{(k)}}(y_1|\sigma_1)$, como veremos mais adiante, é mais conveniente tratar Y_1 como determinística e considerar o estimador de máxima verossimilhança condicional a $Y_1 = y_1$. Deste modo o conjunto completo de variáveis é dado por

$$(y_2, ..., y_T; \sigma_2, ..., \sigma_T)'.$$

Para estimarmos os parâmetros associados ao modelo, modelaremos o ruído de acordo com

$$\epsilon_t|\sigma_t \sim \mathcal{N}\left(0, \frac{\delta^2}{\sigma_t}\right)$$

de modo que,

$$\begin{aligned} Y_t|y_{t-1}, \sigma_t &\sim \mathcal{N}\left(c_0 - c_1 y_{t-1}, \frac{\delta^2}{\sigma_t}\right), \\ \sigma_t &\sim h \end{aligned}$$

para $t = 2, ...T$, então, a distribuição marginal de $Y_t|y_{t-1}$ é a desejada, ie, $Y_t|Y_{t-1} = y_{t-1} \sim SM(c_0 + c_1 y_{t-1}, \zeta)$.

A densidade conjunta de $(Y_2, ..., Y_T; \sigma_2, ..., \sigma_T)'$ é dada por

$$\begin{aligned} p(y_2, ..., y_T; \sigma_2, ..., \sigma_T|y_1) = &\, p(y_T|y_{T-1}, ..., y_1; \sigma_2, ..., \sigma_T) \\ &\cdot p(y_{T-1}|y_{T-2}, ..., y_1; \sigma_2, ..., \sigma_T) \cdots p(y_2|y_1; \sigma_2, ..., \sigma_T) \\ &\cdot p(\sigma_2, ..., \sigma_T) \end{aligned}$$

ou seja,

$$p(y_2, ..., y_T; \sigma_2, ..., \sigma_T) = p(y_T|y_{T-1}; \sigma_T) \cdots p(y_2|y_1; \sigma_2) \cdot \prod_{t=2}^{T} p(\sigma_t).$$

Logo, a log-verossimilhança de $(Y_2, ..., Y_T; \sigma_2, ..., \sigma_T)'$ é dada por

$$\log p(y_2, ..., y_T; \sigma_1, ..., \sigma_T) = \sum_{t=2}^{T} \log p(y_t|y_{t-1}; \sigma_t) + \sum_{t=2}^{T} \log p(\sigma_t).$$

Maximização via Algoritmo EM

Seja $\theta^{(k)}$ a estimativa de θ obtida no j-ésimo passo do algoritmo EM e defina

$$\begin{aligned} Q(\theta|\theta^{(k)}) &= E\{\log p(y_2, ..., y_T; \sigma_2, ..., \sigma_T|\theta)|\mathbf{y}; \theta^{(k)}\} \\ &= \sum_{t=2}^{T} E\{\log p_\theta(y_t|y_{t-1}; \sigma_t)|\mathbf{y}; \theta^{(k)}\} + \sum_{t=2}^{T} E\{\log p_\theta(\sigma_t)|\mathbf{y}; \theta^{(k)}\} \end{aligned}$$

onde as esperanças acima são tomadas com relação à distribuição definida pela densidade

$$\begin{aligned} k(\sigma_2, ..., \sigma_T|\mathbf{y}; \theta^{(k)}) &= \frac{p_{\theta^{(k)}}(\mathbf{y}; \sigma)}{p(\mathbf{y})} \\ &\propto p_{\theta^{(k)}}(\mathbf{y}|\sigma) p_{\theta^{(k)}}(\sigma) \\ &= \prod_{t=2}^{T} p_{\theta^{(k)}}(y_t|y_{t-1}, \sigma_t) \cdot \prod_{t=2}^{T} p_{\theta^{(k)}}(\sigma_t). \end{aligned}$$

Note que, condicionadas no vetor de observações $\mathbf{y}$, as variáveis σ_2,...,σ_T são independentes, cada uma delas com densidade

$$p_{\theta^{(k)}}(y_t|y_{t-1}; \sigma_t) \cdot p_{\theta^{(k)}}(\sigma_t) = \frac{\sigma_t^{1/2}}{\delta} \phi \left(\frac{\sigma_t^{1/2}}{\delta} (c_{0,j} + c_{1,j} y_{t-1}) \right) \cdot p_{\theta^{(k)}}(\sigma_t) \tag{3.13}$$

para $t = 2, ..., T$ e onde ϕ é a densidade associada a distribuição $\mathcal{N}(0, 1)$ e $c_{0,j}$ e $c_{1,j}$ são as estimativas de c_0 e c_1 obtidas no j-ésimo passo do algoritmo EM. Em particular, isto implica no fato que podemos gerar as amostras pseudo aleatórias das variáveis latentes σ_2,...,σ_T separadamente e de acordo com as densidades em (3.13).

3.3.2 Processos AR(p)

Considere agora o modelo auto-regressivo de ordem p,

$$Y_t = c_p Y_{t-p} + ... + c_1 Y_{t-1} + c_0 + \delta \epsilon_t \tag{3.14}$$

e assuma que $\epsilon_t \sim SM(0, \delta; h)$, de modo que $Y_t|y_{t-1}, ..., y_{t-p} \sim SM(c_0 + c_1 y_{t-1} + ... + c_p y_{t-p}, \delta; h)$. Como antes, nosso objetivo é estimar o vetor θ de parâmetros associado ao modelo (3.14) e dado por

$$\theta = (c_p, ..., c_1, c_0, \delta, \zeta')' = (\mathbf{c}', \delta, \zeta')',$$

onde ζ é um eventual vetor de parâmetros associado a h.

Pelos mesmos motivos pelos quais consideramos Y_1 determinístico ao analisar o modelo AR(1) na seção anterior, assumiremos agora que as p primeiras variáveis Y_1,...,Y_p são determinísticas. Ou seja, os valores

y_1,...,y_p não serão considerados como realizações de variáveis aleatórias Y_1,...,Y_p. Deste modo, a densidade conjunta de $(Y_{p+1}, ..., Y_T, \sigma_{p+1}, ..., \sigma_T)'$ é dada por

$$\begin{aligned} p(y_{p+1}, ..., y_T, \sigma_{p+1}, ..., \sigma_T | y_p, ..., y_1) &= p(y_T | y_{T-1}, ..., y_{T-p}; \sigma_T) \cdots \\ &\quad \cdots p(y_{p+1} | y_p, ..., y_1; \sigma_{p+1}) p(\sigma_T) \cdots p(\sigma_{p+1}) \\ &= \prod_{t=p+1}^{T} p(y_t | y_{t-1}, ..., y_{t-p}; \sigma_t) p(\sigma_t) \end{aligned}$$

e, conseqüentemente, a log-verossimilhança por

$$l_c(\boldsymbol{\theta}) = \sum_{t=p+1}^{T} \log p(y_t | y_{t-1}, ..., y_{t-p}; \sigma_t) + \sum_{t=p+1}^{T} \log p(\sigma_t).$$

Como antes, modelaremos os dados por

$$\begin{aligned} Y_t | y_{t-1}, ..., y_{t-p}, \sigma_t &\sim \mathcal{N}\left(c_p y_{t-p} + ... + c_1 y_{t-1} + c_0, \frac{\delta^2}{\sigma_t}\right), \\ \sigma_t &\sim h \end{aligned}$$

para $t = p + 1, ...T$, de modo que o procedimento de estimação se dá de modo análogo ao proposto para processos AR(1), com a diferença de que, agora, a densidade utilizada é

$$\begin{aligned} k(\sigma_2, ..., \sigma_T | \mathbf{y}; \boldsymbol{\theta}^{(k)}) &= \frac{p_{\boldsymbol{\theta}^{(k)}}(\mathbf{y}; \sigma)}{p(\mathbf{y})} \\ &\propto p_{\boldsymbol{\theta}^{(k)}}(\mathbf{y} | \sigma) p_{\boldsymbol{\theta}^{(k)}}(\sigma) \\ &= \prod_{t=p+1}^{T} p_{\boldsymbol{\theta}^{(k)}}(y_t | y_{t-1}, ..., y_{t-p}, \sigma_t) \cdot \prod_{t=p+1}^{T} p_{\boldsymbol{\theta}^{(k)}}(\sigma_t), \end{aligned}$$

onde

$$p_{\boldsymbol{\theta}^{(k)}}(y_t | y_{t-1}, ..., y_{t-p}) = \frac{\sigma_t^{1/2}}{\delta} \phi\left(\frac{\sigma_t^{1/2}}{\delta}(y_t - c_p y_{t-p} - ... - c_1 y_{t-1} - c_0)\right)$$

e $p_{\boldsymbol{\theta}^{(k)}}(\sigma_t) = h_{\boldsymbol{\zeta}_j}(\sigma_t)$. Além disso, evidentemente, a função objetiva passa a ser da forma

$$\begin{aligned} Q(\boldsymbol{\theta} | \boldsymbol{\theta}^{(k)}) &= E\{\log p_{\boldsymbol{\theta}}(y_{p+1}, ..., y_T; \sigma_{p+1}, ..., \sigma_T) | \mathbf{y}; \boldsymbol{\theta}^{(k)}\} \\ &= \sum_{t=p+1}^{T} E\{\log p_{\boldsymbol{\theta}}(y_t | y_{t-1}, ..., y_{t-p}; \sigma_t) | \mathbf{y}; \boldsymbol{\theta}^{(k)}\} \\ &\quad + \sum_{t=p+1}^{T} E\{\log p_{\boldsymbol{\theta}}(\sigma_t) | \mathbf{y}; \boldsymbol{\theta}^{(k)}\}. \end{aligned}$$

Substituindo as densidades condicionais acima pelas suas expressões algébricas, obtemos

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)}) = -\frac{1}{2\delta^2}(\mathbf{y} - \mathbf{Yc})'W(\boldsymbol{\theta}^{(k)})(\mathbf{y} - \mathbf{Yc}) - \frac{T}{2}\log\delta^2 + \sum_{t=p+1}^{T} E_{\boldsymbol{\theta}^{(k)}}\{\log h(\sigma_t|\boldsymbol{\zeta})|\mathbf{y}\},$$

onde $\mathbf{y} = (y_{p+1}, ..., y_T)'$

$$\mathbf{Y} = \begin{bmatrix} y_p & \cdots & y_1 \\ y_{p+1} & \cdots & y_2 \\ \vdots & & \vdots \\ y_{T-1} & \cdots & y_{T-p} \end{bmatrix}$$

e $W(\boldsymbol{\theta}^{(k)}) = \text{diag}(E_{\boldsymbol{\theta}^{(k)}}\{\sigma_{p+1}|\mathbf{y}\}, ..., E_{\boldsymbol{\theta}^{(k)}}\{\sigma_T|\mathbf{y}\})'$.

Os Estimadores no $(k+1)$-ésimo Passo do Algoritmo

Como nos modelos de regressão, as estimativas de $\mathbf{c}$, δ^2 e $\boldsymbol{\zeta}$ são obtidas derivando-se $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)})$ e igualando o resultado a zero. O resultado desta operação pode ser visto no algoritmo 3.3.1.

3.3.3 Desvios Padrões dos Estimadores e Consistência

Consistência

Para demonstrar a consistência dos estimadores, representaremos o modelo AR(p) como um processo vetorial autorregressivo de ordem 1. Para tanto, defina $\mathbf{y}_t = (y_t, y_{t-1}, ..., y_{t-p+1})'$. Note que, assim,

$$\mathbf{Y} = \begin{bmatrix} \mathbf{y}_p' \\ \vdots \\ \mathbf{y}_{T-1}' \end{bmatrix}.$$

Defina também

$$\mathbf{B} = \begin{bmatrix} c_p & c_{p-1} & c_{p-2} & \cdots & c_2 & c_1 \\ 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 1 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & 0 \end{bmatrix}$$

e $\mathbf{u}_t = (\epsilon_t, 0, ..., 0)'$. Note que de acordo com esta notação, o modelo genérico AR(p) pode ser escrito como um modelo autorregressivo de ordem 1,

$$\mathbf{y}_t = \mathbf{B}\mathbf{y}_{t-1} + \mathbf{u}_t. \tag{3.15}$$

O resultado a seguir mostra como ficam as equações de estimação para o modelo (3.14) segundo esta notação.

Algoritmo 3.3.1 Resumo do Algoritmo de Estimação via algoritmo EM para modelos autorregressivos lineares com ruídos distribuídos de acordo com uma mistura de normais.

Dados: $y_1,...,y_p$.

Passo 1. obter uma estimativa inicial para $\mathbf{c}$;

Passo 2. obter uma estimativa inicial para o parâmetro de escala δ^2 e para $\boldsymbol{\zeta}$;

Passo 3. *loop* principal do algoritmo EM.

- Enquanto $|\text{EQM}_k - \text{EQM}_{k-1}| \geq \text{tol}$:

  Passo $(k+1)$**.1:** calcular

$$\mathbf{c}^{(k+1)} = (\mathbf{Y}'W(\boldsymbol{\theta}^{(k)})\mathbf{Y})^{-1}\mathbf{Y}'W(\boldsymbol{\theta}^{(k)})\mathbf{y};$$

  Passo $(k+1)$**.2:** calcular

$$(\delta^2)^{(k+1)} = \frac{1}{T}(\mathbf{y} - \mathbf{Y}\mathbf{c}^{(k+1)})'W(\boldsymbol{\theta}^{(k+1)})(\mathbf{y} - \mathbf{Y}\mathbf{c}^{(k+1)});$$

  Passo $(k+1)$**.3:** no caso do vetor de parâmetros $\boldsymbol{\zeta}$, a estimativa $\boldsymbol{\zeta}^{(k+1)}$ deve satisfazer a equação

$$\frac{\partial}{\partial \boldsymbol{\zeta}} C(\boldsymbol{\zeta}; \boldsymbol{\beta}^{(k+1)}, \delta^{(k+1)}, \boldsymbol{\zeta}^{(k)}) = 0.$$

Proposição 3.3.1. *Se o processo* $\{y_t\}$ *satisfaz o modelo* (3.14)*, com* $c_0 = 0$*, então as equações de estimação obtidas via máxima verossimilhança para* $(c_1, ..., c_p; \delta)$ *são dadas por*

$$\frac{1}{T}\left(\sum_{t=p+1}^{T} \widehat{\omega}_t \mathbf{y}_{t-1}\mathbf{y}_{t-1}'\right)\widehat{\mathbf{B}}' = \frac{1}{T}\sum_{t=p+1}^{T} \widehat{\omega}_t \mathbf{y}_{t-1}\mathbf{y}_t' \tag{3.16}$$

$$\widehat{\Delta} = \sum_{t=p+1}^{T} \widehat{\omega}_t(\mathbf{y}_t - \widehat{\mathbf{B}}\mathbf{y}_{t-1})(\mathbf{y}_t - \widehat{\mathbf{B}}\mathbf{y}_{t-1})' \tag{3.17}$$

onde

$$\Delta = \begin{bmatrix} \delta^2 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & 0 \end{bmatrix}$$

e $\widehat{\omega}_t = E_{\widehat{\mathbf{c}},\widehat{\delta}}\{\sigma_t\}$*, para* $t = p+1, ..., T$.

Demonstração. Em primeiro lugar, note que $(\mathbf{Y}'\widehat{W}\mathbf{Y})\widehat{c} = \mathbf{Y}'\widehat{W}\mathbf{y}$, onde $\widehat{W} = \text{diag}(\widehat{\omega}_{p+1}, ..., \widehat{\omega}_T)$ e, em

segundo lugar, que

$$\mathbf{Y}'\widehat{W}\mathbf{Y} = [\mathbf{y}_p \cdots \mathbf{y}_{T-1}]\widehat{W}\begin{bmatrix} \mathbf{y}_p' \\ \vdots \\ \mathbf{y}_{T-1}' \end{bmatrix} = \sum_{t=p}^{T-1} \widehat{\omega}_{t+1}\mathbf{y}_t\mathbf{y}_t'$$

e

$$\mathbf{Y}'\widehat{W}\mathbf{y} = [\mathbf{y}_p \cdots \mathbf{y}_{T-1}]\widehat{W}\begin{bmatrix} y_{p+1} \\ \vdots \\ y_T \end{bmatrix} = \sum_{t=p}^{T-1} \widehat{\omega}_{t+1}y_{t+1}\mathbf{y}_t.$$

Logo,

$$\frac{1}{T}\left(\sum_{t=p}^{T-1} \widehat{\omega}_{t+1}\mathbf{y}_t\mathbf{y}_t'\right)\widehat{\mathbf{c}} = \frac{1}{T}\sum_{t=p}^{T-1} \widehat{\omega}_{t+1}y_{t+1}\mathbf{y}_t.$$

Observe que $\widehat{\mathbf{c}} = \mathbf{B}_1'$ onde $\mathbf{B}_j'$ é a j-ésima linha de $\mathbf{B}$. Então, analogamente, podemos, de modo mais geral, escrever para uma matriz genérica $\mathbf{B}$,

$$\frac{1}{T}\left(\sum_{t=p}^{T-1} \widehat{\omega}_{t+1}\mathbf{y}_t\mathbf{y}_t'\right)\mathbf{B}' = \frac{1}{T}\sum_{t=p}^{T-1} \widehat{\omega}_{t+1}\mathbf{y}_t\mathbf{y}_{t+1}',$$

de onde sai (3.16).

Para o fator de escala, temos

$$\begin{aligned}\widehat{\delta}^2 &= \frac{1}{T}(\mathbf{y} - \mathbf{Y}\widehat{\mathbf{c}})'\widehat{W}(\mathbf{y} - \mathbf{Y}\widehat{\mathbf{c}}) \\ &= \frac{1}{T}\sum_{t=p+1}^{T} \widehat{\omega}_t(y_t - \widehat{c}_p y_{t-p} - \ldots - \widehat{c}_1 y_{t-1})^2 \\ &= \frac{1}{T}\sum_{t=p+1}^{T} \widehat{\omega}_t(y_t - \widehat{\mathbf{c}}'\mathbf{y}_{t-1})^2.\end{aligned}$$

Logo, observando que

$$(\mathbf{y}_t - \mathbf{B}\mathbf{y}_{t-1})(\mathbf{y}_t - \mathbf{B}\mathbf{y}_{t-1})' = \begin{bmatrix} (y_t - c_p y_{t-p} - \ldots - c_1 y_{t-1})^2 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix},$$

temos (3.17). □

O resultado a seguir garante que as propriedades assintóticas para uma série começando em um dado instante de tempo são idênticas às de uma série definida para todo $t \in \mathbb{Z}$. Para obtermos o resultado, necessitaremos do seguinte lema.

Lema 3.3.1. *Se λ é maior do que o valor absoluto máximo dos autovalores de $\mathbf{B}$, então, existe uma constante c tal que o valor absoluto de uma das componentes de $\mathbf{B}^s$ é menor do que $c\lambda^s$, para todo $s \geq 0$.*

Demonstração. Veja [4], página 191, lema 5.5.1. □

Teorema 3.3.2. *Assuma que* $\{\mathbf{y}_t\}$ *é um processo estocástico satisfazendo* (3.15) *para* $t = p+1, ..., T$, $\{\mathbf{y}_t^*\}$ *é um processo estocástico satisfazendo* (3.15) *para* $t \in \mathbb{Z}$, $\{\mathbf{u}_t\}$ *é uma seqüência de variáveis aleatórias i.i.d. com* $E\mathbf{u}_t = \mathbf{0}$ *e* $E\mathbf{u}_t\mathbf{u}_t' = \Delta$. *Então, se* $\mathbf{B}$ *tem apenas autovalores com valor absoluto menor do que 1 e se* $\{\widehat{\omega}_t\}$ *é uma seqüência de pesos limitada por* $\omega(0) < \infty$ *tal que* $\widehat{\omega}_t = \widehat{\omega}_t(y_1, ..., y_T)$, *segue que*

$$\frac{1}{\sqrt{T}}\left(\sum_{t=p+1}^{T}\widehat{\omega}_t\mathbf{y}_{t-1}^*(\mathbf{y}_{t-1}^*)' - \sum_{t=p+1}^{T}\widehat{\omega}_t\mathbf{y}_{t-1}\mathbf{y}_{t-1}'\right) \xrightarrow{P} 0,$$

$$\frac{1}{\sqrt{T}}\left(\sum_{t=p+1}^{T}\widehat{\omega}_t\mathbf{y}_{t-1}^*(\mathbf{y}_t^*)' - \sum_{t=p+1}^{T}\widehat{\omega}_t\mathbf{y}_{t-1}\mathbf{y}_t'\right) \xrightarrow{P} 0,$$

$$\frac{1}{\sqrt{T}}\left(\sum_{t=p+1}^{T}\widehat{\omega}_t\mathbf{y}_t^*(\mathbf{y}_t^*)' - \sum_{t=p+1}^{T}\widehat{\omega}_t\mathbf{y}_t\mathbf{y}_t'\right) \xrightarrow{P} 0,$$

quando $T \to \infty$.

Demonstração. Note que

$$\mathbf{y}_t^* - \mathbf{y}_t = \mathbf{B}^{t-p}(\mathbf{y}_p^* - \mathbf{y}_p)$$

onde

$$\mathbf{y}_p^* = \sum_{s=0}^{\infty}\mathbf{B}^s\mathbf{u}_{p-s}.$$

Logo,

$$\begin{aligned}\sum_{t=p+1}^{T}\widehat{\omega}_t\mathbf{y}_t^*(\mathbf{y}_t^*)' - \sum_{t=p+1}^{T}\widehat{\omega}_t\mathbf{y}_t\mathbf{y}_t' &= \sum_{t=p+1}^{T}\widehat{\omega}_t\mathbf{B}^{t-p}(\mathbf{y}_p^* - \mathbf{y}_p)(\mathbf{y}_t^*)' + \sum_{t=p+1}^{T}\widehat{\omega}_t\mathbf{y}_t(\mathbf{y}_t^*)' \\ &\quad + \sum_{t=p+1}^{T}\widehat{\omega}_t\mathbf{y}_t[\mathbf{B}^{t-p}(\mathbf{y}_p^* - \mathbf{y}_p)]' - \sum_{t=p+1}^{T}\widehat{\omega}_t\mathbf{y}_t(\mathbf{y}_t^*)' \\ &= \sum_{t=p+1}^{T}\widehat{\omega}_t\mathbf{B}^{t-p}(\mathbf{y}_p^* - \mathbf{y}_p)(\mathbf{y}_t^*)' + \sum_{t=p+1}^{T}\widehat{\omega}_t\mathbf{y}_t(\mathbf{y}_p^* - \mathbf{y}_p)'(\mathbf{B}')^{t-p},\end{aligned}$$

ou seja,

$$\begin{aligned}\sum_{t=p+1}^{T}\widehat{\omega}_t\mathbf{y}_t^*(\mathbf{y}_t^*)' - \sum_{t=p+1}^{T}\widehat{\omega}_t\mathbf{y}_t\mathbf{y}_t' &= \sum_{t=p+1}^{T}\widehat{\omega}_t\mathbf{B}^{t-p}(\mathbf{y}_p^* - \mathbf{y}_p)(\mathbf{y}_p^* - \mathbf{y}_p)'(\mathbf{B}')^{t-p} \\ &\quad + \sum_{t=p+1}^{T}\widehat{\omega}_t\mathbf{B}^{t-p}(\mathbf{y}_p^* - \mathbf{y}_p)\mathbf{y}_t' + \sum_{t=p+1}^{T}\widehat{\omega}_t\mathbf{y}_t(\mathbf{y}_p^* - \mathbf{y}_p)'(\mathbf{B}')^{t-p}.\end{aligned}$$

Note agora que a soma dos valores esperados dos quadrados das componentes da matriz

$$\sum_{t=p+1}^{T} \widehat{\omega}_t \mathbf{B}^{t-p}(\mathbf{y}_p^* - \mathbf{y}_p)\mathbf{y}_t'$$

é dada por

$$\begin{aligned}
&\mathrm{tr}E \sum_{t=p+1}^{T} \widehat{\omega}_t \mathbf{B}^{t-p}(\mathbf{y}_p^* - \mathbf{y}_p)\mathbf{y}_t' \left[\sum_{s=p+1}^{T} \widehat{\omega}_s \mathbf{B}^{s-p}(\mathbf{y}_p^* - \mathbf{y}_p)\mathbf{y}_s'\right]' \\
&\quad = \mathrm{tr}E \sum_{t,s=p+1}^{T} \widehat{\omega}_t \widehat{\omega}_s \mathbf{B}^{t-p}(\mathbf{y}_p^* - \mathbf{y}_p)\mathbf{y}_t'\mathbf{y}_s(\mathbf{y}_p^* - \mathbf{y}_p)'(\mathbf{B}')^{s-p} \\
&\quad = \mathrm{tr}E \sum_{t,s=p+1}^{T} \mathbf{B}^{t-p}[\mathbf{y}_p^*(\mathbf{y}_p^*)' - \mathbf{y}_p^*\mathbf{y}_p' - \mathbf{y}_p(\mathbf{y}_p^*)' + \mathbf{y}_p\mathbf{y}_p'](\mathbf{B}')^{s-p}\widehat{\omega}_t\widehat{\omega}_s\mathbf{y}_t'\mathbf{y}_s \\
&\quad = \sum_{t,s=p+1}^{T} \mathrm{tr}\mathbf{B}^{t-p}(F + \mathbf{y}_p\mathbf{y}_p')(\mathbf{B}')^{s-p}E\widehat{\omega}_t\widehat{\omega}_s\mathbf{y}_t'\mathbf{y}_s,
\end{aligned} \tag{3.18}$$

onde $F = E\mathbf{y}_p^*(\mathbf{y}_p^*)'$. Usando a desigualdade de Cauchy-Schwarz,

$$|E\widehat{\omega}_t\widehat{\omega}_s\mathbf{y}_t'\mathbf{y}_s| \leq (E\widehat{\omega}_t^2\mathbf{y}_t'\mathbf{y}_t)^{1/2}(E\widehat{\omega}_s^2\mathbf{y}_s'\mathbf{y}_s)^{1/2} \leq \omega(0)(E\mathbf{y}_t'\mathbf{y}_t)^{1/2}(E\mathbf{y}_s'\mathbf{y}_s)^{1/2}.$$

Agora,

$$\begin{aligned}
E\mathbf{y}_t\mathbf{y}_t' &= E\left[\sum_{s=0}^{t-p-1} \mathbf{B}^s\mathbf{u}_{t-s} + \mathbf{B}^{t-p}\mathbf{y}_p\right]\left[\sum_{s=0}^{t-p-1} \mathbf{B}^s\mathbf{u}_{t-s} + \mathbf{B}^{t-p}\mathbf{y}_p\right]' \\
&= E\sum_{s=0}^{t-p-1} \mathbf{B}^s\mathbf{u}_{t-s}\mathbf{u}_{t-s}'(\mathbf{B}')^s + E\mathbf{B}^{t-p}\mathbf{y}_p\mathbf{y}_p'(\mathbf{B}')^{t-p} \\
&= \sum_{s=0}^{t-p-1} \mathbf{B}^s\Delta\mathbf{B}^s + \mathbf{B}^{t-p}\mathbf{y}_p\mathbf{y}_p'(\mathbf{B}')^{t-p}.
\end{aligned}$$

Portanto,

$$E\mathbf{y}_t'\mathbf{y}_t = \mathrm{tr}EE\mathbf{y}_t\mathbf{y}_t' \leq \underbrace{\sum_{s=0}^{t-p-1} c^2p^3\lambda^{2s}\max|\delta_{ij}| + c^2p^2\lambda^{2(t-p)}\mathbf{y}_p'\mathbf{y}_p}_{\equiv\kappa_t}.$$

Logo,

$$|E\widehat{\omega}_t\widehat{\omega}_s\mathbf{y}_t'\mathbf{y}_s| \leq \omega(0)\kappa_t\kappa_s.$$

Para terminar de avaliar (3.18), devemos considerar

$$\mathrm{tr}\mathbf{B}^{t-p}(F + \mathbf{y}_p\mathbf{y}_p')(\mathbf{B}')^{s-p},$$

o qual é menor ou igual ao produto entre c^2, p^3, o máximo entre os valores absolutos das componentes de $F + \mathbf{y}_p\mathbf{y}_p'$ e λ^{t+s-2p}. Portanto,

$$\begin{aligned}\sum_{t,s=p+1}^{T} \mathrm{tr}\mathbf{B}^{t-p}(F + \mathbf{y}_p\mathbf{y}_p')(\mathbf{B}')^{s-p} &\leq c^2p^3 \sum_{t,s=p+1}^{T} \lambda^{t+s-2p} \\ &\leq \frac{c^2p^3}{\lambda^{2p}} \sum_{t,s=0}^{T} \lambda^t\lambda^s \\ &= \frac{c^2p^3}{\lambda^{2p}} \left(\sum_{t=0}^{T} \lambda^t\right)^2\end{aligned}$$

o que resulta em

$$\sum_{t,s=p+1}^{T} \mathrm{tr}\mathbf{B}^{t-p}(F + \mathbf{y}_p\mathbf{y}_p')(\mathbf{B}')^{s-p} \leq \frac{c^2p^3}{\lambda^{2p}}\lambda^2 \left(\frac{1-\lambda^T}{1-\lambda}\right)^2 \leq \frac{c^2p^3}{\lambda^{2p-2}} \frac{1}{(1-\lambda)^2}.$$

Então, pela desigualdade de Chebyshev,

$$\frac{1}{\sqrt{T}} \sum_{t=p+1}^{T} \widehat{\omega}_t\mathbf{B}^{t-p}(\mathbf{y}_p^* - \mathbf{y}_p)\mathbf{y}_t' \xrightarrow{P} 0.$$

Obviamente, resulta daí que

$$\frac{1}{\sqrt{T}} \sum_{t=p+1}^{T} \widehat{\omega}_t\mathbf{y}_t(\mathbf{y}_p^* - \mathbf{y}_p)'(\mathbf{B})'^{t-p} \xrightarrow{P} 0.$$

Logo, para mostrar que

$$\frac{1}{\sqrt{T}} \left(\sum_{t=p+1}^{T} \widehat{\omega}_t\mathbf{y}_t^*(\mathbf{y}_t^*)' - \sum_{t=p+1}^{T} \widehat{\omega}_t\mathbf{y}_t\mathbf{y}_t'\right) \xrightarrow{P} 0, \tag{3.19}$$

resta apenas checar que

$$\frac{1}{\sqrt{T}} \sum_{t=p+1}^{T} \widehat{\omega}_t\mathbf{B}^{t-p}(\mathbf{y}_p^* - \mathbf{y}_p)(\mathbf{y}_p^* - \mathbf{y}_p)'(\mathbf{B}')^{t-p} \xrightarrow{P} 0. \tag{3.20}$$

Para tanto,

$$\begin{aligned}&\mathrm{tr}E \sum_{t=p+1}^{T} \widehat{\omega}_t\mathbf{B}^{t-p}(\mathbf{y}_p^* - \mathbf{y}_p)(\mathbf{y}_p^* - \mathbf{y}_p)'(\mathbf{B}')^{t-p} \\ &\quad = \mathrm{tr} \sum_{t=p+1}^{T} \mathbf{B}^{t-p}(F + \mathbf{y}_p\mathbf{y}_p')(\mathbf{B}')^{t-p}E\widehat{\omega}_t \\ &\quad \leq \omega(0) \sum_{t=p+1}^{T} \mathrm{tr}\mathbf{B}^{t-p}(F + \mathbf{y}_p\mathbf{y}_p')(\mathbf{B}')^{t-p}E\widehat{\omega}_t.\end{aligned}$$

Logo,

$$\mathrm{tr}E\sum_{t=p+1}^{T}\widehat{\omega}_t\mathbf{B}^{t-p}(\mathbf{y}_p^*-\mathbf{y}_p)(\mathbf{y}_p^*-\mathbf{y}_p)'(\mathbf{B}')^{t-p}\leq\underbrace{\omega(0)c^2p^3\max_{ij}|(F+\mathbf{y}_p\mathbf{y}_p')_{ij}|}_{\equiv K}\leq\frac{K}{1-\lambda^2},$$

o que demonstra (3.20) e, conseqüentemente, (3.19). As outras duas convergências enunciadas no teorema são conseqüências da que acabamos de demonstrar. □

Como conseqüência do teorema acima, podemos, no que segue, assumir que o processo $\mathbf{y}_t$ satisfaz (3.15) para todo $t\in\mathbb{Z}$. Além, também é verdade que as matrizes $\mathbf{B}$ e Δ podem ser genéricas, de modo que para valer os resultados abaixo, as observações $\{y_t\}$ não devem necessariamente estar relacionadas de acordo com (3.14) e os erros $\mathbf{u}_t$ podem estar arbitrariamente correlacionados. Assumiremos apenas que

Condição 1. $E\mathbf{u}_t\mathbf{u}_t'=\Delta>0$;

Condição 2. $\lim_{T\to\infty}\frac{1}{T}\sum_{t=p+1}^{T}\widehat{\omega}_t\mathbf{u}_t\mathbf{u}_t'=E\{\sigma\}\Delta$;

Condição 3. $\lim_{T\to\infty}\frac{1}{T}\sum_{t=p+1}^{T}\widehat{\omega}_t\mathbf{u}_t\mathbf{u}_{t-s}'=0$, para $s=1,2,...$;

Em particular, note que $\lim_{T\to\infty}\frac{1}{T}\sum_{t=p+1}^{T}\mathbf{u}_t\mathbf{u}_t'=\Delta$.

Lema 3.3.3. *Assumindo as mesmas condições do teorema 3.3.2 e as condições acima, segue que*

$$\frac{1}{T}\sum_{t=p+1}^{T}\widehat{\omega}_t\mathbf{u}_t\mathbf{y}_{t-1}'=0.$$

Demonstração. A soma dos valores esperados dos quadrados das componentes de $\frac{1}{T}\sum_{t=p+1}^{T}\widehat{\omega}_t\mathbf{u}_t\mathbf{y}_{t-1}'$ é igual a

$$\begin{aligned}\mathrm{tr}E\frac{1}{T}\sum_{t=p+1}^{T}\widehat{\omega}_t\mathbf{y}_{t-1}\mathbf{u}_t'\frac{1}{T}\sum_{s=p+1}^{T}\widehat{\omega}_s\mathbf{u}_s\mathbf{y}_{s-1}'&\leq\frac{\omega(0)^2}{T^2}\sum_{t,s=p+1}^{T}E\mathrm{tr}\mathbf{y}_{t-1}\mathbf{u}_t'\mathbf{u}_s\mathbf{y}_{s-1}'\\&=\frac{\omega(0)^2}{T^2}\sum_{t=p+1}^{T}E\mathbf{u}_t'\mathbf{u}_tE\mathbf{y}_{t-1}'\mathbf{y}_{t-1}\\&=\frac{\omega(0)^2}{T^2}\sum_{t=p+1}^{T}\mathrm{tr}\Sigma\mathrm{tr}F.\end{aligned}$$

Logo, o lema segue diretamente do teorema de Chebychev. □

Teorema 3.3.4. *Assumindo as condições (1), (2) e (3), se* $\mathbf{y}_t$ *é definido por* (3.3.1) *para* $t=1,2,...$ *com* $\mathbf{B}$ *tal que seus autovalores são menores do que 1 em valor absoluto, então*

$$\widehat{\mathbf{B}}\to^p\mathbf{B}$$
$$\widehat{\Delta}\to^p E\{\sigma\}\Delta.$$

Demonstração. Para provar as convergências acima, note que

$$\begin{aligned}\widehat{\mathbf{B}}' - \mathbf{B}' &= \left(\frac{1}{T}\sum_{t=p+1}^{T}\widehat{\omega}_t \mathbf{y}_{t-1}\mathbf{y}_{t-1}'\right)^{-1}\left(\frac{1}{T}\sum_{t=p+1}^{T}\widehat{\omega}_t \mathbf{y}_{t-1}\mathbf{y}_t'\right) - \mathbf{B}' \\ &= \left(\frac{1}{T}\sum_{t=p+1}^{T}\widehat{\omega}_t \mathbf{y}_{t-1}\mathbf{y}_{t-1}'\right)^{-1}\left(\frac{1}{T}\sum_{t=p+1}^{T}\widehat{\omega}_t \mathbf{y}_{t-1}(\mathbf{y}_t - \mathbf{B}\mathbf{y}_{t-1})'\right) \\ &= \left(\frac{1}{T}\sum_{t=p+1}^{T}\widehat{\omega}_t \mathbf{y}_{t-1}\mathbf{y}_{t-1}'\right)^{-1}\left(\frac{1}{T}\sum_{t=p+1}^{T}\widehat{\omega}_t \mathbf{y}_{t-1}\mathbf{u}_t'\right) \xrightarrow{P} 0,\end{aligned}$$

e, combinando este resultado com a condição 2, temos que

$$\widehat{\Delta} = \frac{1}{T}\sum_{t=p+1}^{T}\widehat{\omega}_t(\mathbf{y}_t - \widehat{\mathbf{B}}\mathbf{y}_{t-1})(\mathbf{y}_t - \widehat{\mathbf{B}}\mathbf{y}_{t-1})' \xrightarrow{P} E(\sigma)\Delta.$$

□

Embora, pelo teorema acima, $\widehat{\Delta}$ não seja um estimador consistente de Δ, o estimador corrigido $\widehat{\Delta}^* = \widehat{\Delta}/E\{\sigma\}$ satisfaz esta propriedade.

3.4 Sobre a Seleção de Modelos

Nesta seção trataremos da seleção de modelos sob a suposição de que o ruído segue uma mistura na escala. Dado que os parâmetros são estimados via máxima verossimilhança, convém utilizarmos o método AIC de seleção de modelos. Para mais detalhes sobre o AIC e suas variantes, veja [8]. Para um vetor de parâmetros β, o AIC de para a estimativa (AIC $=$ AIC$(\widehat{\boldsymbol{\theta}})$) é dado por

$$\text{AIC}(\widehat{\boldsymbol{\theta}}) = -2l(\widehat{\boldsymbol{\theta}}|\mathbf{y}) + 2K,$$

onde l representa a log-verossimilhança e K o número de parâmetros. A derivação do AIC segue do fato que, como demonstrado em [3], dentro de uma classe de modelos, o elemento que maximiza a log-verossimilhança é aquele que mais se aproxima do modelo real segundo a "distância" de Kullback-Leibler[2] (ou entropia relativa) e que o mesmo ainda é um estimador viesado cujo viés é dado por K. Logo, do ponto de vista prático, escolhemos, dentro os modelos considerados, aquele com menor AIC. Obviamente, podemos usar este indicador para o parâmetro de penalização λ quando estimarmos $\boldsymbol{\theta}$ via máxima verossimilhança penalizada.

[2] a "distância" de Kullback-Leibler, ou entropia relativa, entre duas densidades de probabilidade é definida como

$$D(f,g) = \int f(x)\log\frac{f(x)}{g(x)}dx.$$

Colocamos a palavra 'distância' entre aspas, pois, embora esta medida seja comumente utilizada como uma medida de discriminação entre densidades de probabilidade (ou distribuições de probabilidade em geral), ela não satisfaz todas as propriedades que uma medida de distância deve satisfazer. Mais precisamente, não é sempre verdade que $I(f,g) = I(g,f)$.

Assumindo que o modelo pode ser representado na forma

$$\mathbf{y} = \mathbf{X}\mathbf{c} + \delta\boldsymbol{\epsilon},$$

onde $\mathbf{y} = (y_1, ..., y_T)'$, $\mathbf{X}$ é uma matriz de planejamento $T \times M$, δ é um parâmetro de escala e $\boldsymbol{\epsilon} = (\epsilon_1, ..., \epsilon_T)'$ com ϵ_t distribuido de acordo com uma mistura na escala, temos, condicionando nas variáveis explicativas, que $\boldsymbol{\theta} = (\mathbf{c}', \delta, \boldsymbol{\zeta}')'$, onde $\boldsymbol{\zeta}$ é um vetor de parâmetros associado à densidade $p_{\boldsymbol{\zeta}}$ de ϵ_t, e

$$l(\widehat{\boldsymbol{\theta}}|\mathbf{y}) = -T\log\widehat{\delta} + \sum_{t=1}^{T} \log p_{\widehat{\boldsymbol{\zeta}}}\left(\frac{y_t - \sum_j \widehat{c}_j X_{tj}}{\widehat{\delta}}\right).$$

Logo,

$$\mathrm{AIC}(\widehat{\boldsymbol{\theta}}) = 2T\log\widehat{\delta} - 2\sum_{t=1}^{T} \log p_{\widehat{\boldsymbol{\zeta}}}\left(\frac{y_t - \sum_j \widehat{c}_j X_{tj}}{\widehat{\delta}}\right) + 2K,$$

onde $K = M + 1 + D$ e D é o número de componentes de $\boldsymbol{\zeta}$.

Conforme observado em [8], o AIC pode ser pouco eficiente quando houver muitos parâmetros em relação ao tamanho da amostra. Nestes casos, pode-se utilizar uma variante do AIC, conhecido por AIC de segunda ordem e denotada por AIC_c, onde a única diferença em relação ao AIC está em um ajuste no fator K, de modo que

$$\mathrm{AIC}_c(\widehat{\boldsymbol{\theta}}) = -2l(\widehat{\boldsymbol{\theta}}|\mathbf{y}) + 2K\left(\frac{T}{T-K-1}\right).$$

No caso considerado acima, teríamos que

$$\mathrm{AIC}_c(\widehat{\boldsymbol{\theta}}) = 2T\log\widehat{\delta} - 2\sum_{t=1}^{T} \log p_{\widehat{\boldsymbol{\zeta}}}\left(\frac{y_t - \sum_j \widehat{c}_j X_{tj}}{\widehat{\delta}}\right) + 2K\left(\frac{T}{T-K-1}\right),$$

onde $K = M + 1 + D$.

3.4.1 O Critério AIC Condicional

Note que o algoritmo 3.1.1 depende das distribuições representadas por misturas na escala de normais única e exclusivamente através das ponderações na matriz $W(\boldsymbol{\theta})$ e que, fora tais ponderações, não necessitamos de mais nenhuma informação a respeito destas distribuições. Por outro lado, para o cálculo dos critérios AIC e AIC_c acima necessitamos conhecer p. Nesta seção apresentamos uma variação dos critérios AIC e AIC_c que independe de p e que é identificada unicamente através das ponderações em $W(\boldsymbol{\theta})$.

Modelando os dados de acordo com

$$y_t|\mathbf{x}_t; \sigma_t \sim \mathcal{N}\left(\sum_j c_j X_{tj}, \frac{\delta^2}{\sigma_t}\right)$$
$$\sigma_t \sim h$$

onde h é a densidade de mistura, temos que

$$l(\widehat{\boldsymbol{\theta}}|\mathbf{y},\boldsymbol{\sigma}) \approx -\frac{T}{2}\log\widehat{\delta^2} - \frac{1}{2\widehat{\delta^2}}\sum_{t=1}^{T}\sigma_t(y_t - \sum_j \widehat{c}_j X_{tj})^2 + \log h(\boldsymbol{\sigma}|\widehat{\boldsymbol{\zeta}}),$$

onde "$\approx$" significa igualdade a menos de uma constante. Logo,

$$\mathrm{AIC}(\widehat{\boldsymbol{\theta}}) = T\log\widehat{\delta^2} + \frac{1}{\widehat{\delta^2}}\sum_{t=1}^{T}\sigma_t(y_t - \sum_j \widehat{c}_j X_{tj})^2 - 2\log h(\boldsymbol{\sigma}|\widehat{\boldsymbol{\zeta}}) + 2K \tag{3.21}$$

e

$$\mathrm{AIC}_c(\widehat{\boldsymbol{\theta}}) = T\log\widehat{\delta^2} + \frac{1}{\widehat{\delta^2}}\sum_{t=1}^{T}\sigma_t(y_t - \sum_j \widehat{c}_j X_{tj})^2 - 2\log h(\sigma|\widehat{\boldsymbol{\zeta}}) + 2K\left(\frac{T}{T-K-1}\right). \tag{3.22}$$

Ocorre, no entanto, que as variáveis σ_t são latentes e, portanto, não-observáveis. Sugerimos, deste modo, tomar, o valor esperado das expressões (3.21) e (3.22) de acordo com a densidade a "posteriori" $k(\cdot|y_t,\mathbf{x}_t)$ e, assim, obter

$$\mathrm{EAIC}(\widehat{\boldsymbol{\theta}}) = T\log\widehat{\delta^2} + \frac{1}{\widehat{\delta^2}}(\mathbf{y}-\mathbf{X}\widehat{\mathbf{c}})'W(\widehat{\boldsymbol{\theta}})(\mathbf{y}-\mathbf{X}\widehat{\mathbf{c}}) - 2E\{\log h(\boldsymbol{\sigma}|\widehat{\boldsymbol{\zeta}})|\mathbf{y},\mathbf{x}\} + 2K$$

e

$$\mathrm{EAIC}_c(\widehat{\boldsymbol{\theta}}) = T\log\widehat{\delta^2} + \frac{1}{\widehat{\delta^2}}(\mathbf{y}-\mathbf{X}\widehat{\mathbf{c}})'W(\widehat{\boldsymbol{\theta}})(\mathbf{y}-\mathbf{X}\widehat{\mathbf{c}}) - 2E\{\log h(\boldsymbol{\sigma}|\widehat{\boldsymbol{\zeta}})|\mathbf{y},\mathbf{x}\} + 2K\left(\frac{T}{T-K-1}\right),$$

onde $\mathrm{EAIC}(\widehat{\boldsymbol{\theta}}) = E\{\mathrm{AIC}(\widehat{\boldsymbol{\theta}})|\mathbf{y},\mathbf{x}\}$, $\mathrm{EAIC}_c(\widehat{\boldsymbol{\theta}}) = E\{\mathrm{AIC}_c(\widehat{\boldsymbol{\theta}})|\mathbf{y},\mathbf{x}\}$ e

$$W(\widehat{\boldsymbol{\theta}}) = \mathrm{diag}(E\{\sigma_1|\mathbf{y},\mathbf{x}\},...,E\{\sigma_T|\mathbf{y},\mathbf{x}\}).$$

Sob o ponto de vista prático, devemos escolher o modelo com menor $\mathrm{EAIC}(\boldsymbol{\theta}^*)$ ou $\mathrm{EAIC}_c(\boldsymbol{\theta}^*)$, onde $\boldsymbol{\theta}^*$ é a estimativa obtida ao final das iterações do algoritmo EM. Para justificar o uso de EAIC, notamos que o AIC é uma estimativa do valor esperado da distância de Kullback-Leibler entre o modelo real (densidade ou distribuição) real e a estimada $p(\cdot|\widehat{\boldsymbol{\theta}})$, isto é, de

$$\int p(\mathbf{y},\boldsymbol{\sigma})\int f(x)\log\frac{f(x)}{p(x|\widehat{\boldsymbol{\theta}}(\mathbf{y},\boldsymbol{\sigma}))}dx d\mathbf{y} d\sigma = C - E_{\mathbf{y},\boldsymbol{\sigma}}E_f\{\log p(x|\widehat{\boldsymbol{\theta}}(\mathbf{y},\boldsymbol{\sigma}))\},$$

onde C indica uma constante (com relação aos modelos considerados). Logo, buscar o modelo que minimiza a distância de Kullback-Leibler equivale a buscar o modelo que maximiza

$$E_{\mathbf{y},\boldsymbol{\sigma}}E_f\{\log p(x|\widehat{\boldsymbol{\theta}}(\mathbf{y},\boldsymbol{\sigma}))\}.$$

Como demonstrado em [8][3], a esperança acima pode ser aproximada por

$$E_f\{\log p(x|\widehat{\boldsymbol{\theta}}(x))\} - \mathrm{tr}\left[J(\boldsymbol{\theta})[I(\boldsymbol{\theta})]^{-1}\right], \tag{3.23}$$

[3] capítulo 6.

onde

$$I(\boldsymbol{\theta}) = E_f\left[-\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\log p(x|\boldsymbol{\theta})\right] \text{ e } J(\boldsymbol{\theta}) = E_f\left[\left(\frac{\partial}{\partial\boldsymbol{\theta}}\log p(x|\boldsymbol{\theta})\right)\left(\frac{\partial}{\partial\boldsymbol{\theta}}\log p(x|\boldsymbol{\theta})\right)'\right].$$

Seguindo a mesma argumentação que em [8], observamos que a notação $\boldsymbol{\theta}(x)$ é usada apenas para enfatizar que no lado direito de (3.23) apenas uma variável aleatória aparece, e pode-se assumir que ela inclui os dados reais. Usando agora o fato que $E_f\{\log p(x|\widehat{\boldsymbol{\theta}}(x))\} = E_f\{E\{\log p(x|\widehat{\boldsymbol{\theta}}(x))|\mathbf{y},\mathbf{x}\}\}$, podemos inferir que um critério para a seleção de modelos é da forma

$$E\{\log p(\mathbf{y},\boldsymbol{\sigma}|\widehat{\boldsymbol{\theta}})|\mathbf{y},\mathbf{x}\} - \text{tr}\left[J(\boldsymbol{\theta})[I(\boldsymbol{\theta})]^{-1}\right].$$

Multiplicando-se o critério acima por -2, obtemos o resultado (mais geral)[4] que queríamos.

[4]Sob certas condições de regularidade $J(\boldsymbol{\theta})[I(\boldsymbol{\theta})]^{-1} = K$ e obtemos o critério EAIC definido acima.

Capítulo 4

Aplicação em Modelos de Regressão Semiparamétricos

4.1 Introdução

Dado o modelo

$$y_t = f(x_t) + \delta\epsilon_t \tag{4.1}$$

onde δ representa o parâmetro de escala, nosso objetivo é estimar a função alvo f sob a hipótese de que ϵ_t tem caudas pesadas, modelando, para tanto, o ruído segundo uma distribuição $SM_h(0,1;\psi)$, onde h pode depender de um vetor de parâmetros ζ. Deste modo, assumiremos que

- $t = 1, ..., T$;
- f é uma função desconhecida pertencente a uma determinada classe de funções $\mathcal{H}$;
- $\{\epsilon_t\}$ é uma coleção de variáveis aleatórias i.i.d. cuja densidade pode ser representada na forma (1.3) com $c = 0$, ie, $\epsilon_t \sim SM_h(0,1;\psi)$.

Assuma também que $E\epsilon_t = 0$, para todo t, de modo que $E(y_t|x_t) = f(x_t)$. O modelo acima generaliza os modelos usuais no sentido que permite outras distribuições além da normal para o ruído, incluindo distribuições de cauda pesada e variância infinita como, por exemplo, as distribuições estáveis. Com relação à distribuição de y_t é fácil ver que $y_t|x_t \sim SM_h(f(x_t), \delta; \psi)$[1].

Por hipótese, a distribuição de ϵ_t pode ser representada na forma

$$p(\epsilon_t) = \int_0^\infty \frac{1}{\psi_{\boldsymbol{\theta}}(\sigma_t)} \phi\left(\frac{\epsilon_t}{\psi_{\boldsymbol{\theta}}(\sigma_t)}\right) h(\sigma_t) d\sigma_t,$$

[1] Assim como no capítulo anterior, para não deixar a notação excessivamente carregada, manteremos em mente que as densidades associadas a y_t são condicionadas a x_t e deixaremos, conseqüentemente, de explicitar constantemente este fato.

onde ϕ é densidade da normal padrão, logo, podemos assumir que a variável aleatória $\epsilon_t|\sigma_t$ segue uma distribuição normal com precisão $1/\psi_{\boldsymbol{\theta}}(\sigma_t)^2$ e, conseqüentemente, que

$$(y_t - f(x_t))|\sigma_t \sim \mathcal{N}\left(0, \psi_{\boldsymbol{\theta}}(\sigma_t)^2\right).$$

Observe que, em geral, a função $\psi_{\boldsymbol{\theta}}$ pode depender do vetor de parâmetros $\boldsymbol{\theta}$ associado ao modelo, porém, por simplicidade, nós a denotaremos simplesmente por ψ e explicitaremos a dependência em relação a tal vetor de parâmetros apenas quando necessário. Mais precisamente, o vetor de parâmetros $\boldsymbol{\theta}$ é um elemento pertencente a $\mathcal{H}' \times B \subset \mathcal{H} \times \mathbb{R}^d$, onde d é um inteiro positivo e $\mathcal{H}'$ é um subespaço de $\mathcal{H}$.

Caso a variável aleatória σ fosse observável, e nos fossem dadas as observações $\sigma_1, ..., \sigma_T$ poderíamos estimar a função alvo f maximizando

$$-\frac{1}{T}\sum_{t=1}^{T}\frac{(y_t - f(x_t))^2}{\delta^2\psi(\sigma_t)^2} + J_\lambda(f), \tag{4.2}$$

onde J_λ é uma eventual função de penalização sobre f e λ é o respectivo parâmetro de penalização. Em outras palavras, a função f seria estimada pelo método de máxima verossimilhança penalizada (ou mínimos quadrados penalizados) ponderada pelo vetor $(\psi(\sigma_1), ..., \psi(\sigma_T))$. É este vetor que penaliza observações discrepantes garantindo, desta forma, a robustez do modelo. Talvez o exemplo mais usual de função de penalização seja dado por

$$J_\lambda(f) = -\lambda\int[f''(x)]^2dx. \tag{4.3}$$

A técnica de estimação de f através de funções de penalização do tipo (4.3) é conhecida por *smoothing-splines*. Na prática, a função de penalização J_λ favorece funções mais suaves e o parâmetro λ pode ser visto, neste caso, como um parâmetro de suavização que governa o compromisso entre suavidade e viés do estimador de f. De fato, quando utilizamos *smoothing splines*, o que ocorre se $\mathcal{H}$ é o espaço de Sobolev W_2^m, $\lambda = 0$ resulta numa interpolação dos dados, enquanto que para λ grande, temos uma estimativa próxima ao estimador linear. Convém observar que, definindo $\mathcal{H}$ como um espaço de Sobolev, o problema de otimização (4.2) que, a priori, é infinito dimensional, reduz-se a um problema de dimensão finita, dado que a solução ótima do problema é um spline cúbico. Por outro lado, se definimos $J_\lambda \equiv 0$, então, voltamos para o método de mínimos quadrados usual.

Uma outra maneira de se estimar f e reduzir o problema para espaços de dimensão finita (e pequena), além da técnica de *smoothing splines*, é aproximá-la como uma combinação linear de funções base, tais como ondaletas ou *B-splines*. Denontando tais funções por B_j, para $j = 1, ..., M$, o problema resume-se em determinar o vetor de coeficientes, $(c_1, ..., c_M)'$, que maximiza

$$-\frac{1}{T}\sum_{t=1}^{T}\frac{1}{\delta^2\psi(\sigma_t)^2}\left(y_t - \sum_j c_jB_j\right)^2 + J_\lambda\left(\sum_j c_jB_j\right).$$

Nos parágrafos acima, discutimos a estimação da função alvo sob a hipótese de que σ fosse observável. Ocorre, porém, que isto não é verdade, de modo que os estimadores da função alvo f e do parâmetro de escala δ são aqueles que maximizam a função critério

$$\frac{1}{T}\sum_{t=1}^{T} l(\bar{f}_t, \delta; y_t) + J_\lambda(\bar{f}), \tag{4.4}$$

onde $\bar{f}$ é a aproximação de f segundo o método adotado (*smoothing-splines*, B-*splines*), $l(\bar{f}_t, \delta; y_t)$ é a log-verossimilhança associada a y_t calculada em $\bar{f}$ e δ. Nas seções seguintes discutiremos como obter as estimativas de f e δ que maximizam (4.4).

Teorema 4.1.1. *Assumindo δ conhecido, a solução de* (4.4) *é única.*

Prova. Em primeiro lugar, note que

$$\begin{aligned} l(\bar{f}_t, \delta; y_t) &= \log \int_0^\infty \frac{1}{\delta\psi(\sigma_t)} \phi\left(\frac{y_t - \bar{f}_t}{\delta\psi(\sigma_t)}\right) h(\sigma_t) d\sigma_t \\ &= \log \int_0^\infty \frac{1}{\psi(\sigma_t)} \phi\left(\frac{y_t - \bar{f}_t}{\delta\psi(\sigma_t)}\right) h(\sigma_t) d\sigma_t - \log \delta \\ &\equiv p\left(\frac{y_t - \bar{f}_t}{\delta}\right) - \log \delta, \end{aligned}$$

de modo que maximizar (4.4) equivale a maximizar

$$\frac{1}{T}\sum_{t=1}^{T} p\left(\frac{y_t - \bar{f}_t}{\delta}\right) - \log \delta + J_\lambda(f).$$

O termo $-\log\delta$ atua como uma penalização sobre o parâmetro de escala. Agora para δ fixado e nos restringindo às funções f tais que $J_\lambda = c_0$, onde c_0 é uma constante arbitrária, suponha que $\bar{f}_\delta \neq \bar{g}_\delta$ maximizem (4.4) e defina $\bar{h}_{\delta,\lambda} = \lambda\bar{f}_\delta + (1-\lambda)\bar{g}_\delta$, para $0 < \lambda < 1$, de modo que, usando o fato de que p é estritamente quase-convexa, temos que

$$p\left(\frac{y_t - \bar{h}_t}{\delta}\right) < \max\left\{p\left(\frac{y_t - \bar{f}_t}{\delta}\right), p\left(\frac{y_t - \bar{g}_t}{\delta}\right)\right\} = p\left(\frac{y_t - \bar{f}_t}{\delta}\right).$$

Logo,

$$\frac{1}{T}\sum_{t=1}^{T} p\left(\frac{y_t - \bar{h}_t}{\delta}\right) - \log\delta + c_0 < \frac{1}{T}\sum_{t=1}^{T} p\left(\frac{y_t - \bar{f}_t}{\delta}\right) - \log\delta + c_0,$$

o que implica em contradição. Logo, devemos ter $\bar{f}_\delta = \bar{g}_\delta$. □

4.2 Aplicação do Algoritmo EM

4.2.1 Introdução

Assumindo que a variável observada segue o modelo

$$\begin{aligned} y_t|\sigma_t &\sim \mathcal{N}(f(x_t), \delta^2\psi(\sigma_t)^2), \\ \sigma_t &\sim h, \end{aligned} \tag{4.5}$$

então, $y_t \sim SM_h(f(x_t), \delta; \psi)$. Para estimar os parâmetros do modelo acima (incluindo os parâmetros associados à aproximação de f) e, além disso, incorporar uma função de penalização, propomos o uso do algoritmo EM modificado de [24]. Observe que as variáveis latentes σ_t controlam a variância das observações y_t e, conseqüentemente, suavizam o impacto de observações extremas sobre o estimador. Observe também que, ao obter os parâmetros que maximizam a log-verossimilhança do modelo (4.5), obteremos também, por construção, os parâmetros que maximizam a log-verossimilhança associada às distribuições das observações $(y_1, ..., y_T)'$.[2] É claro que o resultado citado vale quando restringimos o espaço ao qual pertence f (por exemplo, quando fixamos a resolução máxima em uma análise de ondaletas ou quando fixamos os nós em uma análise via B-*splines*). Antes de continuar, no entanto, precisamos introduzir alguma notação:

- $\boldsymbol{\theta}$: vetor de parâmetros de interesse, pertencente a $\mathcal{H}' \times B$;
- $z_t \equiv (y_t, x_t)$: t-ésima observação;
- $\mathbf{y} \equiv (y_1, ..., y_T)'$ e $\mathbf{z} \equiv (z_1, ..., z_T)'$: vetores de observações;
- $\boldsymbol{\sigma} \equiv (\sigma_1, ..., \sigma_T)'$: vetor de variáveis latentes;
- $Q(\boldsymbol{\theta}|\boldsymbol{\theta}') \equiv E_{\boldsymbol{\theta}'}\{\log p(\mathbf{y}, \boldsymbol{\sigma}|\boldsymbol{\theta})\big|\mathbf{z}\}$ onde (i) a esperança é calculada com relação à distribuição conjunta de $\boldsymbol{\sigma}$ dado $\mathbf{z}$ e utilizando-se o vetor de parâmetros da iteração anterior do algoritmo, $\boldsymbol{\theta}'$, e (ii) p é a distribuição conjunta de $(y_1, \sigma_1), ..., (y_T, \sigma_T)$ (dados $(x_1, ..., x_T)'$);
- $Q_t(\boldsymbol{\theta}|\boldsymbol{\theta}') \equiv E_{\boldsymbol{\theta}'}\{\log p(y_t, \sigma_t|\boldsymbol{\theta})\big|z_t\}$. Note que $Q(\boldsymbol{\theta}|\boldsymbol{\theta}') = \sum_{t=1}^{T} Q_t(\boldsymbol{\theta}|\boldsymbol{\theta}')$;
- a exemplo da função de verossimilhança, denotaremos a função de penalização p_λ por $J_\lambda(\boldsymbol{\theta})$.

Assumiremos que os elementos de $\mathcal{H}'$ podem ser escritos como combinações de funções base $\{B_i\}_{i=1}^{K}$, onde $K \in \mathbb{Z}$ ou $K = \infty$, a depender das hipóteses sobre $\mathcal{H}'$. Por exemplo, se assumirmos que $\mathcal{H}'$ é o espaço de polinômios de grau menor ou igual a p, então, poderíamos tomar $K = p$ e $B_j(x) = x^j$, para $j = 0, 1, ..., p$. Por outro lado, se assumirmos que $\mathcal{H}'$ é o espaço de Sobolev $W_2^2[0, 1]$, poderíamos assumir que $K = \infty$ e que os B_j's são os elementos de qualquer seqüência ortonormal completa.

O algoritmo de estimação da função alvo f e dos parâmetros associados a h, denotados por ζ, consiste, portanto, em maximizar com relação a $\boldsymbol{\theta} = (f, \zeta)$,

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}') = \sum_{t=1}^{T} Q_t(\boldsymbol{\theta}|\boldsymbol{\theta}'),$$

ou, em casos que haja um termo de penalização J_λ, em maximizar

$$S(\boldsymbol{\theta}|\boldsymbol{\theta}') \equiv Q(\boldsymbol{\theta}|\boldsymbol{\theta}') + J_\lambda(\boldsymbol{\theta}) = \sum_{t=1}^{T} Q_t(\boldsymbol{\theta}|\boldsymbol{\theta}') + J_\lambda(\boldsymbol{\theta})$$

até a convergência. Note que

$$\begin{aligned} \log p(y_t, \sigma_t|\boldsymbol{\theta}) &= \log p(y_t|\sigma_t, \boldsymbol{\theta}) + \log h(\sigma_t|\boldsymbol{\theta}) \\ &= \log\left[\frac{1}{\delta\psi(\sigma_t)}\phi\left(\frac{y_t - f(x_t)}{\delta\psi(\sigma_t)}\middle|\boldsymbol{\theta}\right)\right] + \log h(\sigma_t|\boldsymbol{\theta}) \end{aligned}$$

[2] veja [9] para mais detalhes.

de modo que

$$Q_t(\boldsymbol{\theta}|\boldsymbol{\theta}') = E_{\boldsymbol{\theta}'}\left\{\log\left[\frac{1}{\delta\psi(\sigma_t)}\phi\left(\frac{y_t - f(x_t)}{\delta\psi(\sigma_t)}\middle|\boldsymbol{\theta}\right)\right] + \log h(\sigma_t|\boldsymbol{\theta})\middle|z_t\right\}.$$

Agora,

$$\frac{1}{\delta\psi(\sigma_t)}\phi\left(\frac{y_t - f(x_t)}{\delta\psi(\sigma_t)}\middle|\boldsymbol{\theta}\right) = \frac{1}{\sqrt{2\pi}\delta\psi(\sigma_t)}e^{-\frac{(y_t - f_{\boldsymbol{\theta}}(x_t))^2}{2\delta^2\psi(\sigma_t)^2}}$$

e, conseqüentemente,

$$\log\left[\frac{1}{\delta\psi(\sigma_t)}\phi\left(\frac{y_t - f(x_t)}{\delta\psi(\sigma_t)}\middle|\boldsymbol{\theta}\right)\right] = -\frac{1}{2}\log(2\pi) - \log\psi(\sigma_t) - \log\delta - \frac{(y_t - f_{\boldsymbol{\theta}}(x_t))^2}{2\delta^2\psi(\sigma_t)^2}.$$

Logo,

$$\begin{aligned} Q_t(\boldsymbol{\theta}|\boldsymbol{\theta}') = &-\frac{1}{2}\log(2\pi) - E_{\boldsymbol{\theta}'}\left\{\log\psi(\sigma_t)|z_t\right\} - \log\delta \\ &- E_{\boldsymbol{\theta}'}\left\{\frac{1}{\psi(\sigma_t)^2}\middle|z_t\right\}\frac{(y_t - f_{\boldsymbol{\theta}}(x_t))^2}{2\delta^2} + E_{\boldsymbol{\theta}'}\left\{\log h(\sigma_t|\boldsymbol{\theta})|z_t\right\}. \end{aligned} \tag{4.6}$$

Como o objetivo é maximizar $Q(\boldsymbol{\theta}'|\boldsymbol{\theta})$, podemos ignorar o termo $-\frac{1}{2}\log(2\pi)$ na expressão (4.6) e considerar

$$\begin{aligned} Q_t(\boldsymbol{\theta}|\boldsymbol{\theta}') = &-E_{\boldsymbol{\theta}'}\left\{\log\psi(\sigma_t)|z_t\right\} - \log\delta - \frac{1}{2\delta^2}E_{\boldsymbol{\theta}'}\left\{\frac{1}{\psi(\sigma_t)^2}\middle|z_t\right\}(y_t - f_{\boldsymbol{\theta}}(x_t))^2 \\ &+ E_{\boldsymbol{\theta}'}\left\{\log h(\sigma_t|\boldsymbol{\theta})|z_t\right\}. \end{aligned}$$

Pela independência entre as variáveis aleatórias $y_1, ..., y_T$,

$$\begin{aligned} Q(\boldsymbol{\theta}|\boldsymbol{\theta}') = &-\sum_{t=1}^{T}E_{\boldsymbol{\theta}'}\{\log\psi(\sigma_t)|z_t\} - T\log\delta - \frac{1}{2\delta^2}\sum_{t=1}^{T}E_{\boldsymbol{\theta}'}\left\{\frac{1}{\psi(\sigma_t)^2}\middle|z_t\right\}(y_t - f_{\boldsymbol{\theta}}(x_t))^2 \\ &+ \sum_{t}E_{\boldsymbol{\theta}'}\{\log h(\sigma_t|\boldsymbol{\theta})|z_t\}. \end{aligned} \tag{4.7}$$

A expressão acima pode ser simplificada, definindo

$$C(\boldsymbol{\zeta};\boldsymbol{\theta}') \equiv \sum_{t=1}^{T}E_{\boldsymbol{\theta}'}\left\{\log\frac{h(\sigma_t|\boldsymbol{\theta})}{\psi(\sigma_t)}\middle|z_t\right\},$$

como a componente que contém apenas os parâmetros associados à distribuição do ruído,

$$W_T(\boldsymbol{\theta}') \equiv \text{diag}\left(E_{\boldsymbol{\theta}'}\left\{\frac{1}{\psi(\sigma_1)^2}\middle|z_t\right\}, ..., E_{\boldsymbol{\theta}'}\left\{\frac{1}{\psi(\sigma_T)^2}\middle|z_t\right\}\right) \tag{4.8}$$

como o vetor de ponderações para a estimação da função alvo f, e

$$\mathbf{y} - \mathbf{f} \equiv (y_1 - f(x_1), ..., y_T - f(x_T))'$$

de modo que

$$\frac{1}{2\delta^2}\sum_{t=1}^{T} E_{\theta'}\left\{\frac{1}{\psi(\sigma_t)^2}\middle| z_t\right\}(y_t - f_\theta(x_t))^2 = \frac{1}{2\delta^2}(\mathbf{y}-\mathbf{f})'W_T(\theta')(\mathbf{y}-\mathbf{f})$$

e

$$Q(\theta|\theta') = C(\zeta;\theta') - T\log\delta - \frac{1}{2\delta^2}(\mathbf{y}-\mathbf{f})'W_T(\theta')(\mathbf{y}-\mathbf{f}). \tag{4.9}$$

Note que,

i. assim como para ψ, a matriz W_T pode depender dos parâmetros em θ;

ii. C independe de f e dos parâmetros associados a sua aproximação;

iii. caso haja um termo de penalização J_λ, a função critério fica dada por

$$S(\theta|\theta') = C(\zeta;\theta') - T\log\delta - \frac{1}{2\delta^2}(\mathbf{y}-\mathbf{f})'W_T(\theta')(\mathbf{y}-\mathbf{f}) + J_\lambda(\theta) \tag{4.10}$$

sendo que a representação da função de penalização J_λ depende de hipóteses mais precisas sobre o espaço ao qual f pertence e sobre as funções base utilizadas para aproximação de f. Tais especificações serão postergadas para a próxima seção.

Com relação às esperanças acima, elas são calculadas baseadas nas distribuições condicionais

$$\begin{aligned} k(\sigma_t|z_t,\theta') &= \frac{p(\sigma_t, y_t|\theta')}{p(y_t|\theta')} \\ &\propto \frac{1}{\delta\psi_{\theta'}(\sigma_t)}\phi\left(\frac{y_t - f_{\theta'}(x_t)}{\delta\psi_{\theta'}(\sigma_t)}\right)h(\sigma_t|\theta') \\ &\approx \frac{1}{\delta\psi_{\theta'}(\sigma_t)}\phi\left(\frac{y_t - \bar{f}_{\theta'}(x_t)}{\delta\psi_{\theta'}(\sigma_t)}\right)h(\sigma_t|\theta'), \end{aligned} \tag{4.11}$$

onde $\bar{f}_{\theta'}$ é a aproximação de f com base nos parâmetros estimados na iteração anterior. Nos casos em que as integrais em (4.7) forem analiticamente intratáveis, o uso de algum método numérico ou de Monte Carlo no passo E do algoritmo deverá ser utilizado para aproximar os valores esperados.

4.2.2 O Caso Canônico

Quando $\psi(\sigma) = \sigma^{-1/2}$, a matriz de ponderações torna-se

$$W_T(\theta') \equiv \text{diag}\left(E_{\theta'}\{\sigma_1|z_t\}, ..., E_{\theta'}\{\sigma_T|z_t\}\right)$$

enquanto que

$$C(\zeta;\theta') \equiv \sum_{t=1}^{T} E_{\theta'}\{\log h(\sigma_t|\theta)|z_t\},$$

uma vez que ψ independe de qualquer parâmetro do modelo. Além disso, a densidade em (4.11) fica da forma

$$\begin{aligned} k(s|y;\boldsymbol{\theta}) &\propto \frac{\sqrt{s}}{\delta}\phi\left(\frac{\sqrt{s}}{\delta}(y - f_{\boldsymbol{\theta}}(x))\right) h(s|\boldsymbol{\theta}) \\ &\approx \frac{\sqrt{s}}{\delta}\phi\left(\frac{\sqrt{s}}{\delta}(y - \bar{f}_{\boldsymbol{\theta}}(x))\right) h(s|\boldsymbol{\theta}) \end{aligned} \tag{4.12}$$

Exemplo 4.2.1 (Distribuição t de Student — ν Conhecido). *Suponha que o ruído seja distribuido de acordo com uma distribuição t com ν graus de liberdade, o qual assumiremos conhecido. Ou seja, assuma que $\epsilon_t \sim t_\nu(0, \zeta^2)$. Isto é o mesmo que assumir $y_t \sim t_\nu(f(x_t), \delta^2)$, de modo que*

$$p(y_t) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\Gamma\left(\frac{1}{2}\right)}\frac{1}{\sqrt{\nu\delta^2}}\left[1 + \frac{1}{\nu}\left(\frac{y_t - f(x_t)}{\delta}\right)^2\right]^{-\frac{\nu+1}{2}}.$$

Então, para estimar a função alvo, modelaremos os dados de acordo com

$$\begin{aligned} y_t|\sigma_t &\sim \mathcal{N}\left(f(x_t), \frac{\delta^2}{\sigma_t}\right), \\ \sigma_t &\sim \Gamma\left(\frac{\nu}{2}, \frac{2}{\nu}\right), \end{aligned}$$

pois, sabe-se que dado o modelo acima, a distribuição marginal de y_t é uma $t_\nu(f(x_t), \delta^2)$, ver [17][3]. Note que, neste caso, $\psi(\sigma) = \frac{\delta}{\sqrt{\sigma}}$ e h é a densidade associada a distribuição $\Gamma\left(\frac{\nu}{2}, \frac{2}{\nu}\right)$, a qual é dada por

$$h(s) = \frac{(\nu/2)^{\nu/2}}{\Gamma(\nu/2)} s^{\frac{\nu}{2}-1} e^{-\frac{\nu}{2}s}, \text{para } s > 0.$$

Camo estamos assumindo ν conhecido e estamos no caso canônico onde ψ independe de quaisquer parâmetros, temos que

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}') = -T\log\delta - \frac{1}{2\delta^2}(\mathbf{y} - \mathbf{f})' W_T(\boldsymbol{\theta}')(\mathbf{y} - \mathbf{f}). \tag{4.13}$$

Agora,

$$E_{\boldsymbol{\theta}'}\left\{\frac{1}{\psi(\sigma_t)^2}\middle| y_t\right\} = E_{\boldsymbol{\theta}'}\{\sigma_t|y_t\} \equiv \bar{\sigma}_t,$$

e

$$W_T(\boldsymbol{\theta}') \equiv diag\{\bar{\sigma}_1, ..., \bar{\sigma}_T\}. \tag{4.14}$$

[3] Aqui estamos assumindo a seguinte definição para a densidade associada a distribuição gama, $\Gamma(\alpha, \beta)$:

$$h(s) = \frac{1}{\beta^\alpha \Gamma(\alpha)} s^{\alpha-1} e^{-s/\beta}$$

Por último, observamos que as esperanças acima são calculadas a partir da densidade definida em (4.11) *a qual, neste caso, assume a forma*

$$k(\sigma_t|y_t,\boldsymbol{\theta}') \propto \frac{\sqrt{\sigma_t}}{\delta}\phi\left(\frac{y_t - f_{\boldsymbol{\theta}'}(x_t)}{\delta/\sqrt{\sigma_t}}\right) h(\sigma_t).$$

Da expressão acima e da definição da densidade da distribuição gama, é fácil ver que

$$\sigma_t|y_t,\boldsymbol{\theta}' \sim \Gamma\left(\frac{1}{2}(\nu+1), \frac{1}{2}(\nu + r_t^2/\delta^2)\right),$$

onde

$$r_t = y_t - f_t.$$

Em particular, as ponderações $\bar{\sigma}_t$ são dadas por

$$\bar{\sigma}_t = \frac{\nu+1}{\nu + r_t^2/\delta^2}. \tag{4.15}$$

Pode-se notar que, por (4.15)*, quanto maior a distância entre a observação e a curva ajustada, menor será o peso dada a esta observação na estimação dos parâmetros do modelo.*

Se $(f_1,...,f_T)' = (f(x_1),...,f(x_T))'$ puder ser escrita como $\mathbf{Bc}$, onde $\mathbf{B}$ é uma determinada matriz de planejamento, tem-se de (4.13) *e* (4.15) *que os estimadores de $\mathbf{c}$ e δ no $(k+1)$-ésimo passo do algoritmo EM são dados por*

$$\mathbf{c}^{(k+1)} = (X'W^{(k)}X)^{-1}X'W^{(k)}\mathbf{y}$$

e

$$(\delta^2)^{(k+1)} = \frac{1}{T}(\mathbf{y} - \mathbf{Bc}^{(k+1)})'W^{(k)}((\mathbf{y} - \mathbf{Bc}^{(k+1)}))$$

onde

$$W_T^{(k)} \equiv diag\{\bar{\sigma}_1^{(k)},...,\bar{\sigma}_T^{(k)}\}$$

e $\bar{\sigma}_T^{(k)}$ é o valor esperado de σ_t dado y_t e os parâmetros $\mathbf{c}$ e δ estimados até o momento. ◊

Exemplo 4.2.2 (Distribuição de Cauchy). *A distribuição de Cauchy é um caso particular da distribuição t de Student onde $\nu = 1$. Logo, temos que*

$$\bar{\sigma}_t = \frac{2}{2 + r_t^2/\delta^2}.$$

◊

O exemplo a seguir explora um outro elemento da classe das distribuições formadas por mistura de normais através da escala, a distribuição exponencial dupla,

$$p(\epsilon) = \frac{1}{2}e^{-|\epsilon|}. \tag{4.16}$$

Note que estimar a função alvo e demais parâmetros através da maximização de (4.16) é equivalente a estimá-los através da minimização da norma em L^1, pois, $\log p(\epsilon) = \text{cte} - |\epsilon| \propto -|\epsilon|$.

Exemplo 4.2.3 (Distribuição Exponencial Dupla). *Para o caso em que a distribuição do ruído é a exponencial dupla, a densidade de mistura é dada por*

$$h(\sigma) = \frac{1}{2\sigma^2}e^{-\frac{1}{2\sigma}}.$$

Deste modo, como descrito anteriormente, os dados devem ser modelados de acordo com

$$y_t|\sigma_t \sim \mathcal{N}\left(f(x_t), \frac{\delta}{\sigma_t}\right),$$
$$\sigma_t \sim h,$$

e $\psi(s) = \frac{1}{\sqrt{s}}$. *Em particular, observe que* h *e* ψ *independem de quaisquer parâmetros, logo, podemos desprezar os termos* $C(\boldsymbol{\theta}|\boldsymbol{\theta}')$ *em* $Q(\boldsymbol{\theta}|\boldsymbol{\theta}')$ *e assumir*

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}') \equiv -T\log\delta - \frac{1}{2\delta^2}(\mathbf{y}-\mathbf{f})'W_T(\boldsymbol{\theta}')(\mathbf{y}-\mathbf{f}).$$

Ou seja, a função alvo será estimada via mínimos quadrados ponderados onde os pesos são dados por

$$\bar{\sigma}_t = E_{\boldsymbol{\theta}'}\left\{\sigma_t\middle|z_t\right\} \tag{4.17}$$

e a matriz de ponderação $W_T(\boldsymbol{\theta}')$ *por* $diag(\bar{\sigma}_1, ..., \bar{\sigma}_T)$.

Os valores esperados em (4.17) *são calculados com relação à função densidade de probabilidade*

$$\begin{aligned} k(\sigma_t|y_t, \boldsymbol{\theta}') &\propto \sigma_t^{1/2}\phi\left(\frac{y_t - f(x_t)}{\delta/\sigma_t^{\frac{1}{2}}}\right)\frac{1}{\sigma_t^2}e^{-\frac{1}{2\sigma_t}} \\ &\propto \sigma_t^{1/2}e^{-\frac{\sigma_t(y_t-f(x_t))^2}{2\delta^2}}\frac{1}{\sigma_t^2}e^{-\frac{1}{2\sigma_t}} \\ &\propto \frac{1}{\sigma_t^{3/2}}e^{-\frac{1}{2}\left(\frac{r_t^2\sigma_t}{\delta^2}+\frac{1}{\sigma_t}\right)}, \end{aligned} \tag{4.18}$$

o que resulta em

$$\begin{aligned} \bar{\sigma}_t &= C_t^{-1}\int_0^\infty \frac{e^{-\frac{1}{2}(\frac{r_t^2\sigma_t}{\delta^2}+\frac{1}{\sigma_t})}}{\sigma_t^{1/2}}d\sigma_t = C_t^{-1}\mathcal{L}\left[\frac{e^{-\frac{1}{2s}}}{\sqrt{s}}\right]\left(\frac{r_t^2}{2\delta^2}\right) \\ &= \frac{\delta}{|r_t|}, \end{aligned}$$

onde $\mathcal{L}[g](x)$ representa a transformada de Laplace calculada da função g calculada em x. A igualdade acima é consequência do fato que

$$\mathcal{L}\left[\frac{e^{-\alpha^2/s}}{\sqrt{s}}\right](x) = \sqrt{\frac{\pi}{x}}e^{-2\alpha\sqrt{x}}$$

para $s > 0$ e para todo $\alpha \in \mathbb{C}$ com $\Re[\alpha] \geq |\Im[\alpha]|$. No nosso caso, dado que $\alpha = 2^{-1/2} \in \mathbb{R}$ obviamente satisfaz a condição exigida, temos que vale a igualdade. Essa igualdade também é consequência do fato que o valor de C_t na expressão acima, a constante de normalização de k, para cada t, em (4.18)*, é igual a*

$$C_t = \sqrt{2\pi}e^{-\frac{|r_t|}{\delta}}. \tag{4.19}$$

Este resultado deriva do fato que

$$\mathcal{L}\left[\frac{e^{-a/s}}{s^{3/2}}\right](x) = \sqrt{\frac{\pi}{a}}e^{-2\sqrt{az}}$$

para $s > 0$ e para todo $a > 0$. Veja [25] para mais detalhes sobre as transformadas de Laplace utilizadas. Note que, quanto mais discrepante o valor de y_t em relação ao valor de $f(x_t)$ (ou, como é na prática, em relação à estimativa da função alvo calculada no ponto x_t), menor é a constante de padronização C_t e menor é o peso $\bar{\sigma}_t$ associado a esta observação. ◊

Note que, nos exemplos acima, o peso associado associado a t-ésima observação, ie, a componente (t, t) da matriz de ponderação $W_T(\boldsymbol{\theta})$, é função decrescente do resíduo. Ou seja, quanto maior o valor $|y_t - f(x_t)|$, menor o valor da respectiva componente em $W_T(\boldsymbol{\theta})$.

Exemplo 4.2.4 (Distribuição Logística). *No caso da distribuição logística, a densidade de mistura é dada por*

$$h(\sigma) = \sum_{k=1}^{\infty}(-1)^{k-1}k^2\sigma^{-2}e^{-k/2\sigma}$$

e, assim como nos exemplos anteriores, teremos

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}') \equiv -T\log\delta - \frac{1}{2\delta^2}(\mathbf{y} - \mathbf{f})'W_T(\boldsymbol{\theta}')(\mathbf{y} - \mathbf{f}),$$

e

$$\bar{\sigma}_t = E\{\sigma_t|y_t\}$$

de modo que $W_T(\boldsymbol{\theta}') = diag(\bar{\sigma}_1, ..., \bar{\sigma}_T)$ onde

$$\begin{aligned} k(\sigma_t|y_t, \boldsymbol{\theta}') &\propto \frac{\sigma_t^{\frac{1}{2}}}{\delta}e^{-\frac{\sigma_t r_t^2}{2\delta^2}}\sum_{k=1}^{\infty}(-1)^{k-1}k^2\sigma_t^{-2}e^{-k/2\sigma_t} \\ &\propto \sum_{k=1}^{\infty}(-1)^{k-1}k^2\sigma_t^{-3/2}e^{-\frac{1}{2}\left(\frac{k}{\sigma_t}+\frac{r_t^2\sigma_t}{\delta^2}\right)}, \end{aligned}$$

onde $r_t = y_t - f(x_t)$. *Logo*

$$\bar{\sigma}_t = C_t^{-1} \sum_{k=1}^{\infty} (-1)^{k-1} k^2 \int_0^{\infty} \frac{e^{-\frac{1}{2}\left(\frac{k}{\sigma_t} + \frac{r_t^2 \sigma_t}{\delta^2}\right)}}{\sigma_t^{1/2}} d\sigma_t,$$

onde C_t *é a constante de normalização associada à densidade* k. *Assim como para a densidade exponencial dupla, podemos mostrar que*

$$\int_0^{\infty} \frac{e^{-\frac{1}{2}\left(\frac{k}{\sigma_t} + \frac{r_t^2 \sigma_t}{\delta^2}\right)}}{\sigma_t^{1/2}} d\sigma_t = \mathcal{L}\left[\frac{e^{-\frac{k}{2s}}}{\sqrt{s}}\right]\left(\frac{r_t^2}{2\delta^2}\right) = \frac{\sqrt{2\pi}\delta}{|r_t|} e^{-\frac{\sqrt{k}|r_t|}{\delta}}$$

e

$$\int_0^{\infty} \frac{e^{-\frac{1}{2}\left(\frac{k}{\sigma_t} + \frac{r_t^2 \sigma_t}{\delta^2}\right)}}{\sigma_t^{3/2}} d\sigma_t = \mathcal{L}\left[\frac{e^{-\frac{k}{2s}}}{s^{3/2}}\right]\left(\frac{r_t^2}{2\delta^2}\right) = \sqrt{\frac{2\pi}{k}} e^{-\frac{\sqrt{k}|r_t|}{\delta}}$$

de modo que

$$\bar{\sigma}_t = \frac{\delta}{|r_t|} \frac{\sum_{k=1}^{\infty} (-1)^{k-1} k^2 e^{-\frac{\sqrt{k}|r_t|}{\delta}}}{\sum_{k=1}^{\infty} (-1)^{k-1} k^{3/2} e^{-\frac{\sqrt{k}|r_t|}{\delta}}}.$$

◊

Exemplo 4.2.5 (Distribuição Normal Contaminada). *No caso da distribuição normal contaminada, a densidade de mistura é dada por*

$$h(\sigma) = \begin{cases} 1 - \xi, & \text{se } \sigma = 1 \\ \xi, & \text{se } \sigma = \lambda^2 \\ 0, & \text{caso contrário} \end{cases}$$

onde $0 < \xi < 1$. *Lembre que, neste caso,*

$$\epsilon_t \sim (1 - \xi)\mathcal{N}(0, 1) + \xi \mathcal{N}\left(0, \frac{1}{\lambda^2}\right).$$

Podemos generalizar o conceito acima para permitir mais de uma fonte de contaminação supondo

$$h(\sigma) = \begin{cases} \pi_0, & \text{se } \sigma = 1 \\ \pi_j, & \text{se } \sigma = \lambda_j^2 \text{ para } j = 1, ..., K \\ 0, & \text{caso contrário} \end{cases}$$

onde $\pi_0, ..., \pi_K > 0$ *e* $\sum_{j=0}^{K} \pi_j = 1$. *Neste caso, os dados são modelados de acordo com*

$$\begin{cases} y_t|\sigma_t \sim \mathcal{N}\left(f(x_t), \frac{\delta^2}{\sigma_t}\right) \\ p(\sigma_t = \lambda_j^2) = \pi_j \end{cases}$$

para $j = 0, 1, ..., K$, *onde* $\lambda_0 = 1$.

Novamente,

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}') \equiv -T \log \delta - \frac{1}{2\delta^2}(\mathbf{y} - \mathbf{f})' W_T(\boldsymbol{\theta}')(\mathbf{y} - \mathbf{f}).$$

e

$$\bar{\sigma}_t = E\{\sigma_t|y_t\}$$

de modo que $W_T(\boldsymbol{\theta}') = diag(\bar{\sigma}_1, ..., \bar{\sigma}_T)$ *onde*

$$k(\sigma_t = \lambda_j^2|y_t, \boldsymbol{\theta}') \propto \frac{\lambda_j}{\delta}\phi\left(\frac{r_t}{\delta/\lambda_j}\right)\pi_j,$$

onde $r_t = y_t - f(x_t)$. *Obviamente, a constante de normalização, neste caso, é dada por*

$$C_t = \sum_{j=0}^{K} \frac{\lambda_j}{\delta}\phi\left(\frac{r_t}{\delta/\lambda_j}\right)\pi_j.$$

Logo, definindo para cada $j = 0, ..., K$, $\pi_{j|t} = \frac{k(\sigma_t = \lambda_j^2|y_t, \boldsymbol{\theta}')}{C_t}$, *teremos que*

$$\bar{\sigma}_t = \sum_{j=0}^{K} \pi_{j|t}\lambda_j.$$

◊

Agora, se aproximarmos f por $\sum_{j=1}^{M} c_j B_j$, onde $\mathbf{c} = (c_1, ..., c_M)'$ e $\mathcal{B} = \{B_j\}_{j=1}^{\infty}$ são como na seção anterior, então

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}') = C(\boldsymbol{\zeta}; \boldsymbol{\theta}') - T \log \delta - \frac{1}{2\delta^2}(\mathbf{y} - \mathbf{Bc})' W_T(\boldsymbol{\theta}')(\mathbf{y} - \mathbf{Bc}). \tag{4.20}$$

Note que, neste caso, $\boldsymbol{\theta} = (\mathbf{c}', \boldsymbol{\zeta}')'$. Em particular, tomando-se o gradiente de Q com relação a $\mathbf{c}$, temos

$$\frac{\partial}{\partial \mathbf{c}} Q(\boldsymbol{\theta}|\boldsymbol{\theta}') = \frac{1}{\delta^2}\left(\mathbf{B}'\mathbf{y} - \mathbf{B}' W_T(\boldsymbol{\theta}')\mathbf{Bc}\right).$$

Logo, dado $\boldsymbol{\theta}'$, a estimativa de $\mathbf{c}$ é

$$\widehat{\mathbf{c}} = (\mathbf{B}' W_T(\boldsymbol{\theta}')\mathbf{B})^{-1}\mathbf{B}' W_T(\boldsymbol{\theta}')\mathbf{y}.$$

Quanto ao parâmetro de escala, temos

$$\frac{\partial}{\partial \delta} Q(\boldsymbol{\theta}|\boldsymbol{\theta}') = -\frac{T}{\delta} + \frac{1}{\delta^2}(\mathbf{y} - \mathbf{Bc})' W_T(\boldsymbol{\theta}')(\mathbf{y} - \mathbf{Bc})$$

o que resulta no estimador

$$\widehat{\delta}^2 = \frac{1}{T}(\mathbf{y} - \mathbf{Bc})' W_T(\boldsymbol{\theta}')(\mathbf{y} - \mathbf{Bc}). \tag{4.21}$$

Finalmente, assumindo que ψ independe de $\boldsymbol{\zeta}$, temos que a estimativa de $\boldsymbol{\zeta}$ é a solução da equação

$$\frac{\partial}{\partial \boldsymbol{\zeta}} Q(\boldsymbol{\theta}|\boldsymbol{\theta}') = \frac{\partial}{\partial \boldsymbol{\zeta}} C(\boldsymbol{\zeta}; \boldsymbol{\theta}') = 0.$$

4.2.3 Critério de Parada

Dado que o objetivo é estimar a função alvo, podemos estabelecer como critério de parada com base no erro quadrático médio

$$\text{EQM}(\widehat{f}) = \frac{1}{T}\sum_{t=1}^{T}(y_t - \widehat{f}(x_t))^2.$$

De fato, podemos inferir que a estimativa da função alvo convergiu para uma determinada curva quando a média da soma dos quadrados das distâncias entre a curva estimada e os pontos observados se mantiverem constantes. Em termos práticos, isto se traduz no seguinte critério: devemos parar as iterações do algoritmo quando $|\text{EQM}(\widehat{f}^{(k+1)}) - \text{EQM}(\widehat{f}^{(k)})| < \text{tol}$, onde tol é um nível de tolerância pré-especificado e $\widehat{f}^{(k)}$ é a estimativa da função alvo na k-ésima iteração do algoritmo.

A metodologia sugerida pode ser vista resumidamente no algoritmo 4.2.1.

4.2.4 Aproximação de f via Smoothing-Splines

Uma alternativa para a aproximação de f que surge naturalmente é o uso de *smoothing-splines*. Isto ocorre por duas razões, sendo a primeira o simples fato que, restritos ao espaço de Sobolev $\mathcal{W}$ das funções com segunda derivada definida, a solução do problema de se minimizar o funcional

$$\sum_{t=1}^{T}\omega_t(y_t - f(x_t))^2 + J_\lambda(f), \tag{4.22}$$

onde $(\omega_1, ..., \omega_T)'$ é um vetor de ponderações e onde a função de penalização é dada por

$$J_\lambda(f) \equiv \lambda \int [f''(t)]^2 dt, \tag{4.23}$$

é um spline cúbico cujos nós são determinados pelas observações $\{x_t\}_{t=1}^T$, veja seção 2.1.4. Em particular, no nosso caso,

$$\omega_t = \frac{1}{2\delta^2} E_{\boldsymbol{\theta}}\left\{\frac{1}{\psi(\sigma_t)}\middle| z_t\right\}.$$

A segunda razão que nos leva a considerar a técnica de *smoothing-splines* é o fato que para a função de penalização J_λ definida em (4.23), a solução do problema de otimização:

$$\text{argmin}_{f\in\mathcal{W}} \sum_{t=1}^{T} l_t(y_t - f(x_t))^2 + J_\lambda(f)$$

Algoritmo 4.2.1 Resumo do Algoritmo de Estimação.

Passo 1. obter uma estimativa inicial para $\mathbf{f} = (f(x_1), ..., f(x_T))$, denotada por $\mathbf{f}^{(0)}$;

- um modo de se obter a estimativa inicial é através do método de estimação usual associado ao suavizador escolhido, ie, supondo o ruído i.i.d. e gaussiano.

Passo 2. obter uma estimativa inicial para o parâmetro de escala δ e para ζ, denotados por $\delta^{(0)}$ e $\zeta^{(0)}$, respectivamente;

Passo 3. *loop* principal do algoritmo EM.

- Enquanto $|\text{EQM}(\widehat{f}^{(k)}) - \text{EQM}(\widehat{f}^{(k-1)})| \geq \text{tol}$:

  Passo $(k+1)$**.1:** $\mathbf{f}^{(k+1)} = \mathbf{B}\mathbf{c}^{(k+1)}$, onde

  $$\mathbf{c}^{(k+1)} = (\mathbf{B}'W_T(\boldsymbol{\theta}^{(k)})\mathbf{B})^{-1}\mathbf{B}'W_T(\boldsymbol{\theta}^{(k)})\mathbf{y};$$

  Passo $(k+1)$**.2:** atualizar a $(k+1)$-ésima estimativa de δ^2 por

  $$\frac{1}{T}(\mathbf{y} - \mathbf{f}^{(k+1)})'W_T(\mathbf{c}^{(k+1)}, \delta^{(k)}, \zeta^{(k)})(\mathbf{y} - \mathbf{f}^{(k+1)});$$

  Passo $(k+1)$**.3:** atualizar a $(k+1)$-ésima estimativa de ζ através da equação

  $$\frac{\partial}{\partial \zeta} C(\zeta; \mathbf{c}^{(k+1)}, \delta^{(k+1)}, \zeta^{(k)}) = 0.$$

onde, para $t = 1, ...T$, l_t é uma função de log-verossimilhança, é um *spline* cúbico. Este fato segue das proposições 2.1.1 e 2.1.2.

Lembre que, na k-ésima iteração do algoritmo EM, o estimador de f é aquele que minimiza, junto com o vetor de parâmetros ζ associado a h, a função critério (4.10). Se mantivermos ζ fixado, temos que o estimador de f é exatamente aquele que minimiza (4.22), cuja solução é obtida através das equações normais

$$(X'W_TX + T\lambda\delta^2\Omega)\mathbf{c} = X'W_T\mathbf{y}, \tag{4.24}$$

onde $X = \{\beta_j(x_i)\}_{i,j=1}^T$ e $\beta_1, ..., \beta_T$ é uma base para o conjunto de *splines* naturais de ordem 4 com nós em $x_1, ..., x_T$ e a matriz Ω é dada por

$$\Omega = \left\{ \int \beta_i''(t)\beta_j''(t)dt \right\}_{i,j=1}^T. \tag{4.25}$$

Da expressão (4.21), a estimativa do parâmetro de escala δ é dada por

$$\widehat{\delta} = \frac{1}{T}(\mathbf{y} - X\mathbf{c})'W_T(\mathbf{y} - X\mathbf{c}).$$

4.2.5 Aproximação via B-Splines

Estimar a função alvo f por *smoothing splines* é um meio de se trazer o problema para dimensões finitas. No entanto, ainda assim, o esforço computacional é relativamente elevado dado que o total de parâmetros é, no mínimo igual à quantidade de observações. Um modo de tornar o processo de estimação computacionalmente mais barato é considerar a aproximação de f via B-*splines* dada por

$$f(x) \approx \sum_{j=1}^{M} c_j B_{m,j}(x)$$

onde as funções $B_{m,j}(x)$ são *B-splines* de ordem m, para $j = 1, ..., M$. Deste modo, nosso objetivo passa a ser determinar as estimativas $\widehat{c}_1, ..., \widehat{c}_M$ dos coeficientes $c \equiv (c_1, ..., c_M)'$ e, conseqüentemente, estimar f por

$$\widehat{f}(x) = \sum_{j=1}^{M} \widehat{c}_j B_{m,j}(x).$$

Logo, neste caso, o vetor de parâmetros é dado por $\boldsymbol{\theta} = (c', \boldsymbol{\zeta}')'$.

Assim como na seção 4.2.4, nosso objetivo continua a ser estimar f de modo a minimizar o funcional (4.22). Para determinar as equações normais neste caso, basta notar que o vetor de aproximações via B-*splines* de f nos pontos $(x_1, ...,_T)$ é dado por

$$\bar{\mathbf{f}} = (\bar{f}(x_1), .., \bar{f}(x_T)) = B_M \mathbf{c},$$

onde $\mathbf{c} = (c_1, ..., c_M)'$ e $B_M \equiv (B_{m,j}(x_i))_{i,j}$ é a matriz $T \times M$ formada pelos elementos da base de *B-splines* calculados nos pontos $x_1, ..., x_T$, e que, substituindo $f''(t)$ em (4.23) pela segunda derivada da aproximação de f via *B-splines*, temos

$$\begin{aligned} J_\lambda(f) &= \lambda \sum_{j,k=1}^{M} c_j c_k \int (B''_{m,j}(t) B''_{m,k}(t)) dt \\ &= \lambda \mathbf{c}' \Omega_M \mathbf{c}, \end{aligned}$$

onde

$$\Omega_M \equiv (\int (B''_{m,j}(t) B''_{m,k}(t)) dt)_{1 \leq j,k \leq M}. \tag{4.26}$$

De fato, como na seção 4.2.4, na k-ésima iteração do algoritmo EM, se mantivermos $\boldsymbol{\zeta}$ fixado, temos que o estimador de f é obtido através da minimização de

$$\frac{1}{2\delta^2}(\mathbf{y} - B_M \mathbf{c})' W_T (\mathbf{y} - B_M \mathbf{c}) + \mathbf{c}' \Omega_M \mathbf{c}$$

de modo que as equações normais ficam dadas por

$$(B'_M W_T B_M + 2\lambda \delta^2 \Omega)\mathbf{c} = B'_M W_T \mathbf{y}.$$

Assim como para *smoothing-splines*, o parâmetro de escala é estimado por

$$\widehat{\delta} = \frac{1}{T}(\mathbf{y} - B_M \mathbf{c})' W_T (\mathbf{y} - B_M \mathbf{c}).$$

4.2.6 Bandas de Confiança via Bootstrap

Uma maneira de se obter bandas de confiança para a estimativa da função alvo, do parâmetro de escala e do vetor de parâmetros ζ é através da técnica *bootstrap*. Uma amostra *bootstrap* é obtida via monte-carlo amostrando-se o ruído ϵ T vezes para, com o resultado, construir uma nova amostra de variáveis resposta. Abaixo, descrevemos duas metodologias *bootstrap* distintas e conhecidas por *bootstrap* paramétrico e *bootstrap* não paramétrico, respectivamente. A primeira metodologia leva em consideração a distribuição assumida para o ruído, a qual, no nosso caso, é uma mistura na escala de normais, enquanto que a segunda utiliza a distribuição empírica dos resíduos. Embora estejamos assumindo uma determinada distribuição para o ruído, na prática, é interessante o uso do *bootstrap* não paramétrico.

Bootstrap Paramétrico

Para tanto, consideramos o fato que

$$\epsilon_t|\sigma_t \sim \mathcal{N}(0, \psi(\sigma_t)^2),$$
$$\sigma_t \sim h.$$

Neste caso, podemos obter por simulação os valores $\sigma_1^*, ..., \sigma_T^*$ de acordo com a densidade h e depois, para cada $t = 1, ..., T$, obter os valores $\epsilon_1^*, ..., \epsilon_T^*$ de acordo com a distribuição $\mathcal{N}(0, \psi(\sigma_t^*)^2)$. O passo seguinte consiste em definir

$$y_t^* \equiv \widehat{f}(x_t) + \widehat{\delta}\epsilon_t^*, \text{ para } t = 1, ..., T$$

onde $\widehat{f}$ e δ são as estimativas de f e δ, respectivamente, obtidas com a amostra original $y_1, ..., y_T$. A estimativa *bootstrap* de f é, então, obtida usando-se esta nova amostra $y_1^*, ..., y_T^*$.

O procedimento acima é repetido B vezes e, para cada amostra *bootstrap*, obtemos uma estimativa *bootstrap* $\widehat{\mathbf{f}}_j^*$ e, com estas estimativas, construímos as bandas de confiança desejadas.

Bootstrap Não-Paramétrico

O procedimento não-paramétrico é análogo ao descrito acima, com a diferença de que a amostragem agora é feita sem assumir nenhuma distribuição específica para o ruído. Ou seja, ϵ_t^* é amostrado via amostragem simples com reposição a partir da coleção de resíduos obtida após o ajuste inicial do modelo.

Observação 4.2.1. *Como notado em [28] e nas referências lá contidas, na hipótese do ruído possuir variância infinita, tanto para modelos de regressão quanto para séries temporais, os métodos usuais de bootstrap, como os descritos acima, falham drasticamente caso a condição $B = o(T)$ seja violada. Em particular, no nosso caso, este cuidado deve ser tomado quando assumirmos uma distribuição estável para o ruído.*

4.3 Aproximação de $S(\theta|\theta')$ via Monte-Carlo

Como descrito na seção 2.2.5, quando a esperança $Q(\theta|\theta^{(j)})$ não pode ou é difícil de ser calculada analiticamente, uma possibilidade é aproximá-la via Monte-Carlo. Para tanto, seja $\sigma_{t,1}^{(j)}, ..., \sigma_{t,J(j)}^{(j)}$ uma amostra pseudo

aleatória da variável σ_t obtida segundo a densidade condicional (4.11) com $\theta^{(j)}$ no lugar de θ', e considere as seguintes aproximações do vetor de ponderações W e de $C(\zeta;\theta')$, respectivamente,

$$\widehat{W}_j \equiv \text{diag}\left\{\sum_{i=1}^{J(j)} \frac{1}{\psi_\zeta(\sigma_{1,i}^{(j)})^2}, ..., \sum_{i=1}^{J(j)} \frac{1}{\psi_\zeta(\sigma_{T,i}^{(j)})^2}\right\}. \tag{4.27}$$

e

$$\widehat{C}(\zeta;\theta^{(j)}) = \sum_{t=1}^{T}\sum_{i=1}^{J(j)} \log \frac{h(\sigma_{t,i}^{(j)}|\zeta)}{\psi_\zeta(\sigma_{t,i}^{(j)})}. \tag{4.28}$$

Logo, a função critério (4.10) pode ser aproximada por

$$\widehat{S}(\theta|\theta^{(j)}) = \widehat{C}(\zeta;\theta^{(j)}) - T\log\delta - \frac{1}{2}(\mathbf{y}-\bar{\mathbf{f}}_\theta)'\widehat{W}_j(\mathbf{y}-\bar{\mathbf{f}}_\theta) + J_\lambda(\theta), \tag{4.29}$$

onde $\bar{\mathbf{f}}_\theta$ é o vetor de aproximações de f nos pontos $(x_1,...,x_T)'$ com base nos parâmetros em θ.

Exemplo 4.3.1 (Distribuição t de Student — ν Conhecido). *No caso da distribuição t, temos*

$$\widehat{S}(\theta|\theta^{(j)}) = -T\log\delta - \frac{\kappa_j}{2\delta^2}$$

onde $\kappa_j \equiv \sum_{i=1}^{T}\bar{\sigma}_i^{(j)}(y_i - f(x_i))^2$ *e onde* $\bar{\sigma}_t \equiv E_{\theta^{(j)}}\{\sigma_t|y_t\}$. *Note que* $\kappa_j \geq 0$. *Derivando* $\widehat{S}_r$ *e igualando o resultado a zero, obtemos o estimador de* ζ *para* j*-ésima iteração do algoritmo EM,*

$$\widehat{\delta}_j = \sqrt{\frac{\kappa_j}{T}}.$$

Observe também que $\widehat{\delta}_j$ *é o estimador de máxima verossimilhança para o parâmetro de escala* δ *ponderado pelos pesos* $\bar{\sigma}_1^{(j)},...,\bar{\sigma}_T^{(j)}$. ◊

4.3.1 Smoothing Splines

Substituindo W_T por $\widehat{W}_j$ em (4.24), o $(j+1)$-ésimo passo do algoritmo EM aproximado resulta nas estimativas que maximizam

$$\widehat{S}(\theta|\theta^{(j)}) = \widehat{C}(\zeta;\theta^{(j)}) - \frac{1}{2}(\mathbf{y}-X\mathbf{c})'\widehat{W}_j(\mathbf{y}-X\mathbf{c}) - \lambda\mathbf{c}'\Omega\mathbf{c}, \tag{4.30}$$

onde Ω é dado por (4.25). O procedimento de maximização é quebrado em três passos, como mostrado no algoritmo 4.3.1.

Algoritmo 4.3.1 Procedimento de maximização quando usamos *smoothing splines*.

Passo 1. Manter $\zeta = \zeta^{(j)}$, maximizar $\widehat{S}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(j)})$ com relação a $\mathbf{c}$. O estimador resultante, $\mathbf{c}^{(j+1)}$ é a solução do sistema de equações

$$(X'\widehat{W}_j X + T\lambda\Omega)\mathbf{c} = X'\mathbf{y}.$$

Passo 2. Usar a estimativa $\mathbf{c}^{(j+1)}$,

$$\delta^{(j+1)} = \frac{1}{T}(\mathbf{y} - X\mathbf{c}^{(j+1)})'W_j(\mathbf{y} - X\mathbf{c}^{(j+1)})$$

Passo 3. Usar as estimativas $\mathbf{c}^{(j+1)}$ e $\delta^{(j+1)}$ para estimar $\zeta^{(j+1)}$ maximizando-se a função critério atualizada $\widehat{S}(\boldsymbol{\theta}|\mathbf{f}^{(j+1)}, \delta^{(j+1)}, \zeta^{(j)})$.

4.3.2 B-Splines

O algoritmo de estimação de f e ζ via B-*splines* é perfeitamente análogo ao obtido via *smoothing-splines*. De fato, substituindo W_T por $\widehat{W}_j$ em (4.24), o $(j+1)$-ésimo passo do algoritmo EM aproximado consiste em se maximizar com relação a $\mathbf{c}$ e ζ

$$\widehat{S}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(j)}) = \widehat{C}(\zeta; \boldsymbol{\theta}^{(j)}) - T\log\delta - \frac{1}{2\delta^2}(\mathbf{y} - B_M\mathbf{c})'\widehat{W}_j(\mathbf{y} - B_M\mathbf{c}) - \lambda\mathbf{c}'\Omega_M\mathbf{c}, \tag{4.31}$$

onde Ω é dado por (4.26). Novamente, o procedimento de maximização é quebrado em 3 passos como ilustrado no algoritmo 4.3.2.

Algoritmo 4.3.2 Procedimento de maximização quando usamos B-*splines*

Passo 1. Manter $\zeta = \zeta^{(j)}$, maximizar $\widehat{S}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(j)})$ com relação a $\mathbf{c}$. O estimador resultante, $\mathbf{c}^{(j+1)}$ é a solução do sistema de equações

$$(B_M'\widehat{W}_j B_M + 2\lambda\delta^2\Omega)\mathbf{c} = B_M'\widehat{W}_j\mathbf{y}.$$

Passo 2. Usar a estimativa $\mathbf{c}^{(j+1)}$,

$$\delta^{(j+1)} = \frac{1}{T}(\mathbf{y} - B_M\mathbf{c}^{(j+1)})'W_j(\mathbf{y} - B_M\mathbf{c}^{(j+1)})$$

Passo 3. Usar a estimativa $\mathbf{c}^{(j+1)}$ para estimar $\zeta^{(j+1)}$ maximizando-se a função critério atualizada $\widehat{S}(\boldsymbol{\theta}|\mathbf{f}^{(j+1)}, \zeta^{(j)})$.

Observação 4.3.1. (Maximização com relação a ζ — parte 1) *Note que as componentes* $\lambda\mathbf{c}'\Omega\mathbf{c}$ *e* $T\log\delta$ *de* $\widehat{S}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(j)})$ *independem de* ζ *e, portanto, podemos apenas considerar*

$$\widehat{S}_r(\boldsymbol{\theta}|\boldsymbol{\theta}^{(j)}) = \widehat{C}(\zeta; \boldsymbol{\theta}^{(j)}) - \frac{1}{2}(\mathbf{y} - X\mathbf{c})'\widehat{W}_j(\mathbf{y} - X\mathbf{c}) \tag{4.32}$$

ao estimar ζ. ◊

Observação 4.3.2. (Maximização com relação a ζ — parte 2) *O vetor de parâmetros ζ tem, em geral, dimensão pequena e a maximização com relação aos elementos deste vetor pode ser quebrada em uma pequena seqüência de maximizações univariadas,*

$$\begin{aligned}
\delta_1^{(j+1)} &= argmax_{\delta_1} \widehat{S}(\delta_1; \mathbf{f}^{(j+1)}, \delta_2^{(j)}, ..., \delta_p^{(j)}) \\
\delta_2^{(j+1)} &= argmax_{\delta_2} \widehat{S}(\delta_2; \mathbf{f}^{(j+1)}, \delta_1^{(j+1)}, \delta_2, ..., \delta_p^{(j)}) \\
&\vdots \\
\delta_p^{(j+1)} &= argmax_{\delta_p} \widehat{S}(\delta_p; \mathbf{f}^{(j+1)}, \delta_1^{(j+1)}, \delta_2^{(j+1)}, ..., \delta_p),
\end{aligned} \tag{4.33}$$

onde $\zeta = (\delta_1, ..., \delta_p)'$ *e onde usamos* (4.32) *no lugar da função critério original.* ◊

4.3.3 Algoritmo Genérico de Estimação

No algoritmo 4.3.3 descrevemos, de modo um pouco mais conciso, os passos do procedimento de estimação da função alvo f e do vetor de parâmetros ζ.

Caso as esperanças do passo E do algoritmo e a distribuição condicional de σ dado y possam ser calculadas analiticamente, os passos $k.1$ a $k.5$ podem ser substituídos pelos passos E e M usuais do algoritmo, ie, sem simulações de Monte Carlo.

4.3.4 Sobre a Implementação do Algoritmo de Metropolis-Hastings

O objetivo é amostrar a partir da densidade condicional de σ, logo usando o fato de que esta densidade tem suporte em $[0, \infty)$ e que ela é o produto da densidade marginal $h(\cdot|\theta)$ com a densidade da normal padrão calculada em um determinado ponto (ver abaixo), podemos tomar como densidade proposta a própria densidade marginal h[4], ie,

$$q(\sigma_t^{(j)}, s) \equiv h(s|\theta),$$

[4]estamos obviamente assumindo que h é de fato uma densidade de probabilidades, o que, no caso de misturas na escala de normais, não é sempre verdade

Algoritmo 4.3.3 Algoritmo genérico de estimação para o modelo semiparamétrico

Passo 1. obter uma estimativa inicial para $\mathbf{f} = (f(x_1), ..., f(x_T))$, denotada por $\mathbf{f}^{(0)}$.

- um modo de se obter a estimativa inicial é através do método de estimação usual associado ao suavizador escolhido, ie, supondo o ruído i.i.d. e gaussiano.

Passo 2. obter uma estimativa inicial para ζ, denotada por $\zeta^{(0)}$

Passo 3. iniciar o *loop* principal do algoritmo EM.

k-ésima iteração: dados $\mathbf{f}^{(k-1)}$ e $\zeta^{(k-1)}$:

Passo k.1: para cada $t = 1, ..., T$ gerar uma amostra pseudo-aleatória, através do algoritmo de Gibbs ou via Metropolis-Hastings, de $\sigma_t^{(1)}, ..., \sigma_t^{(J(j))}$ de acordo com a distribuição

$$\bar{k}(\sigma_t|y_t, \boldsymbol{\theta}^{(k-1)}) \propto \frac{1}{\delta^{(k-1)}\psi_{\boldsymbol{\theta}^{(k-1)}}(\sigma_t)} \cdot \phi\left(\frac{y_t - f_t^{(k-1)}}{\delta^{(k-1)}\psi_{\boldsymbol{\theta}^{(k-1)}}(\sigma_t)}\right) h(\sigma_t|\boldsymbol{\theta}^{(k-1)})$$

onde $f_t^{(k-1)}$ é a aproximação obtida para f no ponto x_t na iteração $(k-1)$.

Passo k.2: calcular $\widehat{W}_k$ de acordo com (4.27) e com $\zeta = \zeta^{(k-1)}$;

Passo k.3: calcular $\mathbf{f}^{(k)}$ através das equações normais apropriadas;

Passo k.4: calcular $\zeta^{(k)}$ usando o resultado do passo anterior;

Passo k.5: se $|L(\boldsymbol{\theta}^{(k)}) - L(\boldsymbol{\theta}^{(k-1)})| < \varepsilon$, terminar o algoritmo. Caso contrário, ir para o passo $(k+1)$.**1**;

---

onde $\boldsymbol{\theta}$ é o vetor de parâmetros do modelo estimado na última iteração, digamos a iteração de número k, do algoritmo EM. É fácil ver que, neste caso, a probabilidade de aceitação $\rho(\sigma_t^{(j)}, s)$ é dada por

$$\begin{aligned}\rho(\sigma_t^{(j)}, s) &= \min\left\{\frac{\psi(\sigma_t^{(j)})}{\psi(s)} \frac{\phi\left(\frac{y_t - f^{(k)}(x_t)}{\delta^{(k)}\psi(s)}\right)}{\phi\left(\frac{y_t - f^{(k)}(x_t)}{\delta^{(k)}\psi(\sigma_t^{(j)})}\right)}, 1\right\} \\ &= \min\left\{\frac{\psi(\sigma_t^{(j)})}{\psi(s)} \exp\left[-\frac{(y_t - f^{(k)}(x_t))^2}{2(\delta^{(k)})^2}\left(\frac{1}{\psi(s)^2} - \frac{1}{\psi(\sigma_t^{(j)})^2}\right)\right], 1\right\},\end{aligned}$$

de modo que a amostra de interesse é gerada de acordo com a regra

$$\sigma_t^{(j+1)} = \begin{cases} s & , \text{ com probabilidade } \rho(\sigma_t^{(j)}, s) \\ \sigma_t^{(j)} & , \text{ com probabilidade } 1 - \rho(\sigma_t^{(j)}, s) \end{cases}.$$

Em particular, o algoritmo acima pode ser utilizada para modelos cujos ruídos são distribuidos de acordo

com uma distribuição estável, pois, devemos apenas ser capazes de gerar uma amostra pseudo-aleatória segundo esta distribuição e, desta forma, evitamos as dificuldades oriundas da maximização da respectiva função de verossimilhança.

Exemplo 4.3.2 (Distribuição t de Student — ν Conhecido). *Embora, no caso em que o ruído segue uma distribuição t de Student não seja necessário recorrer a simulações para obter as esperanças do passo E do algoritmo, temos que, neste caso, $\psi(\sigma) = \zeta/\sqrt{\sigma}$, de modo que a probabilidade de aceitação ficaria igual a*

$$\rho(\sigma_t^{(j)}, s) = \min\left\{\sqrt{\frac{s}{\sigma_t^{(j)}}}\exp\left[-\frac{(y_t - f^{(k)}(x_t))^2}{2\zeta^2}(s - \sigma_t^{(j)})\right], 1\right\}$$

◊

4.4 Estudo de Simulação

Para este estudo, usamos as seguintes funções alvo:

i. **Função alvo 1:** (figura 4.1 (a))

$$f(x) = \text{sen}(2x) \tag{4.34}$$

ii. **Função alvo 2:** (figura 4.1 (b))

$$f(x) = 10x + 15\exp(-12x^2) \tag{4.35}$$

iii. **Função alvo 3:** (figura 4.1 (c))

$$f(x) = \frac{\text{sen}(20(x + 0.2))}{x + 0.2} \tag{4.36}$$

4.4.1 Simulação e estimação via B-Splines

Nesta seção, avaliamos via simulação a metodologia sugerida no caso em que a aproximação da função alvo é feita através de *B-splines*. Neste primeiro estudo consideramos a função alvo

$$f(x) = \text{sen}(2x)$$

(figura 4.1 (a)) e assumimos como parâmetro de escala $\delta = 1,5$. Os dados foram obtidos via simulação assumindo-se que o ruído seguia uma distribuição t de Student com $1,5$ graus de liberdade de modo a gerar uma amostra de tamanho 200. Foram realizados 3 ajustes, sendo que, para todos eles, consideramos $M = 3$ e obtivemos amostras *bootstrap* de tamanho 100 com base nas quais construímos intervalos de confiança (95%)

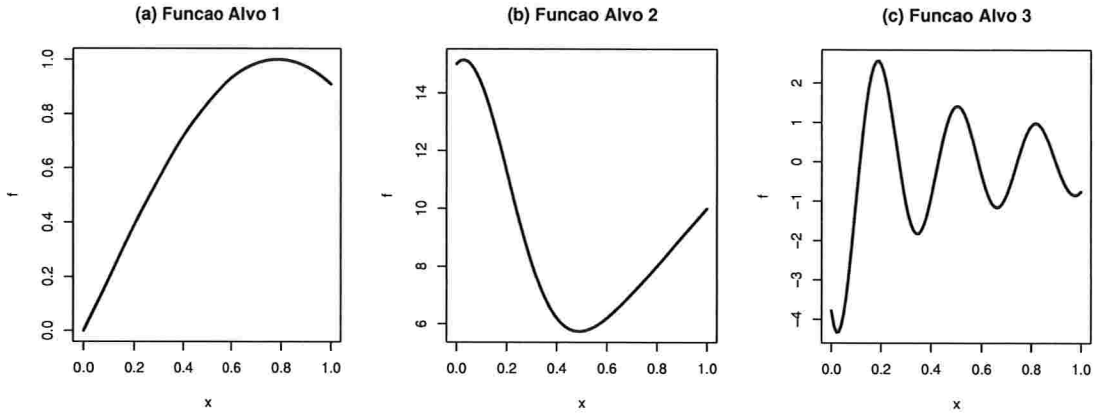


Figura 4.1: Funções alvo utilizadas nos estudos de simulação para o modelo de regressão semiparamétrico sugerido.

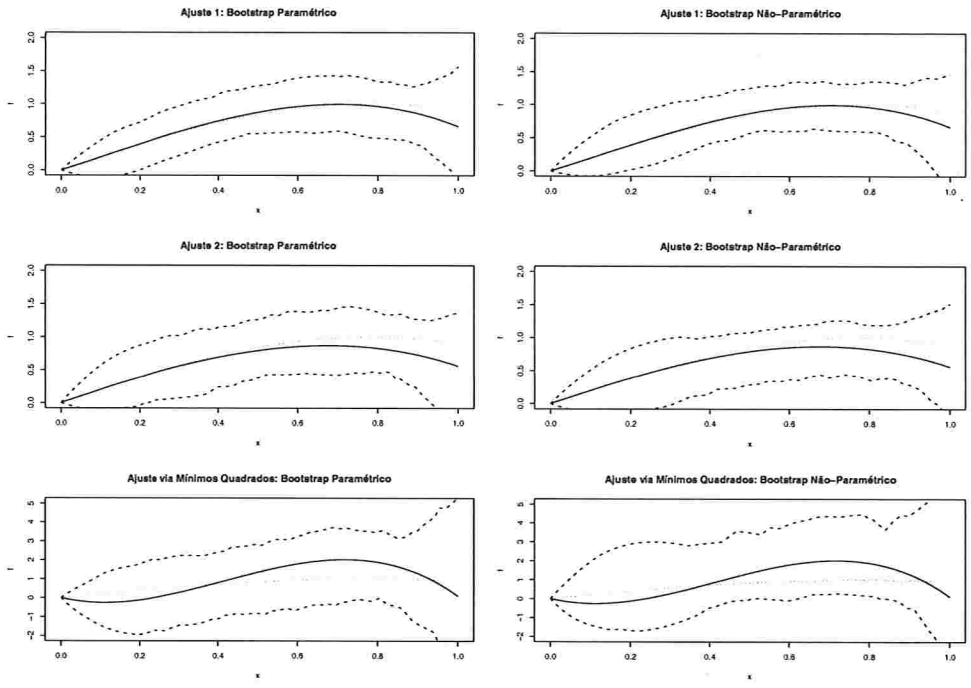


Figura 4.2: Estimativas da função alvo 1 segundo a metodologia proposta, Ajuste 1 (gráficos superiores — correspondem a $\nu = 1,5$) e Ajuste 2 (gráficos intermediários — corresponde a $\nu = 3,5$), e via mínimos quadrados, Ajuste 3 (gráficos inferiores).

para δ e bandas de confiança (95%) para as estimativas da função alvo. O primeiro ajuste (Ajuste 1) foi feito segundo a metodologia proposta na tese e tomando $\nu = 1,5$, o segundo ajuste também foi realizado segundo a metodologia proposta, porém, com $\nu = 3,5$ e, finalmente, o terceiro ajuste (Ajuste 3) foi feito via mínimos quadrados. O resultado, relativo à estimativa da função alvo, deste estudo pode ser visto na figura 4.2. Quanto ao parâmetro de escala, obtivemos os resultados expostos na tabela 4.1. Note que o erro quadrático médio é, de fato, mais baixo quando tomamos o valor correto para ν. No entanto, não há uma alteração significativa se o subestimamos, isto é se escolhemos um valor maior para ν. Essa diferença, por outro lado, é bastante significativa se comparamos o resultado obtido para as estimativas de mínimos quadrados. O mesmo vale para os valores estimados do parâmetro de escala em cada ajuste.

Tabela 4.1: Estimativas e intervalos de confiança para o parâmetro de escala $\delta = 1,5$ e o erro quadrático médio para cada ajuste obtidas no primeiro estudo de simulação onde o ruído foi gerado segundo uma distribuição t de Student com 1,5 graus de liberdade.

Ajuste	*bootstrap*	$\widehat{\delta}$	$IC(95\%)$	EQM
1	paramétrico	1,536	[1,326;1,722]	0,0044
1	não-paramétrico	1,536	[1,360;1,687]	0,0044
2	paramétrico	2,018	[1,842;2,199]	0,0168
2	não-paramétrico	2,018	[1,816;2,336]	0,0168
M.Q.	paramétrico	10,011	[9,191;10,824]	0,3632
M.Q.	não-paramétrico	10,011	[4,806;,14,848]	0,3632

Em um segundo estudo de simulação, testamos a metodologia proposta em um caso em que a função alvo apresenta um comportamento menos comportado do que aquela utilizada no exemplo anterior. Para tanto, usamos a função definida em 4.1(c). Além disso, desta vez, simulamos os dados assumindo que o ruído obedecia a uma distribuição de Cauchy e, como anteriormente, assumimos o parâmetro de escala $\delta = 1,5$. O resultado da simulação descrita foi uma amostra de tamanho 300. Para efeito de análise, foram realizados 3 ajustes, sendo que, para todos eles, consideramos $M = 13$ e obtivemos amostras *bootstrap* de tamanho 100 com base nas quais construímos intervalos de confiança (95%) para δ e bandas de confiança (95%) para as estimativas da função alvo. O primeiro ajuste (Ajuste 1) foi feito segundo a metodologia proposta na tese usando a própria distribuição de Cauchy, o segundo ajuste também foi realizado segundo a metodologia proposta, porém, assumindo uma distribuição t com $\nu = 3$ graus de liberdade e, finalmente, o terceiro ajuste (Ajuste 3) foi feito via mínimos quadrados. O resultado, relativo à estimativa da função alvo, deste estudo pode ser visto na figura 4.3. Quanto ao parâmetro de escala, obtivemos os resultados contidos na tabela 4.2. O fenômeno observado no primeiro estudo de simulação repete-se aqui ainda mais evidente. Nota-se que a distância entre as estimativas e erros quadráticos médios obtidos usando-se distribuições com caudas pesadas porém distintas é significativamente menor do que aquela entre qualquer um destes resultados o aqueles obtidos quando obtemos as estimativas via mínimos quadrados.

4.4.2 Aplicação a Outros Métodos de Aproximação

Abaixo seguem os resultados de mais algumas simulações realizadas com base no algoritmo sugerido, mas utilizando outros métodos de suavização diferentes de *splines*. O objetivo é simplesmente ilustrar o quão

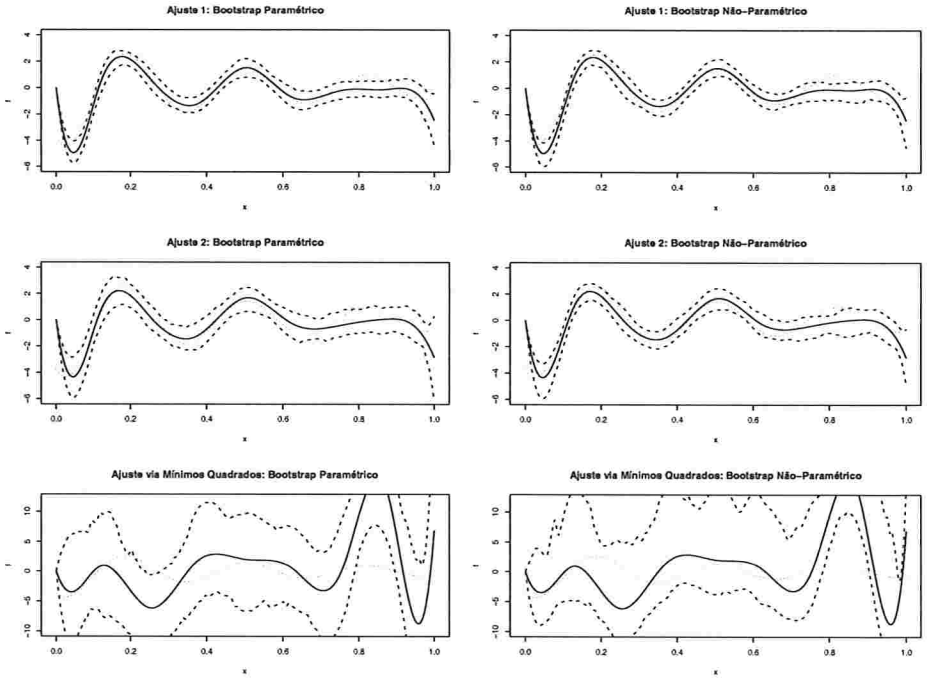


Figura 4.3: Estimativas da função alvo 2 segundo a metodologia proposta, Ajuste 1 (gráficos superiores — correspondem ao ajuste com $\nu = 1$) e Ajuste 2 (gráficos intermediários — corresponde ao ajuste com $\nu = 3$), e via mínimos quadrados, Ajuste 3 (gráficos inferiores).

bem a metodologia proposta adequa-se a outras técnicas de aproximação e como ele se comporta quando os parâmetros da distribuição associada ao ruído fixados a priori não correspondem exatamente aos de sua distribuição. Em todas as simulações abaixo, consideramos a função alvo 2, figura (4.1)(b).

Tabela 4.2: Estimativas e intervalos de confiança para o parâmetro de escala $\delta = 1,5$ e o erro quadrático para cada ajuste obtidos no segundo estudo de simulação onde o ruído foi gerado segundo uma distribuição de Cauchy.

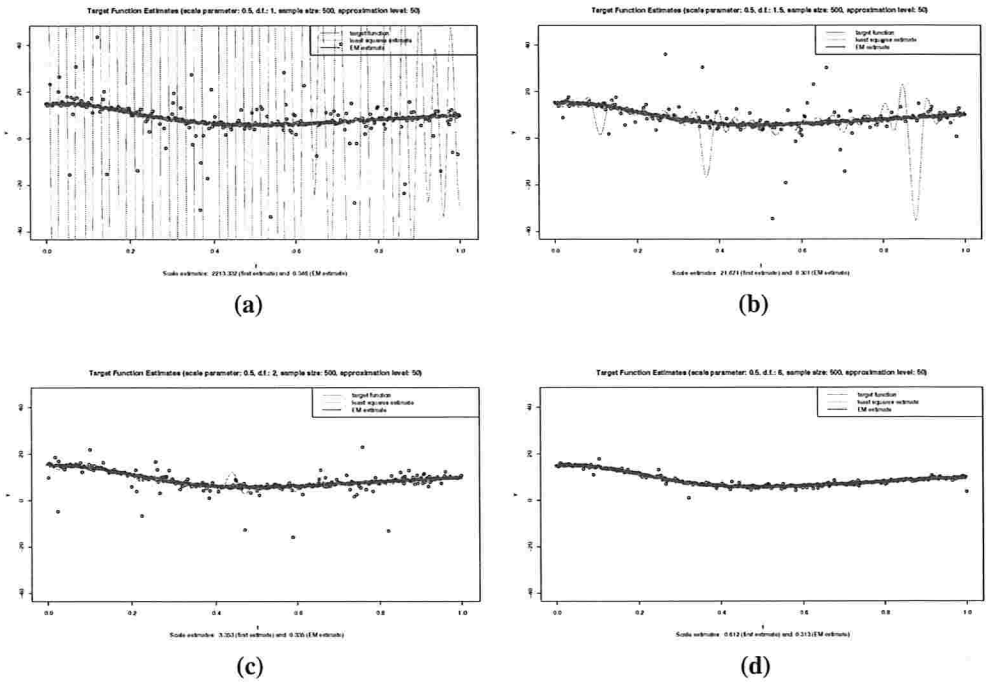
Ajuste	*bootstrap*	$\widehat{\delta}$	$IC(95\%)$	EQM
1	paramétrico	1,593	[1,328;1,743]	0,3788
1	não-paramétrico	1,593	[1,327;1,768]	0,3788
2	paramétrico	2,807	[2,499;3,038]	0,4243
2	não-paramétrico	2,807	[2,420;3,215]	0,4243
M.Q.	paramétrico	28,059	[26,069;29,111]	26,0889
M.Q.	não-paramétrico	28,059	[8,494;45,566]	26,0889

Séries Trigonométricas

Em todas as simulações abaixo (relativas às séries trigonométricas) tomamos $\delta = 0,5$.

• **Simulação 1:**

A figura 4.4 contém os resultados das estimativas obtidas sobre uma amostra de tamanho 500 gerada de acordo com uma distribuição de Cauchy e com uma t de Student com graus de liberdade iguais a $1,5$; 2 e 6, respectivamente. Além disso, as variáveis independentes foram escolhidas de modo a formar um conjunto de pontos igualmente espaçados no intervalo $[0, 1]$. Em todas estas simulações usamos 50 funções base.



(a) (b)

(c) (d)

Figura 4.4: (**Simulação 1**) Estimativas via algoritmo EM e por mínimos quadrados (usando séries trigonométricas) nos casos em que o ruído segue uma distribuição t de Student com graus de liberdade iguais a (a) $\nu = 1,0$, (b) $\nu = 1,5$, (c) $\nu = 2,0$ e (d) $\nu = 6,0$.

Note que a curva estimada pelo algoritmo EM ajusta-se perfeitamente à curva real, enquanto que o estimador de mínimos quadrados é extremamente sensível aos valores extremos. Conforme aumentamos ν, a distribuição do ruído aproxima-se da distribuição gaussiana (embora uma t de Student com $\nu = 6$ ainda tenha caudas bem pesadas quando comparada a uma distribuição normal) e, como pode-se notar, os estimadores usual e EM ajustam-se bem aos dados. Com relação ao parâmetro de escala, obtivemos os resultados expostos na tabela 4.3. Assim como no caso das curvas estimadas, podemos notar através dos parâmetros de escala obtidos que o método iterativo é extremamente menos sensível aos valores extremos do que o método de mínimos quadrados.

• **Simulação 2:**

Na figura 4.5, geramos 100 amostras para as quais o ruído associado segue a distribuição $\mathcal{N}(0, \delta^2)$, com $\delta = 0,5$, e obtivemos estimativas com $\nu = 1$, $\nu = 1.5$, $\nu = 2$ e $\nu = 6$. Em todas as estimativas, utilizamos as 20 primeiras funções base e 1000 iterações. Pode-se notar que, embora não tão quanto o ajuste via mínimos quadrados, as estimativas via EM aproximam-se bem da curva real mesmo quando o ruído não tem caudas pesadas e seguem uma distribuição gaussiana. Como era de se esperar, pode-se também notar que a estimativa do fator de escala aproxima-se do valor real do desvio padrão conforme aumentamos os graus de liberdade.

• **Simulação 3:**

Finalmente, consideramos o caso em que o ruído é gerado de acordo uma distribuição de Cauchy e estimamos a f assumindo graus de liberdade superiores a 1. O resultado destas estimativas está contido na figura 4.6. Embora, as estimativas apresentem um descolamento maior em relação à curva teórica quando comparadas às outras estimativas, podemos notar que, ainda assim, elas são bastante resistentes a valores extremos. Observamos também um aumento sensível na estimativa do fator de escala. Isto era esperado, pois, uma vez que aumentamos os graus de liberdade associados ao estimador, mais nos aproximamos por máxima verossimilhança supondo a normalidade do ruído.

Ondaletas

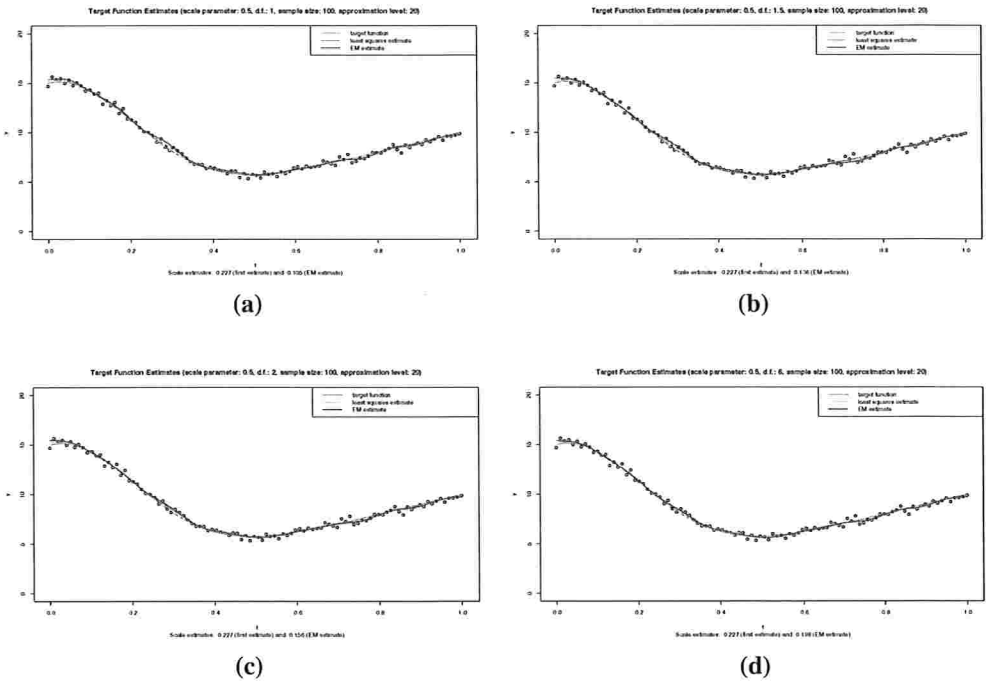
Nesta seção usamos ondaletas para aproximar a função alvo f. Como se sabe, qualquer função em $L^2[0,1]$ pode ser representada por uma série de ondaletas de modo que

$$f(x) = c_0\phi_{00}(x) + \sum_{j=0}^{\infty}\sum_{k=0}^{2^j-1} d_{jk}\psi_{jk}(x),$$

onde ϕ_{jk} e ψ_{jk} são as funções de escala e ondaletas, respectivamente, e c_{jk} e d_{jk} seus respectivos coeficientes. Para mais detalhes a respeito de aproximações via ondaletas, recomendamos [30]. Antes de realizar o estudo de simulação, gostariamos de observar que a presença de valores extremos tem forte impacto não apenas sobre estimativas obtidas via mínimos quadrados, mas também sobre as estimativas obtidas via transformadas discretas de ondaletas. Este fato pode ser observado quando comparamos as figuras 4.7 e 4.8. Elas

Tabela 4.3: Estimativas do parâmetro de escala nas estimativas da simulação 1. A segunda coluna, $\widehat{\delta}$, são as estimativas obtidas pelo método iterativo, enquanto que a terceira coluna, M.Q., são as estimativas obtidas via mínimos quadrados.

Graus de Liberdade	$\widehat{\delta}$	M.Q.
1	0,59	47,05
1,5	0,55	4,66
2	0,58	1,83
6	0,56	0,78



(a) (b)

(c) (d)

Figura 4.5: (**Simulação 2**) Estimativas da função alvo e do parâmetro de escala no caso em que o ruído segue uma distribuição normal (usando séries trigonométricas) e assumindo os seguintes graus de liberdade para o processo de estimação (a) $\nu = 1,0$, (b) $\nu = 1,5$, (c) $\nu = 2,0$ e (d) $\nu = 6,0$.

correspondem à análise de multirresolução do sinal

$$y_t = \text{sen}(10t) + \epsilon_t, \tag{4.37}$$

onde $\epsilon_t \sim \mathcal{N}(0,1)$, figura 4.7, e $\epsilon_t \sim t_2$, figura 4.8. Note que no primeiro caso, isto é, o gaussiano, os maiores coeficientes de ondaletas (detalhes) correspondem ao sinal real, enquanto que no segundo caso, os maiores coeficientes de ondaletas correspondem ao ruído. Logo, a aplicação dos limiares usuais não seria muito efetiva no sentido de que não eliminaria a "sujeira" deixada pelo ruído. Lembramos que a expressão para o limiar universal, derivado em [15], é

$$\delta\sqrt{2\log T},$$

onde δ é o parâmetro de escala (desvio-padrão para o ruído gaussiano). A dedução deste limiar é baseada na desigualdade de Borel através da qual podemos inferir que para $Z_0, \ldots, Z_{T-1}$ gaussianos com média zero e variâncias δ_t^2, vale a desigualdade

$$P\left(\max_{0\leq t<T}\left|\frac{Z_t}{\delta_t}\right| > \sqrt{2\log T}\right) \longrightarrow 0$$

quando $T \to \infty$. Por outro lado, dado que podemos representar a distribuição t de Student como uma mistura de normais através da escala, nós podemos através de uma aplicação da desigualdade de Borel obter o limiar

universal "equivalente" para sinais com ruídos distribuidos de acordo com uma t de Student com ν graus de liberdade, o qual deve ser proporcional a

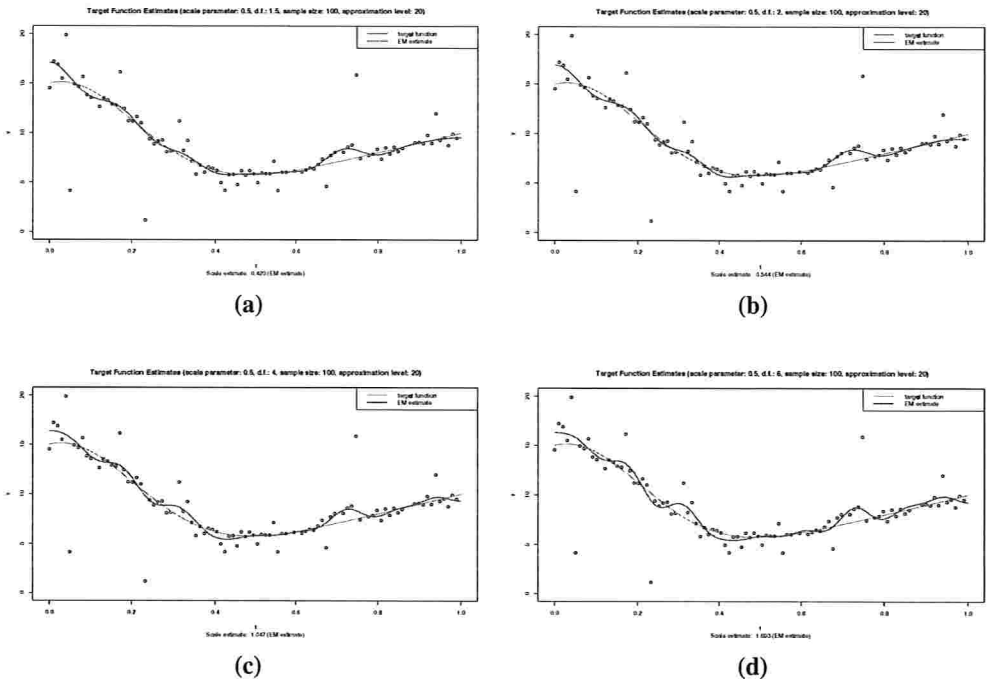
$$\sqrt{T^{\nu/2} - 1}. \tag{4.38}$$

Um limiar como este anularia todos os coeficientes de ondaletas. A dedução completa da expressão (4.38) encontra-se no apêndice ao final deste capítulo.

Para o estudo de simulação, dado que ondaletas são extremamente eficientes em captar descontinuidades da função alvo, escolhemos como função alvo

$$f(x) = \begin{cases} 0, & \text{caso } 0 \leq x < \frac{1}{4} \\ 2, & \text{caso } \frac{1}{4} \leq x < \frac{1}{2} \\ -3, & \text{caso } \frac{1}{2} \leq x < \frac{3}{4} \\ 6, & \text{caso } \frac{3}{4} \leq x < 1 \end{cases} \tag{4.39}$$

a qual está representada na figura 4.9. Optamos, finalmente, em tomar um ruído, $\{\epsilon_t\}$, distribuido de acordo com uma distribuição de Cauchy e tomar como parâmetro de escala $\delta = 1,5$. Fixados estes parâmetros,



(a) (b)

(c) (d)

Figura 4.6: (**Simulação 3**) Estimativas da função alvo e do parâmetro de escala no caso em que o ruído segue uma distribuição de Cauchy (usando séries trigonométricas) e assumindo os seguintes graus de liberdade para o processo de estimação (a) $\nu = 1,5$, (b) $\nu = 2,0$, (c) $\nu = 4,0$ e (d) $\nu = 6,0$.

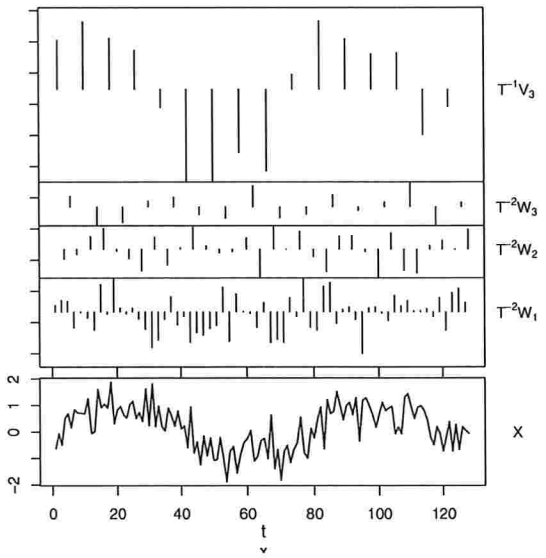


Figura 4.7: Análise de multirresolução correspondente ao sinal (4.37) com ruído distribuido de acordo com uma normal padrão.

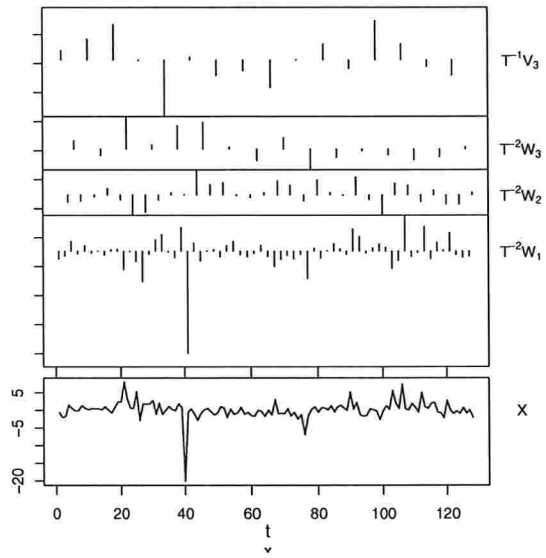


Figura 4.8: Análise de multirresolução correspondente ao sinal (4.37) com ruído distribuido de acordo com uma t de Student com 2 graus de liberdade.

geramos uma amostra pseudo-aleatória de tamanho $T = 512$ de acordo com o modelo

$$y_t = f(x_t) + \delta\epsilon_t$$

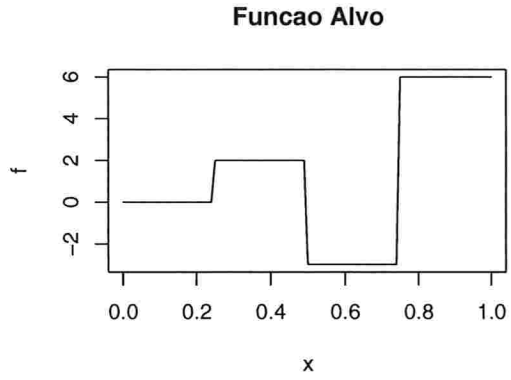


Figura 4.9: Gráfico da função alvo (4.9) usada na simulação via ondaletas.

onde os x_t's são igualmente espaçados em $[0, 1]$.

Para estimar a função alvo e o parâmetro de escala optamos por usar a família Haar de ondaletas. Os resultados das estimativas obtidas podem ser vistas na figura 4.10. Em ambos os gráficos a curva pontilhada representa a função alvo original. O gráfico à esquerda representa um ajuste simples via mínimos quadrados. É evidente o efeito dos valores extremos sobre a estimativa obtida. Isto fica claro, principalmente, no segundo degrau da estimativa o qual é "jogado" para baixo devido a presença de valores extremos. O gráfico a direita na mesma figura representa a estimativa obtida iterativamente. Note que ela capta perfeitamente a forma da função alvo ajustando-se perfeitamente à mesma. Além disso, obtivemos via bootstrap as bandas de confiança (95%) para esta estimativa, as quais estão representadas pelas curvas tracejadas. Quanto ao parâmetro de escala, obtivemos a estimativa $\widehat{\delta} = 1,6616$ (contra $34,3743$ quando tal estimativa é extraída diretamente da coeficientes estimados via mínimos quadrados) e o intervalo de confiança a 95% (bootstrap) dado por $[1,3085; 2,3271]$.

4.5 Modelos Parcialmente Lineares

Possivelmente, a extensão mais simples do modelo univariado (4.1) para modelos multivariados seja os modelos parcialmente lineares. Estes modelos são caracterizados pela presença de uma componente não-paramétrica, em geral, univariada, e uma componente paramétrica linear multivariada. A expressão matemática para estes modelos é dada por

$$y_t = f(x_t) + \boldsymbol{\beta}' \mathbf{u}_t + \epsilon_t. \tag{4.40}$$

Duas aplicações interessantes para (4.40) são, em primeiro lugar, testar se a contribuição de uma determinada variável explicativa é linear ou não e, em segundo lugar, permitir a introdução de outras covariáveis no modelo.

O objetivo desta seção é mostrar como podemos estender a metodologia aplicada ao modelo (4.1) para

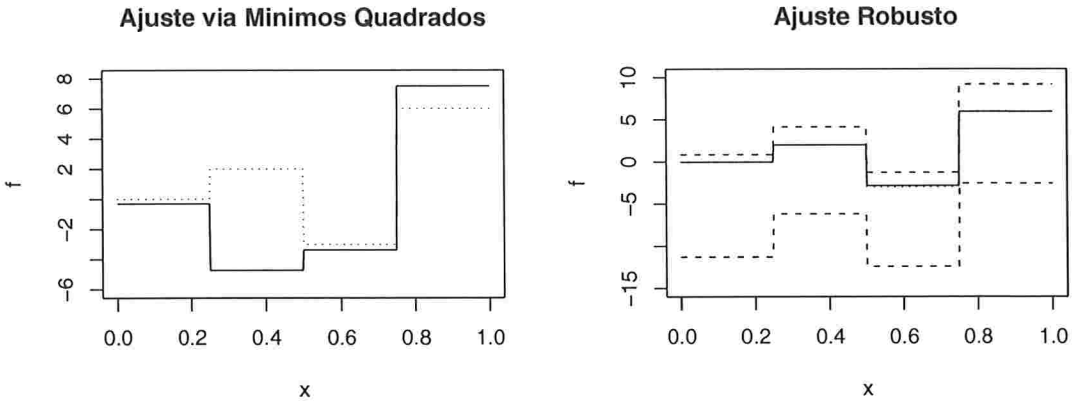


Figura 4.10: Ajuste da função escada (4.9) usando ondaletas via mínimos quadrados (a esquerda) e através da metodologia sugerida (a direita). Em ambos os gráficos, a linha pontilhada corresponde à função alvo original e as linhas tracejadas às bandas de confiança obtidas via *bootstrap*.

o caso do modelo (4.40). Ou seja, mostraremos como adaptá-la para estimar a função alvo f, o vetor de parâmetros (coeficientes) $\boldsymbol{\beta}$ e o vetor de parâmetros $\boldsymbol{\zeta}$ associado à distribuição dos ruídos ϵ_t.

Começamos a análise modelando os dados de acordo com

$$\begin{aligned} y_t|\sigma_t &\sim \mathcal{N}(f(x_t) + \boldsymbol{\beta}'\mathbf{u}_t, \delta^2\psi(\sigma_t)^2), \\ \sigma_t &\sim h. \end{aligned}$$

Aproximando f por $\bar{f} \equiv \sum_{j=1}^{M} c_j B_j$, temos que os parâmetros do modelo são $\boldsymbol{\theta} \equiv (\mathbf{c}, \boldsymbol{\beta}, \delta, \boldsymbol{\zeta})'$, onde $\boldsymbol{\zeta}$ é o vetor de parâmetros associado à densidade h. Procedendo analogamente às seções anteriores, temos

$$\begin{aligned} Q(\boldsymbol{\theta}|\boldsymbol{\theta}_0) &= E_{\boldsymbol{\theta}_0}\{\log p(\mathbf{y}, \boldsymbol{\sigma})|\mathbf{z}\} \\ &= \sum_{t=1}^{T} E_{\boldsymbol{\theta}_0}\{\log p(y_t, \sigma_t|z_t\} \\ &\equiv \sum_{t=1}^{T} Q_t(\boldsymbol{\theta}|\boldsymbol{\theta}_0). \end{aligned}$$

onde $z_t = (y_t, x_t, \mathbf{u}_t')'$. Agora,

$$\begin{aligned} Q_t(\boldsymbol{\theta}|\boldsymbol{\theta}_0) = &-\frac{1}{2}\log(2\pi) - E_{\boldsymbol{\theta}_0}\{\log\psi(\sigma_t)|z_t\} - \log\delta \\ &- \frac{1}{2\delta^2} E_{\boldsymbol{\theta}_0}\left\{\frac{1}{\psi(\sigma_t)}\middle| z_t\right\}(y_t - \bar{f}_{\boldsymbol{\theta}}(x_t))^2 + E_{\boldsymbol{\theta}_0}\{\log h(\sigma_t|\boldsymbol{\theta})|z_t\}. \end{aligned}$$

Logo, eliminando o primeiro termo da expressão acima, já que o mesmo não interfere no resultado da

maximização de Q em função dos parâmetros do modelo,

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}_0) = C(\boldsymbol{\zeta};\boldsymbol{\theta}_0) - T\log\delta - \frac{1}{2\delta^2}(\mathbf{y} - B\mathbf{c} - U\boldsymbol{\beta})'W_T(\boldsymbol{\theta}_0)(\mathbf{y} - B\mathbf{c} - U\boldsymbol{\beta})$$

onde, como antes,

$$C(\boldsymbol{\zeta};\boldsymbol{\theta}_0) = \sum_{t=1}^{T} E_{\boldsymbol{\theta}_0}\left\{\log\frac{h(\sigma_t|\boldsymbol{\zeta})}{\psi(\sigma_t)}\middle| z_t\right\}$$

e

$$W_T(\boldsymbol{\theta}_0) = \text{diag}\left(E_{\boldsymbol{\theta}_0}\left\{\frac{1}{\psi(\sigma_1)^2}\middle| z_1\right\}, ..., E_{\boldsymbol{\theta}_0}\left\{\frac{1}{\psi(\sigma_T)^2}\middle| z_1\right\}\right).$$

Para o caso em que estivermos considerando a função de penalização $J_\lambda(\boldsymbol{\theta})$, a função objetiva será dada por

$$S(\boldsymbol{\theta}|\boldsymbol{\theta}_0) = Q(\boldsymbol{\theta}|\boldsymbol{\theta}_0) + J_\lambda(\boldsymbol{\theta}).$$

Com relação à distribuição associada às esperanças em $W_T(\boldsymbol{\theta}_0)$,

$$\begin{aligned} k(\sigma_t|z_t,\boldsymbol{\theta}_0) &\propto \frac{1}{\delta\psi(\sigma_t)}\phi\left(\frac{y_t - f(x_t) - \boldsymbol{\beta}'u_t}{\delta\psi(\sigma_t)}\right)h(\sigma_t|\boldsymbol{\theta}_0) \\ &\approx \frac{1}{\delta\psi(\sigma_t)}\phi\left(\frac{y_t - \bar{f}_{\boldsymbol{\theta}_0}(x_t) - \boldsymbol{\beta}'u_t}{\delta\psi(\sigma_t)}\right)h(\sigma_t|\boldsymbol{\theta}_0). \end{aligned}$$

Ao derivar Q em relação a $\mathbf{c}$, temos

$$\frac{\partial}{\partial\mathbf{c}}Q(\boldsymbol{\theta}|\boldsymbol{\theta}_0) = \frac{1}{\delta^2}\left(B'(\mathbf{y} - U\boldsymbol{\beta}) - B'W_T(\boldsymbol{\theta}_0)B\mathbf{c}\right),$$

enquanto que, ao derivar Q em relação a $\boldsymbol{\beta}$, obtemos

$$\frac{\partial}{\partial\boldsymbol{\beta}}Q(\boldsymbol{\theta}|\boldsymbol{\theta}_0) = \frac{1}{\delta^2}\left(U'(\mathbf{y} - B\mathbf{c}) - U'W_T(\boldsymbol{\theta}_0)U\boldsymbol{\beta}\right).$$

Por último, a derivada de Q em relação ao parâmetro de escala δ é dada por

$$\frac{\partial}{\partial\delta}(\boldsymbol{\theta}|\boldsymbol{\theta}_0) = \frac{1}{\delta^3}(\mathbf{y} - B\mathbf{c} - U\boldsymbol{\beta})'W_T(\boldsymbol{\theta}_0)(\mathbf{y} - B\mathbf{c} - U\boldsymbol{\beta}) - \frac{T}{\delta}.$$

Os passos para a aplicação do algoritmo EM para estimar $\boldsymbol{\theta}$ estão descritos no algoritmo 4.5.1.

4.5.1 Estimando f via P-splines

Uma outra possibilidade para se estimar a função alvo e os demais parâmetros é via regularização. Aqui, nós sugerimos o uso de P-splines e tomando $J_\lambda(\mathbf{c}) = \lambda\mathbf{c}'K\mathbf{c}$, onde a matriz K é descrita em (6.4) e (6.5), dependendo da ordem, temos

$$S(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)}) = Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)}) + \lambda\mathbf{c}'K\mathbf{c},$$

Algoritmo 4.5.1 Esquematização do algoritmo EM para o tratamento de modelos parcialmente lineares na presença de um erro distribuído de acordo com uma distribuição t de Student.

Dados $\mathbf{c}^{(k)}$, $\boldsymbol{\beta}^{(k)}$, $\delta^{(k)}$ e $\boldsymbol{\zeta}^{(k)}$:

1. Calcular
$$\mathbf{c}^{(k+1)} = (B'W_T(\boldsymbol{\theta}^{(k)})B)^{-1}B'W_T(\boldsymbol{\theta}^{(k)})(\mathbf{y} - U\boldsymbol{\beta}^{(k)});$$
2. Calcular
$$\boldsymbol{\beta}^{(k+1)} = (U'W_T(\boldsymbol{\theta}^{(k)})U)^{-1}U'W_T(\boldsymbol{\theta}^{(k)})(\mathbf{y} - B\mathbf{c}^{(k+1)});$$
3. Calcular
$$(\delta^2)^{(k+1)} = \frac{1}{T}(\mathbf{y} - B\mathbf{c}^{(k+1)} - U\boldsymbol{\beta}^{(k+1)})'W_T(\boldsymbol{\theta}^{(k)})(\mathbf{y} - B\mathbf{c}^{(k+1)} - U\boldsymbol{\beta}^{(k+1)})$$
4. Obter $\boldsymbol{\zeta}^{(k+1)}$ como solução de
$$\frac{\partial}{\partial \boldsymbol{\zeta}}Q(\boldsymbol{\zeta}; \mathbf{c}^{(k+1)}, \boldsymbol{\beta}^{(k+1)}, \delta^{(k+1)}|\boldsymbol{\theta}^{(k)}) = 0$$

de modo que o único passo no algoritmo descrito na seção anterior que deve ser corrigido em virtude da penalização imposta é o primeiro. De fato,

$$\frac{\partial}{\partial \mathbf{c}}S(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)}) = \frac{1}{\delta^2}B'W_T(\boldsymbol{\theta}^{(k)})(\mathbf{y} - U\boldsymbol{\beta}) - \frac{1}{\delta^2}B'W_T(\boldsymbol{\theta}^{(k)})B\mathbf{c} + 2\lambda K\mathbf{c}$$

de modo que

$$\begin{aligned}\mathbf{c}^{(k+1)} &= \left(\frac{1}{\delta^2}B'W_T(\boldsymbol{\theta}^{(k)})B + 2\lambda K\right)^{-1}\frac{1}{\delta^2}B'W_T(\boldsymbol{\theta}^{(k)})\mathbf{y} \\ &= \left(B'W_T(\boldsymbol{\theta}^{(k)})B + 2\lambda\delta^2 K\right)^{-1}B'W_T(\boldsymbol{\theta}^{(k)})\mathbf{y}\end{aligned}$$

4.5.2 Estudo de Simulação

No primeiro estudo de simulação para modelos parcialmente lineares, consideramos a função alvo da figura 4.11 e assumimos que a variável explicativa associada a componente linear $\mathbf{u}$ é um vetor de dimensão 3 com $\boldsymbol{\beta} = (0,45; 1,23; -3)'$. As colunas da matriz de planejamento $\mathbf{U}$ foram geradas de acordo com uma variável uniforme definida no intervalo $(3(c-1), 3c)$, onde c representa a numeração das colunas de $\mathbf{U}$. Em relação ao ruído, assumimos que ele é distribuido de acordo com uma distribuição t de Student com 3 graus de liberdade e parâmetro de escala $\delta = 1,5$. As bandas e intervalos de confiança foram obtidos via bootstrap (paramétrico e não-paramétrico), onde para cada ajuste fizemos 200 simulações. Ao todo, fizemos quatro estudos:

Estimativa 1. Neste estudo assumimos uma distribuição t para o ruído com 3 graus de liberdade e usamos o bootstrap paramétrico para o cálculo das bandas e intervalos de confiança.

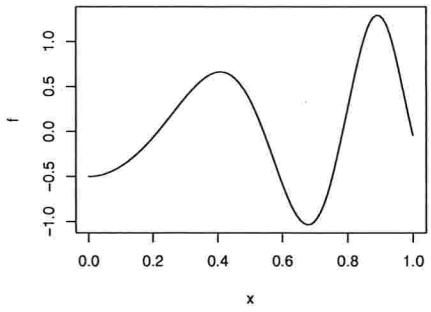


Figura 4.11: Função alvo utilizada: $x^2 + \text{sen}(10x^2) - 0,5$.

Estimativa 2. Neste estudo assumimos uma distribuição t para o ruído com 3 graus de liberdade e usamos o bootstrap não-paramétrico para o cálculo das bandas e intervalos de confiança.

Estimativa 3. Neste estudo assumimos uma distribuição t para o ruído com 5 graus de liberdade e usamos o bootstrap paramétrico para o cálculo das bandas e intervalos de confiança.

Estimativa 4. Neste estudo assumimos uma distribuição t para o ruído com 5 graus de liberdade e usamos o bootstrap não-paramétrico para o cálculo das bandas e intervalos de confiança.

O resultado deste ajuste pode ser visto na figura 4.12. Com relação aos demais parâmetros do modelo, obtivemos para o parâmetro de escala os resultados contidos na tabela 4.4, enquanto que, com relação aos

Tabela 4.4: Tabela com os intervalos de confiança obtidos via bootstrap para o parâmetro de escala δ.

	$\widehat{\delta}$	**5%**	**95%**
Estimativa 1.	1,3913	1,2373	1,4918
Estimativa 2.	1,3913	1,2221	1,4680
Estimativa 3.	1,5500	1,3831	1,6670
Estimativa 4.	1,5500	1,3930	1,6385

coeficientes associados à componente linear do modelo, ie, β, obtivemos os resultados contidos nas tabelas 4.5 e 4.6 Note que apesar da presença de valores extremos devido à distribuição assumida para o ruído, as curvas estimadas ajustam-se muito bem aos dados e o modelo ainda é capaz de captar o parâmetro de escala δ e os coeficientes associados à componente linear com precisão como se pode ver comparando-se o valor real com os valores estimados e com os intervalos de confiança obtidos.

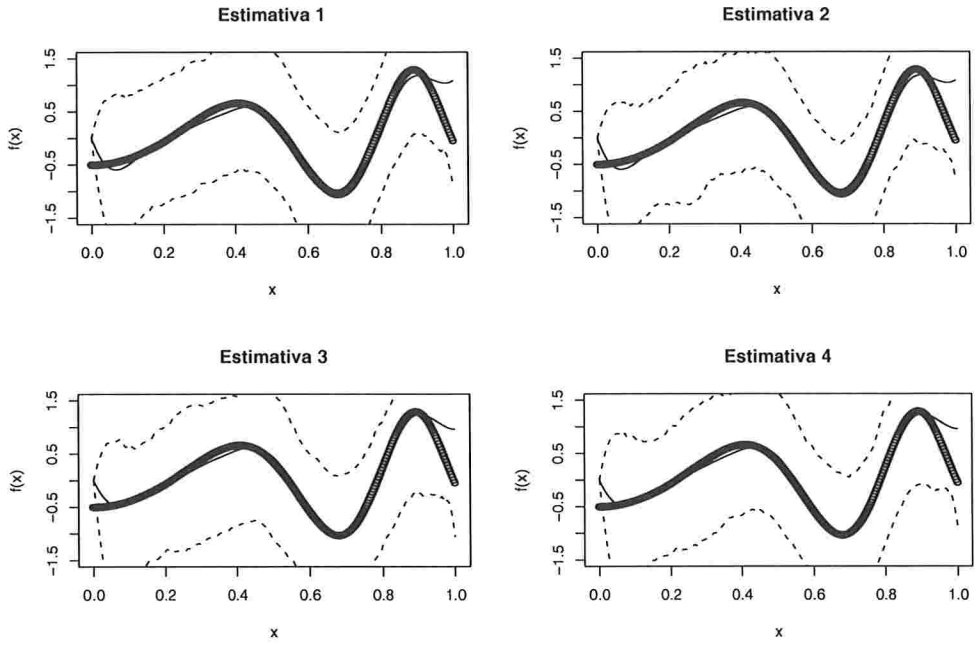


Figura 4.12: Estimativas obtidas para a função alvo da figura 4.11. A figura no canto superior esquerdo (Estimativa 1) corresponde à estimativa obtida de acordo com a metodologia sugerida assumindo a mesma distribuição que gerou o ruído. As bandas de confiança foram obtidas via bootstrap paramétrico com 200 simulações. Analogamente, porém com bootstrap não-paramétrico, temos a figura no canto superior direito. As figuras inferiores seguem o mesmo padrão, mas asssumindo-se que a distribuição subjacente tem 5 graus de liberdade, ao invés de 3.

Tabela 4.5: Tabela com os intervalos de confiança obtidos via bootstrap para o vetor de coeficientes $\beta = (0,45; 1,23; -3)'$ associado à componente linear do modelo.

	$\widehat{\beta}_1$	$\widehat{\beta}_2$	$\widehat{\beta}_3$
Estimativas 1 e 2.	0,3446	1,1893	-2,9446
Estimativa 3 e 4.	0,3304	1,1753	-2,9380

4.6 Apêndice

4.6.1 Dedução de (4.38)

A Distribuição t Multivariada

Dizemos que um vetor aleatório $\mathbf{Z} = (Z_0, ..., Z_{T-1})'$ segue uma *distribuição t multivariada com ν graus de liberdade, média μ e matriz de correlação R (ou Σ como matriz de covariância)* se

$$f(\mathbf{z}) = \frac{\Gamma\left(\frac{\nu+T}{2}\right)}{(\pi\nu)^{T/2}\Gamma\left(\frac{\nu}{2}\right)|R|^{1/2}}\left[1 + \frac{1}{\nu}(\mathbf{z}-\boldsymbol{\mu})'R^{-1}(\mathbf{z}-\boldsymbol{\mu})\right]^{-\frac{\nu+T}{2}}.$$

Tabela 4.6: Tabela com os intervalos de confiança obtidos via bootstrap para o vetor de coeficientes $\boldsymbol{\beta} = (0,45; 1,23; -3)'$ associado à componente linear do modelo.

	β_1		β_2		β_3	
	5%	**95%**	**5%**	**95%**	**5%**	**95%**
Estimativa 1.	0,2802	0,6587	1,0786	1,3904	-3,1617	-2,8689
Estimativa 2.	0,2710	0,6508	1,0227	1,3806	-3,1333	-2,8755
Estimativa 3.	0,2719	0,6627	1,0265	1,4042	-3,1223	-2,8560
Estimativa 4.	0,2655	0,6389	1,0659	1,4056	-3,1561	-2,8466

No caso em que $\boldsymbol{\mu} = \mathbf{0}$, dizemos que a distribuição é central e, no caso geral, usamos a notação $\mathbf{Z} \sim t_{\nu,T}(\boldsymbol{\mu}, R)$.

Por outro lado, se $(\mathbf{Y}, S)$ segue uma distribuição normal-gama multivariada com parâmetros $\boldsymbol{\mu}$, R, ν e τ, i.e., se

$$\begin{cases} \mathbf{Y}|S = s \sim \mathcal{N}_T\left(\boldsymbol{\mu}, \frac{R}{s}\right), \\ S \sim \Gamma\left(\frac{\nu}{2}, \frac{\tau\nu}{2}\right), \end{cases}$$

então, $\mathbf{Y} \sim t_{\nu,T}(\boldsymbol{\mu}, \sigma R)$. Em particular, se $\sigma = 1$, então $\mathbf{Y} \sim t_{\nu,T}(\boldsymbol{\mu}, R)$.

A Desigualdade de Borel

Abaixo temos o teorema, conhecido como desigualdade de Borel, cuja prova pode ser encontrada em [2].

Teorema 4.6.1. *Seja $\mathbf{Z} = \{Z_t\}_{t\in T}$ um processo gaussiano de média zero e com trajetórias amostrais limitadas quase-certamente. Defina $\|\mathbf{Z}\|_1 \equiv \sup_{t\in T} Z_t$. Então, $E\|\mathbf{Z}\|_1 < \infty$ e, para todo $\lambda > 0$,*

$$P\{\|\mathbf{Z}\|_1 - E\|\mathbf{Z}\|_1 > \lambda\} \leq 2e^{-\frac{\lambda^2}{2\sigma_T^2}},$$

onde $\sigma_T^2 \equiv \sup_{t\in T} EZ_t^2$.

Dedução do Limiar

Denotemos

$$Z = S^{-1}\mathbf{Y},$$

onde

$$\begin{cases} \mathbf{Y} \sim \mathcal{N}_T(\mathbf{0}, \Sigma), \\ \chi \equiv \frac{\nu S^2}{\sigma^2} \sim \chi^2_\nu, \end{cases}$$

de modo que, assumindo-se que R é a matriz de correlação de $\mathbf{Y}$, teremos $\mathbf{Z} \sim t_{\nu,T}(\mathbf{0}, R)$. Assumiremos também que $E\|\mathbf{Y}\|_1 = \mathbf{0}$. Agora, aplicando a desigualdade de Borel sobre o vetor aleatório $\mathbf{Y}$ e definindo

$\Lambda \equiv \{\mathbf{z} : \|\mathbf{z}\|_1 > \lambda\}$,

$$\begin{aligned} P\{\|\mathbf{Z}\|_1 > \lambda\} &= P\{\Lambda\} = EI(\Lambda) \\ &= EE[I(\Lambda)|S] = EP\{\|\mathbf{Y}\|_1 > S\lambda|S\} \\ &\leq 2Ee^{-\frac{\lambda^2 S^2}{2\sigma_T^2}} \equiv 2Ee^{\kappa\chi}, \end{aligned}$$

onde I representa a função indicadora, $\kappa \equiv -\frac{\lambda^2\sigma^2}{2\sigma_T^2\nu} < 0$ e $\chi \sim \chi^2_\nu$, onde $\sigma_T^2 \equiv \sup_{0\leq t\leq T} EY_t^2$. Então, usando o fato que

$$Ee^{\kappa\chi} = \left(\frac{1}{1-2\kappa}\right)^{\nu/2} = \left(\frac{\sigma_T^2\nu}{\sigma_T^2\nu + \lambda^2\sigma^2}\right)^{\nu/2} = \left(\frac{\nu^2}{\nu^2 + (\nu-2)\lambda^2}\right)^{\nu/2},$$

pois, $\text{Var}\, Y_i = \frac{\sigma^2\nu}{\nu-2}$. Logo,

$$P\{\|\mathbf{Z}\|_1 > \lambda\} \leq 2\left(\frac{\nu^2}{\nu^2 + (\nu-2)\lambda^2}\right)^{\nu/2}. \tag{4.41}$$

Consequentemente, para se obter uma taxa de decaimento da probabilidade (4.41) equivalente ao caso gaussiano, o limiar deve ser estabelecido como sendo

$$\lambda_{\nu,T} = \frac{\nu}{\sqrt{\nu-2}}\sqrt{T^{\nu/2} - 1}.$$

Note que, para T fixado, $\lim_{\nu\to 2} \lambda_{\nu,T} = \infty$.

Capítulo 5

Modelos Autorregressivos — Enfoque Semiparamétrico

5.1 Modelo Semiparamétrico

Neste capítulo aplicaremos um procedimento análogo ao sugerido no capítulo 4, mas para dados de séries temporais supostamente não-lineares.

5.1.1 Modelo Não-Linear

Considere o modelo

$$Y_t = f(Y_{t-1}) + \epsilon_t \tag{5.1}$$

onde ϵ_t segue uma mistura na escala de distribuições gaussianas. Como nos demais capítulos, sua função densidade de probabilidade será dada por

$$p(\epsilon_t) = \int_0^\infty \frac{1}{\psi_{\boldsymbol{\theta}}(\sigma_t)} \phi\left(\frac{\epsilon_t}{\psi_{\boldsymbol{\theta}}(\sigma_t)}\right) h(\sigma_t) d\sigma_t,$$

onde h é a densidade de mistura. Em particular, temos que a distribuição condicional de $Y_t|Y_{t-1} = y_{t-1}$ é uma mistura de gaussianas na escala cuja função densidade de probabilidades é dada por (5.2). O objetivo é, portanto, estimar a função alvo f e os parâmetros associados à h. Como antes, denotaremos o vetor com a totalidade de parâmetros a serem estimados por $\boldsymbol{\theta}$ e o vetor de parâmetros associados exclusivamente a h por $\boldsymbol{\zeta}$. Para cumprir o objetivo, modelamos os dados de acordo com

$$\begin{aligned} Y_t|y_{t-1}; \sigma_t &\sim \mathcal{N}(f(y_{t-1}), \delta^2 \psi_{\boldsymbol{\theta}}(\sigma_t)^2), \\ \sigma_t &\sim h, \end{aligned}$$

para todo $t = 2, ..., T$ e onde as variáveis $(\sigma_2, ..., \sigma_T)'$ são latentes, ie, não observáveis. Como no caso paramétrico, consideraremos y_1 determinístico, de maneira que, neste caso, as distribuições condicionais

$p(y_t|y_{t-1})$ sejam bem-definidas e dadas por

$$\begin{aligned} p(y_t|y_{t-1}) &= \int_0^\infty p(y_t, \sigma_t|y_{t-1})d\sigma_t \\ &= \int_0^\infty \frac{1}{\delta\psi_{\boldsymbol{\theta}}(\sigma_t)}\phi\left(\frac{y_t - f(y_{t-1})}{\delta\psi_{\boldsymbol{\theta}}(\sigma_t)}\right) h(\sigma_t|\boldsymbol{\zeta})d\sigma_t. \end{aligned} \tag{5.2}$$

Para aplicar o algoritmo EM, é preciso calcular o valor esperado

$$Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(k)}) = E_{\boldsymbol{\theta}^{(k)}}\{\log p(\mathbf{y}, \boldsymbol{\sigma}|\boldsymbol{\theta})|\mathbf{y}\}$$

com relação à densidade condicional

$$k(\boldsymbol{\sigma}|y_T, ..., y_1, \boldsymbol{\theta}^{(k)}),$$

onde $\boldsymbol{\sigma} \equiv (\sigma_2, ..., \sigma_T)'$, $\mathbf{y} \equiv (y_2, ..., y_T)'$ e $\boldsymbol{\theta}^{(k)}$ é a estimativa para $\boldsymbol{\theta}$ obtida no k-ésimo passo do algoritmo. Expandindo e desenvolvendo algebricamente as densidades $k(\boldsymbol{\sigma}|y_T, ..., y_1; \boldsymbol{\theta}^{(k)})$ e $p(\mathbf{y}, \boldsymbol{\sigma}|\boldsymbol{\theta}))$, respectivamente, obtemos

$$\begin{aligned} k(\boldsymbol{\sigma}|y_T, ..., y_1, \boldsymbol{\theta}^{(k)}) &= \frac{p(\mathbf{y}, \boldsymbol{\sigma}|\boldsymbol{\theta}^{(k)})}{p(\mathbf{y}|\boldsymbol{\theta}))} \propto p(\mathbf{y}|\boldsymbol{\sigma}; \boldsymbol{\theta})h(\boldsymbol{\sigma}|\boldsymbol{\zeta}) \\ &= \prod_{t=2}^{T} p(y_t|y_{t-1}, \sigma_t; \boldsymbol{\theta}))h(\sigma_t|\boldsymbol{\zeta}) \\ &\propto \prod_{t=2}^{T} \frac{1}{\psi_{\boldsymbol{\theta}}(\sigma_t)}\phi\left(\frac{y_t - f(y_{t-1})}{\psi_{\boldsymbol{\theta}}(\sigma_t)}\right) h(\sigma_t|\boldsymbol{\zeta})) \end{aligned}$$

e

$$\begin{aligned} p(\mathbf{y}, \boldsymbol{\sigma}|\boldsymbol{\theta}) &= p(y_T, ..., y_1, \boldsymbol{\sigma}|\boldsymbol{\theta}) \\ &= p(y_T|y_{T-1}, ..., y_1, \boldsymbol{\sigma}; \boldsymbol{\theta}) \cdots p(y_2|y_1, \boldsymbol{\sigma}; \boldsymbol{\theta})h(\boldsymbol{\sigma}|\boldsymbol{\zeta}) \\ &= \prod_{t=2}^{T} p(y_t|y_{t-1}, \sigma_t; \boldsymbol{\theta})h(\sigma_t|\boldsymbol{\zeta}). \end{aligned}$$

Logo,

$$\log p(\mathbf{y}, \boldsymbol{\sigma}|\boldsymbol{\theta}) = \sum_{t=2}^{T} \log p(y_t|y_{t-1}, \sigma_t; \boldsymbol{\theta}) + \sum_{t=2}^{T} \log h(\sigma_t|\boldsymbol{\zeta}).$$

Em particular, note que, assim como no caso de regressão, as variáveis latentes $\sigma_2, ..., \sigma_T$ permanecem independentes entre si mesmo após condicionadas com relação a $y_1, ..., y_T$. Agora,

$$\sum_{t=2}^{T} \log p(y_t|y_{t-1}, \sigma_t; \boldsymbol{\theta}) \approx -(T-1)\log\delta + \sum_{t=2}^{T} \log\frac{1}{\psi_{\boldsymbol{\theta}}(\sigma_t)} - \frac{1}{2\delta^2}\sum_{t=2}^{T}\frac{r_t^2}{\psi_{\boldsymbol{\theta}}(\sigma_t)}$$

onde $r_t = y_t - f(y_{t-1})$ e '$\approx$' significa igualdade a menos de uma constante. Conseqüentemente[1]

$$\begin{aligned} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)}) = &-\frac{1}{2}\sum_{t=2}^{T} E_{\boldsymbol{\theta}^{(k)}}\left\{\frac{1}{\psi_{\boldsymbol{\theta}}(\sigma_t)^2}\right\}\Big| y_t, y_{t-1}\Big\} r_t^2 \\ &+ \sum_{t=2}^{T} E_{\boldsymbol{\theta}^{(k)}}\left\{\log\frac{h(\sigma_t|\boldsymbol{\zeta})}{\psi_{\boldsymbol{\theta}}(\sigma_t)}\Big| y_t, y_{t-1}\right\}. \end{aligned} \tag{5.3}$$

Suponha que a função alvo possa ser representa como uma combinação linear (possivelmente de dimensão infinita) de funções base da forma

$$f(y) = \sum_{j=1}^{J} c_j B_j(y),$$

onde $J \in \mathbb{Z}_+ \cup \infty$ e que usemos a aproximação

$$f_{\boldsymbol{\theta}}(y) = \sum_{j=1}^{M} c_j B_j(y),$$

onde $M \in \mathbb{Z}_+$. Defina a matriz de planejamento e o vetor de coeficientes

$$\mathbf{B} = \begin{bmatrix} B_1(y_1) & B_2(y_1) & ... & B_M(y_1) \\ \vdots & \vdots & ... & \vdots \\ B_1(y_{T-1}) & B_2(y_1) & ... & B_M(y_{T-1}) \end{bmatrix} \text{ e } \mathbf{c} = \begin{bmatrix} c_1 \\ \vdots \\ c_M \end{bmatrix},$$

de modo que $(f_{\boldsymbol{\theta}}(y_1), ..., f_{\boldsymbol{\theta}}(y_{T-1})) = \mathbf{Bc}$. Note que agora o vetor $\mathbf{c}$ está contido em $\boldsymbol{\theta}$. Além disso, defina

$$C(\boldsymbol{\zeta}|\boldsymbol{\theta}^{(k)}) \equiv \sum_{t=2}^{T} E_{\boldsymbol{\theta}^{(k)}}\left\{\log\frac{h(\sigma_t|\boldsymbol{\zeta})}{\psi_{\boldsymbol{\theta}}(\sigma_t)}\Big| y_t, y_{t-1}\right\}$$

e

$$W_T(\boldsymbol{\theta}^{(k)}) \equiv \text{diag}\left(E_{\boldsymbol{\theta}^{(k)}}\left\{\frac{1}{\psi_{\boldsymbol{\theta}}(\sigma_2)^2}\Big| y_2, y_1\right\}, ..., E_{\boldsymbol{\theta}^{(k)}}\left\{\frac{1}{\psi_{\boldsymbol{\theta}}(\sigma_T)^2}\Big| y_T, y_{T-1}\right\}\right),$$

de modo que

$$Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(k)}) = -(T-1)\log\delta - \frac{1}{2\delta^2}(\mathbf{y}_{2:T} - \mathbf{Bc})' W_T(\boldsymbol{\theta}^{(k)})(\mathbf{y}_{2:T} - \mathbf{Bc}) + C(\boldsymbol{\zeta}; \boldsymbol{\theta}^{(k)})$$

onde $\mathbf{y}_{2:T} \equiv (y_2, ..., y_T)'$.

Derivando Q em relação a $\mathbf{c}$ e igualando o resultado a zero temos que, dado $\boldsymbol{\theta}^{(k)}$, o estimador de $\mathbf{c}$ é dado por

$$\mathbf{c}^{(k+1)} = (\mathbf{B}' W_T(\boldsymbol{\theta}^{(k)})\mathbf{B})^{-1}\mathbf{B}' W_T(\boldsymbol{\theta}^{(k)})\mathbf{y}_{2:T}.$$

[1] O último termo a direita da igualdade na expressão (5.3) deve ser considerada apenas se estivermos interessados em $\boldsymbol{\zeta}$ e/ou ψ depender de $\boldsymbol{\theta}$.

Anaogamente e usando a estimativa $\mathbf{c}^{(k+1)}$, temos que

$$(\delta^2)^{(k+1)} = \frac{1}{T-1}(\mathbf{y}_{2:T} - \mathbf{B}\mathbf{c}^{(k+1)})' W_T(\boldsymbol{\theta}^{(k)})(\mathbf{y}_{2:T} - \mathbf{B}\mathbf{c}^{(k+1)}).$$

Finalmente, o estimador de $\boldsymbol{\zeta}$, dado $\boldsymbol{\theta}^{(k)}$, é obtido derivando-se $C(\boldsymbol{\zeta}|\boldsymbol{\theta}^{(k)})$ com relação a $\boldsymbol{\zeta}$ e igualando-se o resultado a zero.

O Caso Canônico

Assim como no caso do modelo de regressão semiparamétrica, consideraremos aqui principalmente o caso em que $\psi(s) = s^{-1/2}$, isto é, o caso canônico. Nestas circunstâncias,

$$W_T(\boldsymbol{\theta}) = \text{diag}\left(E_{\boldsymbol{\theta}}\{\sigma_2|y_2, y_1\}, ..., E_{\boldsymbol{\theta}}\{\sigma_T|y_T, y_{T-1}\}\right)$$

e

$$k(\sigma_T, ..., \sigma_2|y_T, ..., y_1; \boldsymbol{\theta}) \propto \prod_{t=2}^{T} \sqrt{\sigma_t}\phi\left(\frac{r_t}{\delta/\sqrt{\sigma_t}}\right) h(\sigma_t|\boldsymbol{\zeta}). \tag{5.4}$$

Exemplos

Os exemplos abaixo foram retirados do capítulo 4 sobre regressão. Eles diferem daqueles pelo simples fato de o par (y_t, x_t) é agora dado por (y_t, y_{t-1}). Em todos eles $r_t = y_t - f(y_{t-1})$.

Exemplo 5.1.1 (Distribuição t de Student — ν Conhecido). *Assim como no caso do modelo de regressão semiparamétrica, suponha que o ruído do modelo seja distribuido de acordo com uma distribuição t com ν graus de liberdade, o qual assumiremos conhecido, e com parâmetro de escala ζ^2. Ou seja, assuma que $\epsilon_t \sim t_\nu(0, \zeta^2)$. Isto é o mesmo que assumir que $y_t|y_{t-1} \sim t_\nu(f(y_{t-1}), \delta^2)$. Então, para estimar a função alvo, modelaremos os dados de acordo com*

$$y_t|y_{t-1}; \sigma_t \sim \mathcal{N}\left(f(y_{t-1}), \frac{\delta^2}{\sigma_t}\right),$$
$$\sigma_t \sim \Gamma\left(\frac{\nu}{2}, \frac{2}{\nu}\right),$$

pois, sabe-se que dado o modelo acima, a distribuição marginal de $y_t|y_{t-1}$ é uma $t_\nu(f(y_{t-1}), \delta^2)$, ver [17]. Como antes, $\psi(\sigma) = \frac{1}{\sqrt{\sigma}}$ e h é a densidade associada à distribuição $\Gamma\left(\frac{\nu}{2}, \frac{2}{\nu}\right)$.

Procedendo de modo exatamente análogo ao realizado no caso do modelo semiparamétrico de regressão, temos que

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}') = -\frac{(T-1)}{2}\log\delta^2 - \frac{1}{2\delta^2}(\mathbf{y}_{2:T} - \mathbf{B}\mathbf{c})' W_T(\boldsymbol{\theta}')(\mathbf{y}_{2:T} - \mathbf{B}\mathbf{c}), \tag{5.5}$$

onde

$$W_T(\boldsymbol{\theta}') \equiv diag(\bar{\sigma}_1, ..., \bar{\sigma}_T) \tag{5.6}$$

e $\bar{\sigma}_t = E_{\boldsymbol{\theta}'}\{\sigma_t | y_t, y_{t-1}\}$. *Além disso, teremos*

$$\sigma_t | y_t, y_{t-1}; \boldsymbol{\theta}' \sim \Gamma\left(\frac{\nu+1}{2}, \frac{2}{\nu + r_t^2/\delta^2}\right),$$

de modo que

$$\bar{\sigma}_t = \frac{\nu+1}{\nu + r_t^2/\delta^2}. \tag{5.7}$$

Assim como para regressão, expressão (5.7) *indica que quanto maior a distância entre a observação e a curva ajustada, menor será o peso dada a esta observação durante o processo de estimação. O parâmetro de escala tem efeito semelhante, no entanto, sua atuação é uniforme sobre todas as variáveis.* ◊

O exemplo a seguir explicita os pesos obtidos quando assumimos uma distribuição de Cauchy para o ruído.

Exemplo 5.1.2 (Distribuição de Cauchy). *Como a distribuição de Cauchy é um caso particular da distibuição t de Student* ($\nu = 1$), *temos que*

$$\sigma_t | y_t, y_{t-1}; \boldsymbol{\theta}' \sim \Gamma\left(2, \frac{2}{1 + r_t^2/\delta^2}\right)$$

e que, conseqüentemente,

$$\bar{\sigma}_t = \frac{2}{2 + r_t^2/\delta^2}.$$

◊

O exemplo a seguir explora a distribuição exponencial dupla,

$$p(\epsilon) = \frac{1}{2} e^{-|\epsilon|}.$$

Como observado anteriormente, usar a distribuição exponencial dupla equivale a estimar os parâmetros através da minimização da norma L^1.

Exemplo 5.1.3 (Distribuição Exponencial Dupla). *Seguindo os mesmos passos do exemplo 4.2.3, temos que*

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}') = -(T-1)\log\delta - \frac{1}{2\delta^2}(\mathbf{y}_{2:T} - \mathbf{B}\mathbf{c})' W_T(\boldsymbol{\theta}')(\mathbf{y}_{2:T} - \mathbf{B}\mathbf{c})$$

e que $W_T(\boldsymbol{\theta}') = diag(\bar{\sigma}_1, ..., \bar{\sigma}_T)$ *onde as ponderações* $\bar{\sigma}_t$ *são iguais a* $E_{\boldsymbol{\theta}'}\left\{\sigma_t \middle| z_t\right\}$ *e dadas por*

$$\bar{\sigma}_t = \frac{\delta}{|r_t|}.$$

◊

Para o caso da distribuição logística, temos

Exemplo 5.1.4 (Distribuição Logística). *De modo análogo ao exemplo 4.2.4, temos que para a distribuição logística,*

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}') = -(T-1)\log\delta - \frac{1}{2\delta^2}(\mathbf{y}_{2:T} - \mathbf{Bc})' W_T(\boldsymbol{\theta}')(\mathbf{y}_{2:T} - \mathbf{Bc})$$

e que $W_T(\boldsymbol{\theta}') = diag(\bar{\sigma}_1, ..., \bar{\sigma}_T)$ *onde as ponderações* $\bar{\sigma}_t$ *são iguais a* $E_{\boldsymbol{\theta}'}\left\{\sigma_t \middle| z_t\right\}$ *e dadas por*

$$\bar{\sigma}_t = \frac{\delta}{|r_t|} \frac{\sum_{k=1}^{\infty}(-1)^{k-1}k^2 e^{-\frac{\sqrt{k}|r_t|}{\delta}}}{\sum_{k=1}^{\infty}(-1)^{k-1}k^{3/2} e^{-\frac{\sqrt{k}|r_t|}{\delta}}}.$$

◊

Finalmente, consideramos a distribuição normal contaminada.

Exemplo 5.1.5 (Distribuição Normal Contaminada). *No caso da distribuição normal contaminada, o operador* $Q(\boldsymbol{\theta}|\boldsymbol{\theta}')$ *e a matriz* $W_T(\boldsymbol{\theta}')$ *são iguais aos dos exemplos acima. Com relação às ponderações* $\bar{\sigma}_t$*, temos, assim como no exemplo 4.2.5, para regressão, que*

$$\bar{\sigma}_t = \sum_{j=0}^{K} \pi_{j|t}\lambda_j.$$

onde $\pi_{j|t} = \frac{k(\sigma_t = \lambda_j^2|y_t, \boldsymbol{\theta}')}{C_t}$ *e* λ_j^2 *representa a variância da* j*-ésima fonte de contaminação, para* $j = 0, ..., K$*, onde*

$$C_t = \sum_{j=0}^{K} \frac{\lambda_j}{\delta}\phi\left(\frac{r_t}{\delta/\lambda_j}\right)\pi_j.$$

Lembramos que K *representa a quantidade de fontes de ruído e que a distribuição a posteriori* $k(\cdot|y_t, \boldsymbol{\theta}')$ *é discreta (veja o exemplo 4.2.5 para mais detalhes).* ◊

Algoritmo 5.1.1 Resumo do algoritmo de estimação para dados de séries temporais.

Passo 1. obter uma estimativa inicial para $\mathbf{f} = (f(y_1), ..., f(y_{T-1}))$, denotada por $\mathbf{f}^{(0)}$;

- um modo de se obter a estimativa inicial é através do método de estimação usual associado ao suavizador escolhido, ie, supondo o ruído i.i.d. e gaussiano.

Passo 2. obter uma estimativa inicial para o parâmetro de escala δ e para ζ, denotados por $\delta^{(0)}$ e $\zeta^{(0)}$, respectivamente;

Passo 3. *loop* principal do algoritmo EM.

- Enquanto $|\text{EQM}(\hat{f}^{(k)}) - \text{EQM}(\hat{f}^{(k-1)})| \geq \text{tol}$:
  Passo $(k+1)$**.1:** $\mathbf{f}^{(k+1)} = \mathbf{B}\mathbf{c}^{(k+1)}$, onde

$$\mathbf{c}^{(k+1)} = (\mathbf{B}'W_T(\boldsymbol{\theta}^{(k)})\mathbf{B})^{-1}\mathbf{B}'W_T(\boldsymbol{\theta}^{(k)})\mathbf{y}_{2:T};$$

  Passo $(k+1)$**.2:** atualizar a $(k+1)$-ésima estimativa de δ^2 por

$$\frac{1}{T-1}(\mathbf{y}_{2:T} - \mathbf{f}^{(k+1)})'W_T(\mathbf{c}^{(k+1)}, \delta^{(k)}, \zeta^{(k)})(\mathbf{y}_{2:T} - \mathbf{f}^{(k+1)});$$

  Passo $(k+1)$**.3:** atualizar a $(k+1)$-ésima estimativa de ζ através da equação

$$\frac{\partial}{\partial \zeta}C(\zeta; \mathbf{c}^{(k+1)}, \delta^{(k+1)}, \zeta^{(k)}) = 0.$$

5.1.2 O Algoritmo de Estimação

A metodologia sugerida acima para estimação da função alvo usando dados de séries temporais pode ser vista resumidamente no algoritmo 5.1.1. Neste algoritmo usamos o mesmo critério de parada sugerido na seção 4.2.3.

5.1.3 Bandas de Confiança via Bootstrap

Novamente, uma maneira de se obter bandas de confiança para a estimativa da função alvo, do parâmetro de escala e do vetor de parâmetros ζ é através da técnica *bootstrap*. No caso em que a componente estocástica é comandada por inovações estocásticas, uma amostra *bootstrap* é obtida via Monte-Carlo amostrando-se o ruído, ou o resíduo, ϵ um número de vezes para, com o resultado, construir uma nova amostra de variáveis resposta.

Quando tratamos de dados dependentes, certos cuidados devem ser tomados, pois, ao contrário do caso i.i.d., o processo gerador dos dados não é, em geral, completamente especificado. Logo, não existe uma única maneira de reamostrar os dados. De qualquer modo, a reamostragem deve ocorrer de tal modo que a estrutura de dependência original seja capturada nas novas amostras. Em alguns modelos, tais como os

modelos finito-dimensionais ARMA, isto pode ser facilmente obtido considerando-se os resíduos i.i.d., mas em outros modelos isto não ocorre assim tão facilmente. Entre os métodos mais populares de *bootstrap* para dados dependentes estão o *bootstrap* por blocos, *sieve*, de regressão, local e *wild*. Todos são métodos não paramétricos. Para mais detalhes, veja [18] e as referências lá contidas.

Aqui consideraremos o *bootstrap* por regressão, onde a amostra *bootstrap* y_t^*, $t = 1, 2, ...$, é obtida amostrando-se com reposição ϵ_t^* (*bootstrap* não-paramétrico), $t = 1, 2, ...$, a partir dos resíduos centralizados[2] no modelo (5.1), de modo que

$$y_t^* = \hat{f}(y_{t-1}) + \hat{\delta}\epsilon_t^*$$

onde $\hat{f}$ é uma estimativa obtida para f. Note que, deste modo, um modelo de regressão não-paramétrico é gerado com planejamento (condicionalmente) fixado, daí o nome deste tipo de *bootstrap*. Como a série temporal original é usada como covariáveis em um problema de regressão, o *bootstrap* por regressão tem a vantagem de não ser tão sensível em relação à estimativa da função alvo f.

Bootstrap Paramétrico

No caso do *bootstrap* paramétrico, usamos a mesma metodologia acima, mas agora considerando o fato que

$$\epsilon_t|\sigma_t \sim \mathcal{N}(0, \psi(\sigma_t)^2),$$
$$\sigma_t \sim h.$$

Neste caso, podemos obter por simulação os valores $\sigma_1^*, ..., \sigma_T^*$ de acordo com a densidade h e depois, para cada $t = 1, ..., T$, obter os valores $\epsilon_1^*, ..., \epsilon_T^*$ de acordo com a distribuição $\mathcal{N}(0, \psi(\sigma_t^*)^2)$. O passo seguinte consiste em definir

$$y_t^* \equiv \hat{f}(y_{t-1}) + \hat{\delta}\epsilon_t^*, \text{ para } t = 2, ..., T.$$

O procedimento acima é repetido B vezes e, para cada amostra *bootstrap*, obtemos uma estimativa *bootstrap* $\hat{\mathbf{f}}_j^*$ e, com estas estimativas, construímos as bandas de confiança desejadas.

Observação 5.1.1. *Como notado em [28] e nas referências lá contidas, na hipótese do ruído possuir variância infinita, tanto para modelos de regressão quanto para séries temporais, os métodos usuais de bootstrap, como os descritos acima, falham drasticamente caso a condição $B = o(T)$ seja violada. Em particular, no nosso caso, este cuidado deve ser tomado quando assumirmos uma distribuição estável para o ruído.*

5.2 Estudos de Simulação

Nesta seção apresentamos uma série de simulações para avaliar empiricamente a metodologia descrita acima. Os dois primeiros estudos comparam a metodologia sugerida com as estimativas usuais via mínimos quadrados. Nos demais estudos avaliamos a qualidade das estimativas usando bandas de confiança obtidas via

[2] o t-ésimo resíduo centralizado $\hat{\epsilon}_t$ é definido por

$$y_t - \hat{f}(y_{t-1}) - \frac{1}{T-1}\sum_{j=2}^{T}(y_j - \hat{f}(y_{j-1})).$$

bootstrap. As funções alvo utilizadas nos dois primeiros estudos são dadas pelas expressões (5.8) e (5.9), representadas nas figuras 5.1 e 5.2, respectivamente.

Função alvo 1: (figura 5.1)

$$f(x) = \text{sen}(2x); \tag{5.8}$$

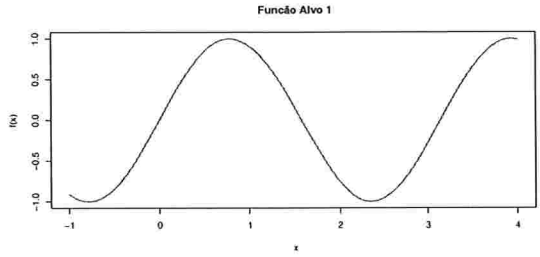


Figura 5.1: Gráfico da função alvo (5.8).

Função alvo 2: (figura 5.2)

$$f(x) = 10y(1-y) \tag{5.9}$$

onde $y \equiv x \bmod 1$;

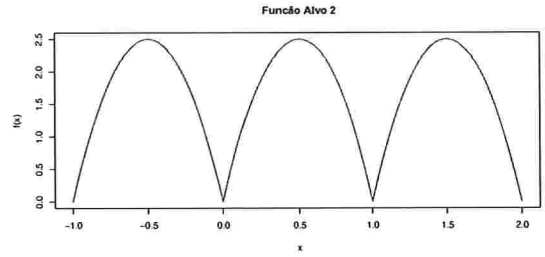


Figura 5.2: Gráfico da função alvo (5.9).

Estudo 1:

Nesta primeira bateria de simulações consideramos um ruído distribuido de acordo com a distribuição t de Student com graus de liberdade conhecido e utilizamos *B-Splines* para aproximar a função alvo, neste caso, dada por (5.8).

A figura 5.3 contém os resultados das estimativas obtidas sobre uma amostra de tamanho 1000 gerada de acordo com uma t de Student com $1,0$ (distribuição de Cauchy), $1,5$, $2,0$ e $6,0$ graus de liberdade, respectivamente. Em todas estas simulações, o parâmetro de escala, δ, foi fixado igual a $0,5$. Para estimar a função alvo e o parâmetro de escala, *B-splines* com 25 graus de liberdade. Para estas simulações, as estimativas para

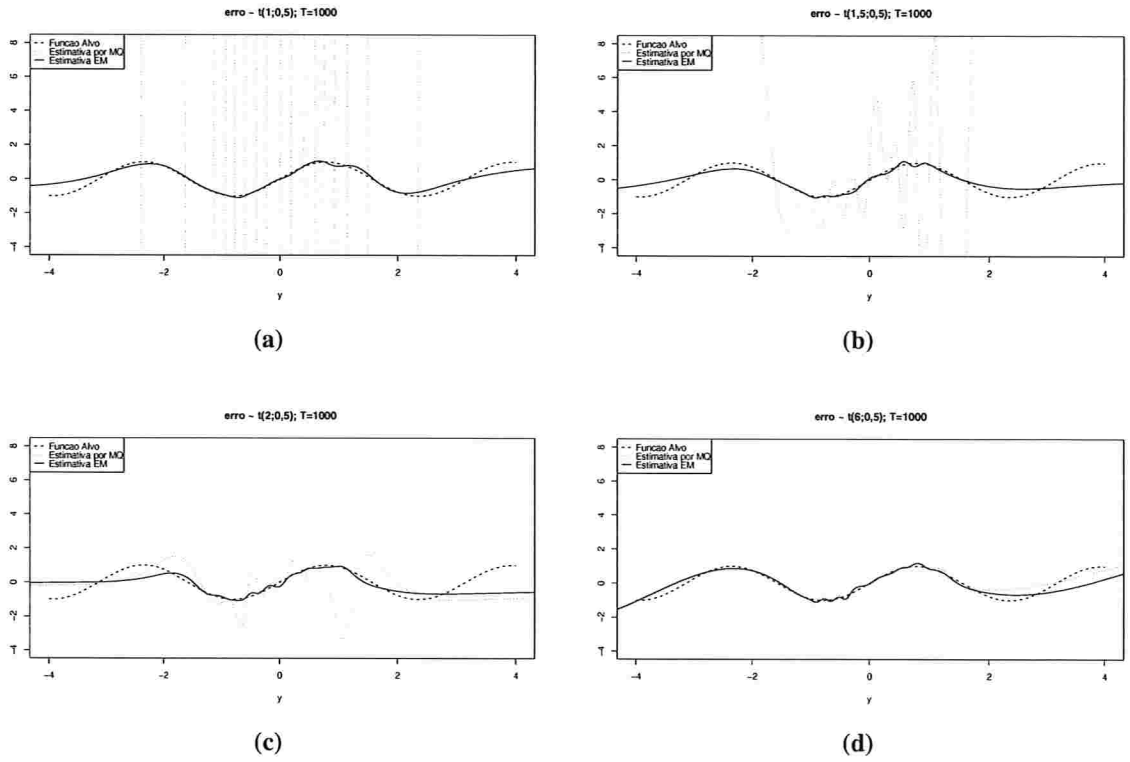


Figura 5.3: Estimativas via o algoritmo EM e por mínimos quadrados para o estudo de simulação #1 nos casos em que o ruído segue uma distribuição t de Student com graus de liberdade iguais a (a) $\nu = 1,0$, (b) $\nu = 1,5$, (c) $\nu = 2,0$ e (d) $\nu = 6,0$. A linha cheia representa a estimativa de acordo com o método iterativo, a linha tracejada a função alvo e a linha pontilhada, a estimativa via mínimos quadrados.

o parâmetro de escala, δ, estão resumidas na tabela 8.1 Tanto através dos gráficos quanto das estimativas obtidas para o parâmetro de escala, podemos notar notar a superioridade do método iterativo sobre as estimativas de mínimos quadrados, principalmente quando a variância do ruído é maior. Note que ambas as estimativas convergem quando o número de graus de liberdade associado à distribuição t aumenta.

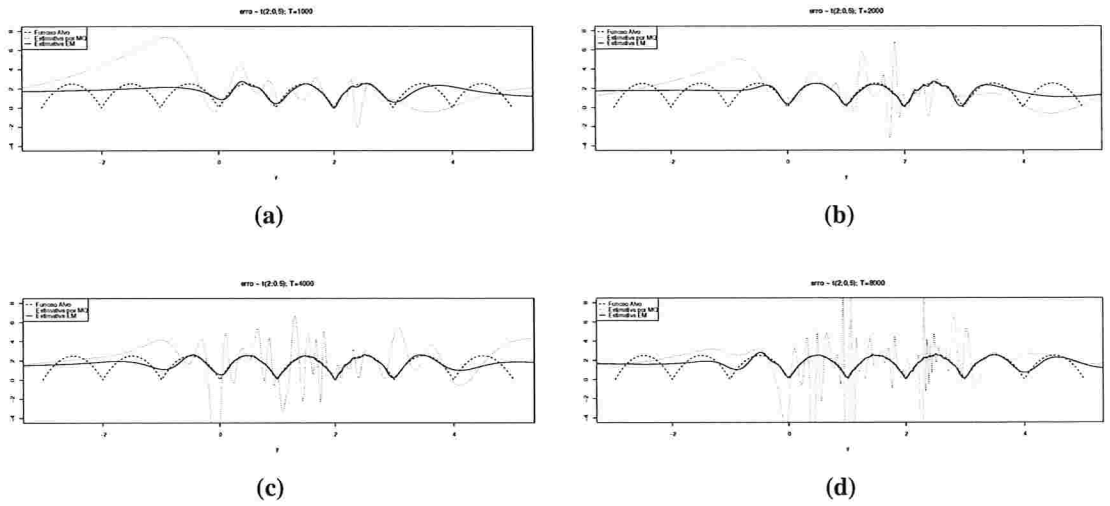
Estudo 2:

Neste estudo de simulação avaliamos o efeito do tamanho amostral sobre as estimativas obtidas, tanto pelo método sugerido quanto via mínimos quadrados. Nestas simulações mantivemos fixados os graus de liberdade

Tabela 5.1: Estimativas do parâmetro de escala, $\delta = 0,5$, para ruídos com diferentes graus de liberdade obtidas pelo método sugerido e via mínimos quadrados para o estudo de simulação #1.

ν	**Estimativa EM**	**Estimativa M.Q.**
1	0,8448	∞
1,5	0,8309	14272,62
2	0,7009	70,4637
6	0,6460	0,9412

da distribuição t de Student em $\nu = 2$ e variamos o tamanho da amostra considerando os seguintes valores $T = 1000$, $T = 2000$, $T = 4000$ e $T = 8000$. Em todas as simulações assumimos $\delta = 0.5$ e tomamos (5.9) como função alvo. Além disso, variamos os graus de liberdade associados às *B-splines* de modo que eles fossem aproximadamente iguais a $\sqrt{T}$. As estimativas obtidas para a função alvo podem ser visualizadas na figura 5.4. Note que, com relação às estimativas via mínimos quadrados há uma deterioração na qualidade das mesmas conforme aumentamos o tamanho amostral — o mesmo não ocorre com relação à metodologia sugerida — devido ao aumento de valores extremos que ocorre como consequência do aumento no número de observações. Além disso, é interessante notar o que ocorre nos extremos dos gráficos na figura 5.4. Nestas localizações, as estimativas perdem qualidade e isto ocorre devido à ausência de observações nestes extremos. Obviamente, conforme aumentamos o número de observações, a estimativa obtida identifica-se com um trecho maior da função alvo, a qual está definida para toda a reta real. O mesmo efeito ocorre com relação às



(a) (b)

(c) (d)

Figura 5.4: Estimativas da função alvo utilizada no estudo de simulação #2 via o algoritmo EM e por mínimos quadrados nos casos em que os tamanhos amostrais são iguais a 1000, 2000, 4000 e 8000 (gráficos (a), (b), (c) e (d), respectivamente). As linhas cheias representam a estimativa obtida de acordo com o método sugerido, as linhas tracejadas representam função alvo e as linhas pontilhadas, as estimativas obtidas via mínimos quadrados.

estimativas obtidas para o parâmetro de escala, δ, as quais estão resumidas na tabela 5.2

Tabela 5.2: Estimativas do parâmetro de escala obtidas no estudo de simulação #2, $\delta = 0,5$, para ruídos distribuidos de acordo uma distribuição t de Student com 2 graus de liberdade obtidos pelo método sugerido e via mínimos quadrados para diversos tamanhos amostrais.

T	Estimativa EM	Estimativa M.Q.
1000	0,8298	99,2108
2000	0,7633	114,8684
4000	0,7700	327,836
8000	0,7514	3632,441

Estudo 3:

O objetivo deste estudo é avaliar a qualidade das estimativas obtivas através de bandas de confiança obtidas via bootstrap. Nele consideramos a função alvo

$$f(x) = \frac{\text{sen}(20(x + 0.2))}{x + 0.2}$$

(figura 4.1 (c)) e assumimos como parâmetro de escala $\delta = 0,5$. Os dados foram obtidos via simulação assumindo-se que o ruído seguia uma distribuição Cauchy de modo a gerar uma amostra de tamanho 300. Foram realizados 3 ajustes, sendo que, para todos eles, consideramos $M = 40$ (graus de liberdade associados à base *B-splines* utilizada) e obtivemos amostras *bootstrap* de tamanho 200 com base nas quais construímos intervalos de confiança (95%) para δ e bandas de confiança (95%) para as estimativas da função alvo. O primeiro ajuste (Ajuste 1) foi feito segundo a metodologia proposta na tese e tomando $\nu = 1$, o segundo ajuste também foi realizado segundo a metodologia proposta, porém, com $\nu = 3$ e, finalmente, o terceiro ajuste (Ajuste 3) foi feito via mínimos quadrados. O resultado, relativo à estimativa da função alvo, deste estudo pode ser visto na figura 5.5. Quanto ao parâmetro de escala, obtivemos os resultados expostos na tabela 5.3. Pode-se notar, tanto pelos ajustes da função alvo, quanto pelas estimativas do parâmetro de escala

Tabela 5.3: Estimativas e intervalos de confiança para o parâmetro de escala $\delta = 0,5$ e o erro quadrático para cada ajuste obtidas no estudo de simulação #3 onde o ruído foi gerado segundo uma distribuição de Cauchy.

Ajuste	*bootstrap*	$\hat{\delta}$	$IC(95\%)$	EQM
1	paramétrico	0,524	[0,371;0,506]	0,2290
1	não-paramétrico	0,524	[0,372;0,540]	0,2290
2	paramétrico	1,056	[0,872;1,071]	0,2124
2	não-paramétrico	1,056	[0,817;1,169]	0,2124
M.Q.	paramétrico	5,770	[5,036;5,780]	2,8824
M.Q.	não-paramétrico	5,770	[3,684;6,805]	2,8824

que o método sugerido adequa-se melhor à presença de valores extremos no conjunto de observações. Isto ocorre mesmo quando a distribuição assumida para o ruído não corresponde à real distribuição.

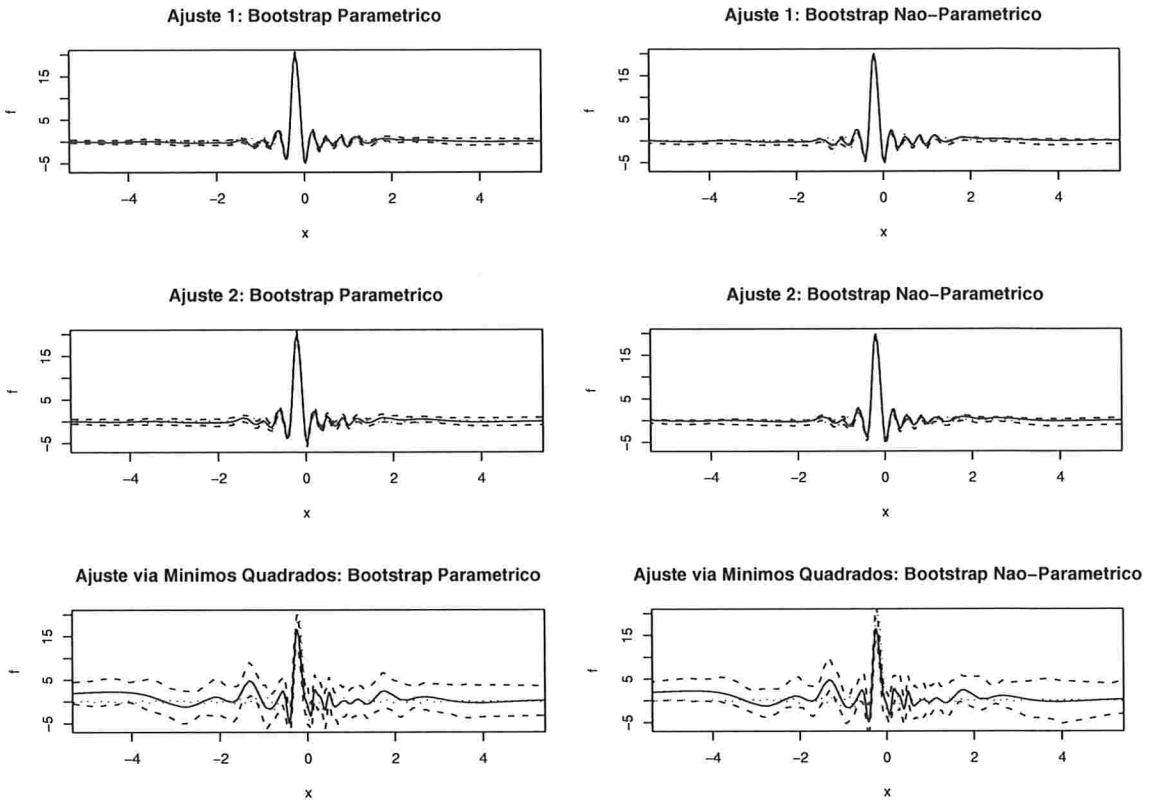


Figura 5.5: Estimativas da função alvo para o estudo de simulação #3 segundo a metodologia proposta, (Ajuste1) e (Ajuste 2), e via mínimos quadrados, (Ajuste 3) e suas respectivas bandas de confiança obtidas via *bootstrap*.

5.3 Extensão para o Modelo Parcialmente Linear

Consideraremos agora uma extensão da metodologia usada nas seções anteriores para séries autorregressivas não lineares de ordem 1 para modelos autorregressivos de ordem p. A extensão mais simples é o modelo parcialmente linear, dado por

$$y_t = f(y_{t-1}) + \boldsymbol{\beta}' \mathbf{u}_t + \delta \epsilon_t.$$

O vetor $\mathbf{u}_t$ pode ao mesmo tempo representar realizações passadas do processo estocástico Y de modo que, por exemplo, $\mathbf{u}_t = (y_{t-2}, ..., y_{t-p})'$, para algum $p \geq 2$, como pode também representar outros regressores ou covariáveis, ou pode até mesmo ser uma mistura de ambos. Embora estejamos convencionando que a componente não-linear esteja associada à realização do processo Y imediatamente anterior àquela observada, isto não deve servir de regra e, portanto, no instante t, podemos assumir que y_{t-k} seja a componente não-linearmente relacionada com y_t, para algum $k \in \mathbb{Z}$.

Analogamente ao modelo parcialmente linear de regressão, modelamos os dados de acordo com

$$y_t|y_{t-1}, \mathbf{u}_t; \sigma_t \sim \mathcal{N}(f(y_{t-1}) + \boldsymbol{\beta}'\mathbf{u}_t, \delta^2\psi(\sigma_t)^2),$$
$$\sigma_t \sim h,$$

para $t = p^* + 1, ..., T$, onde p^* representa a quantidade de realizações passadas de Y consideradas na componente linear e na componente não-linear. Por exemplo, no caso em que $\mathbf{u}_t = (y_{t-2}, ..., y_{t-p})'$, temos $p^* = p$ e y_1,...,y_p são considerados como determinísticos. Note que, em qualquer caso, $p^* \geq 2$. Aproximando a função alvo f como a soma de funções base, $(B_1, ..., B_M)$, teremos, analogamente à seção 4.5, o vetor de parâmetros $\boldsymbol{\theta} = (\mathbf{c}', \boldsymbol{\beta}', \delta^2, \boldsymbol{\zeta}')'$ onde $\mathbf{c}$ corresponde aos coeficientes da expansão (aproximação) de f em função de B_1,...,B_M. Conseqüentemente, no k-ésimo passo do algoritmo,

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k-1)}) = C(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k-1)}) - (T - p^*)\log\delta$$
$$- \frac{1}{2\delta^2}(\mathbf{y}^* - \mathbf{B}\mathbf{c} - \mathbf{U}\boldsymbol{\beta})'W_T(\boldsymbol{\theta}^{(k-1)})(\mathbf{y}^* - \mathbf{B}\mathbf{c} - \mathbf{U}\boldsymbol{\beta}),$$

onde $\mathbf{y}^* = \mathbf{y}_{(p^*+1):T}$,

$$C(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k-1)}) = \sum_{t=p^*}^{T} E_{\boldsymbol{\theta}^{(k-1)}}\left\{\log\frac{h(\sigma_t|\boldsymbol{\zeta}^{(k-1)})}{\psi_{\boldsymbol{\theta}}(\sigma_t)}\middle| y_t, \mathbf{u}_t\right\}$$

e

$$W_T(\boldsymbol{\theta}^{(k-1)}) = \text{diag}\left(E_{\boldsymbol{\theta}^{(k-1)}}\left\{\frac{1}{\psi_{\boldsymbol{\theta}}(\sigma_t)^2}\middle| y_t; \mathbf{u}_t\right\}\right)_{t=p^*+1}^{T}.$$

As esperanças acima são calculadas de acordo com a densidade "a posteriori"

$$k(\sigma_t|y_t, \mathbf{u}_t) \propto \frac{1}{\delta\psi_{\boldsymbol{\theta}}(\sigma_t)}\phi\left(\frac{y_t - \sum_{j=1}^{M} c_j B_j(y_{t-1}) - \boldsymbol{\beta}'\mathbf{u}_t}{\delta\psi_{\boldsymbol{\theta}}(\sigma_t)}\right) h(\sigma_t|\boldsymbol{\zeta}^{(k-1)}),$$

para $t = p^* + 1, ..., T$. Em particular, no caso em que $\mathbf{u}_t = (y_{t-2}, ..., y_{t-p})'$, a matriz $\mathbf{U}$ (de dimensão $(T-p) \times (p-1)$) é dada por

$$\mathbf{U} = \begin{bmatrix} y_{p-1} & y_{p-2} & \cdots & y_1 \\ y_p & y_{p-1} & \cdots & y_2 \\ \vdots & \vdots & & \vdots \\ y_{T-2} & y_{T-3} & \cdots & y_{T-p} \end{bmatrix}$$

e

$$k(\sigma_t|y_t, \mathbf{u}_t) \propto \frac{1}{\delta\psi_{\boldsymbol{\theta}}(\sigma_t)}\phi\left(\frac{y_t - \sum_{j=1}^{M} c_j B_j(y_{t-1}) - \sum_{j=1}^{p-1}\beta_j y_{t-1-j}}{\delta\psi_{\boldsymbol{\theta}}(\sigma_t)}\right)$$
$$\cdot\, h(\sigma_t|\boldsymbol{\zeta}^{(k-1)}).$$

Note que, neste caso, as primeiras p observações são tomadas como determinísticas.

Tomando as derivadas parciais de Q, temos, com relação a $\mathbf{c}$,

$$\frac{\partial}{\partial\mathbf{c}}Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k-1)}) = \frac{1}{\delta}(\mathbf{B}'W_T(\boldsymbol{\theta}^{(k-1)})(\mathbf{y}^* - \mathbf{U}\boldsymbol{\beta}) - \mathbf{B}'W_T(\boldsymbol{\theta}^{(k-1)})\mathbf{B}\mathbf{c}),$$

com relação a β,

$$\frac{\partial}{\partial \beta} Q(\theta|\theta^{(k-1)}) = \frac{1}{\delta}(\mathbf{U}'W_T(\theta^{(k-1)})(\mathbf{y}^* - \mathbf{Bc}) - \mathbf{U}'W_T(\theta^{(k-1)})\mathbf{U}\beta),$$

e com relação a δ,

$$\frac{\partial}{\partial \delta} Q(\theta|\theta^{(k-1)}) = \frac{1}{\delta^3}(\mathbf{y}^* - \mathbf{Bc} - \mathbf{U}\beta)'W_T(\theta^{(k-1)})(\mathbf{y}^* - \mathbf{Bc} - \mathbf{U}\beta) - \frac{T-p^*}{\delta},$$

o que resulta no algoritmo 5.3.1.

Algoritmo 5.3.1 Esquematização do algoritmo EM para o tratamento de modelos parcialmente lineares na presença de um erro distribuído de acordo com uma distribuição t de Student para dados autocorrelacionados.

- Defina $\mathbf{y}^* = \mathbf{y}_{(p^*+1):T}$.
- Dados $\mathbf{c}^{(k)}$, $\beta^{(k)}$, $\delta^{(k)}$ e $\zeta^{(k)}$:

1. Calcular
$$\mathbf{c}^{(k+1)} = (\mathbf{B}'W_T(\theta^{(k)})\mathbf{B})^{-1}\mathbf{B}'W_T(\theta^{(k)})(\mathbf{y}^* - \mathbf{U}\beta^{(k)});$$
2. Calcular
$$\beta^{(k+1)} = (\mathbf{U}'W_T(\theta^{(k)})\mathbf{U})^{-1}\mathbf{U}'W_T(\theta^{(k)})(\mathbf{y}^* - \mathbf{Bc}^{(k+1)});$$
3. Calcular
$$(\delta^2)^{(k+1)} = \frac{1}{T-p^*}(\mathbf{r}^{(k+1)})'W_T(\theta^{(k)})\mathbf{r}^{(k+1)}$$
onde
$$\mathbf{r}^{(k+1)} = \mathbf{y}^* - \mathbf{Bc}^{(k+1)} - \mathbf{U}\beta^{(k+1)};$$
4. Obter $\zeta^{(k+1)}$ como solução de
$$\frac{\partial}{\partial \zeta} Q(\zeta; \mathbf{c}^{(k+1)}, \beta^{(k+1)}, \delta^{(k+1)}|\theta^{(k)}) = 0.$$

- se $EQM(\widehat{f}) < tol$, parar iteração.

5.3.1 Estudo de Simulação

Os dados utilizados no estudo de simulação abaixo foram obtidos assumindo-se um ruído distribuido conforme uma t de Student com $\nu = 1,5$ graus de liberdade e parâmetro de escala $\delta = 1,5$. A componente linear foi escolhida de modo que $\mathbf{u}_t = (y_{t-2}, y_{t-3})$ e seus coeficientes definidos por $\beta_2 = 0,64$ e $\beta_3 = 0,20$. A componente não-linear é representada pela função alvo

$$f(x) = \frac{x}{2(1+x^2)}$$

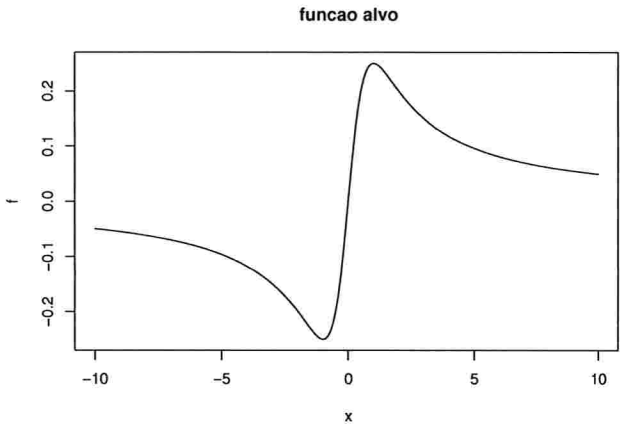


Figura 5.6: Função alvo utilizada no estudo de simulação para o modelo parcialmente linear.

representada na figura 5.6. Para aproximar a curva, optamos por utilizar B-splines, cujos graus de liberdade, M, foram escolhido com base no critério EAIC_c. Assumindo $\nu = 1, 5$, ajustamos curvas para M entre 6 e 16 e obtivemos o resultado (EAIC_c) exposto na figura 5.7, de modo que optamos por $M = 9$. O resultado da

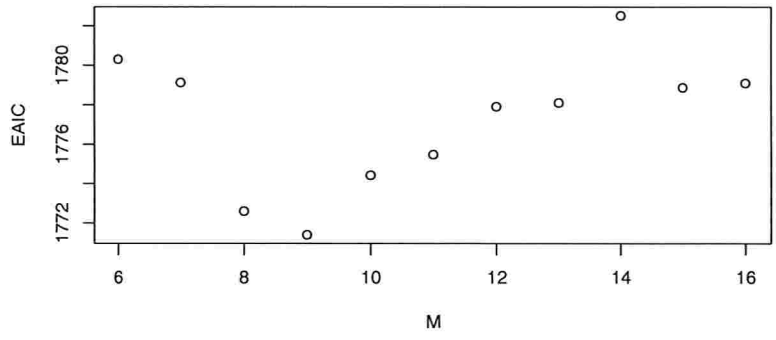


Figura 5.7: EAIC_c obtidos para M entre 6 e 16 no estudo de simulação para o modelo parcialmente linear.

estimativa para a função alvo pode ser visualizado na figura 5.8. Lá o gráfico a esquerda representa a estimativa obtida através do método sugerido, enquanto que no gráfico a direita temos o resultado obtido via mínimos quadrados. Em ambos os gráficos podem ser visualizada bandas de confiança a 95%. Como se pode notar, embora pouco satisfatória quando comparada a outros estudos de simulação nesta tese, ainda supera em muito a estimativa via mínimos quadrados. Em ambos os gráficos, consideramos um intervalo correspondente a 90% das observações. Nos extremos, como era de se esperar, ambas as estimativas são muito deterioradas pela

escassez de observações. Pelas bandas de confiança, podemos observar que a estimativa robusta é, de fato, bem sensível a valores extremos. As estimativas obtidas para os coeficientes associados à componente linear estão

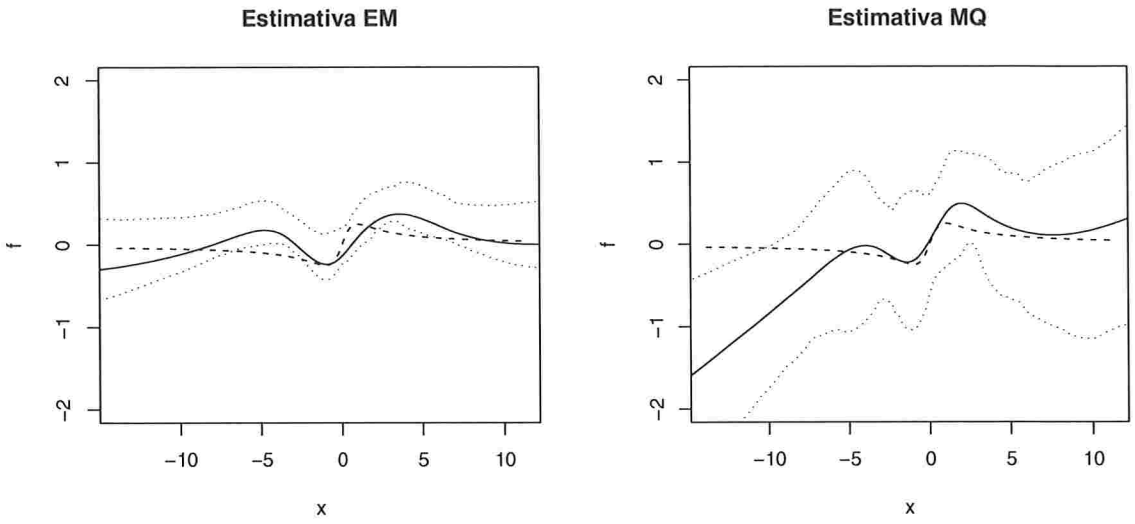


Figura 5.8: Estimativas obtidas para a função alvo através do método sugerido (esquerda) e via mínimos quadrados (direita) e suas respectivas bandas de confiança a 95% obtidas por *bootstrap*. Em ambas, assumimos $\nu = 1,5$.

na tabela 5.4. Novamente as estimativas obtidas segundo o método sugerido demonstram ser muito menos

Tabela 5.4: Estimativas dos coeficientes da componente linear do modelo calculados pelo método robusto (Estimativa EM), assumindo uma distribuição t de Student com $\nu = 1,5$, e via mínimos quadrados e seus respectivos intervalos de confiança a 95%.

	β	Estimativa EM	IC95%	Estimativa MQ	IC95%
β_1	0,64	0,6334	[0,6144;0,6492]	0,6421	[0,5923;0,6828]
β_2	0,20	0,2049	[0,1874;0,2280]	0,1147	[0,0549;0,1702]

suscetíveis a valores extremos do que as estimativas via mínimos quadrados e, embora as estimativas para β_1 tenham sido muito próximas em ambos os métodos, a estimativa para β_2 via EM se aproximou muito mais do valor real do que a obtida via M.Q.. Mesmo no caso de β_1, o intervalo de confiança para a estimativa robusta indica que valores estimados segundo o método robusto devem oscilar muito mais próximos do valor real do que estimativas obtidas via M.Q. Finalmente, para o parâmetro de escala, $\delta^2 = (1,5)^2 = 2,25$, obtivemos os resultados expostos na tabela 5.5. Assim como para a componente linear do modelo, a diferença entre as estimativas obtidas segundo o método robusto e o via M.Q. é imensa refletindo a diferença de sensibilidade em relação à presença de valores extremos entre ambos os métodos.

Para testar o efeito de se subestimar ν, repetimos a análise assumindo $\nu = 4$. Novamente, escolhemos o modelo de acordo com o critério EAIC_c (figura 5.9) e optamos por $M = 8$. O resultado das estimativas para

Tabela 5.5: Estimativas do parâmetro de escala, δ^2, calculados pelo método robusto (Estimativa EM), assumindo uma distribuição t de Student com $\nu = 1,5$, e via mínimos quadrados e seus respectivos intervalos de confiança a 95%.

	δ^2	Estimativa EM	IC95%	Estimativa MQ	IC95%
δ^2	2,25	2,1220	[2,0419;2,1114]	35,5020	[34,8863;35,4166]

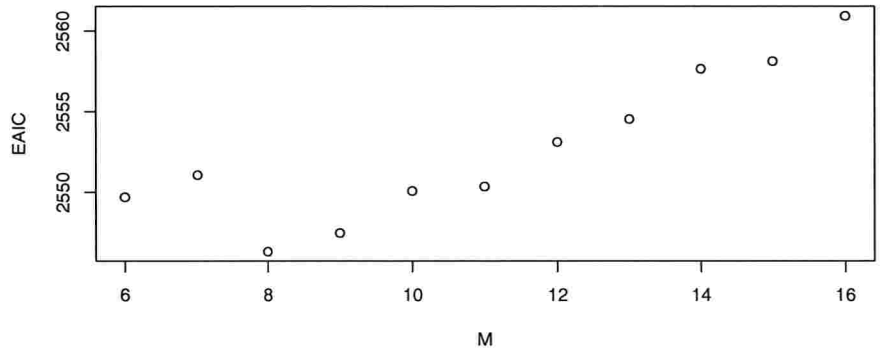


Figura 5.9: EAIC_c obtidos para M entre 6 e 16 no estudo de simulação para o modelo parcialmente linear para estimativas obtidas assumindo-se $\nu = 4$.

a função alvo e respectivas bandas de confiança podem ser visualizadas na figura 5.10. Os resultados obtidos para os coeficientes associados à componente linear, β, e para o parâmetro de escala, δ^2, podem ser vistos nas tabelas 5.6 e 5.7, respectivamente. Embora, em relação ao caso em que assumimos $\nu = 1,5$, as estimativas

Tabela 5.6: Estimativas dos coeficientes da componente linear do modelo calculados pelo método robusto (Estimativa EM), assumindo uma distribuição t de Student com $\nu = 4$, e via mínimos quadrados e seus respectivos intervalos de confiança a 95%.

	β	Estimativa EM	IC95%	Estimativa MQ	IC95%
β_1	0,64	0,6362	[0,6172;0,6540]	0,6362	[0,5999;0,7002]
β_2	0,20	0,2050	[0,1824;0,2287]	0,1137	[0,0503;0,1805]

Tabela 5.7: Estimativas do parâmetro de escala, δ^2, calculados pelo método robusto (Estimativa EM), assumindo uma distribuição t de Student com $\nu = 4$, e via mínimos quadrados e seus respectivos intervalos de confiança a 95%.

	δ^2	Estimativa EM	IC95%	Estimativa MQ	IC95%
δ^2	2,25	4,5514	[4,5269;4,6175]	35,5173	[34,3210;35,5173]

tenham se deteriorado, o resultado obtido ainda foi significativamente superior ao obtido via M.Q., indicando

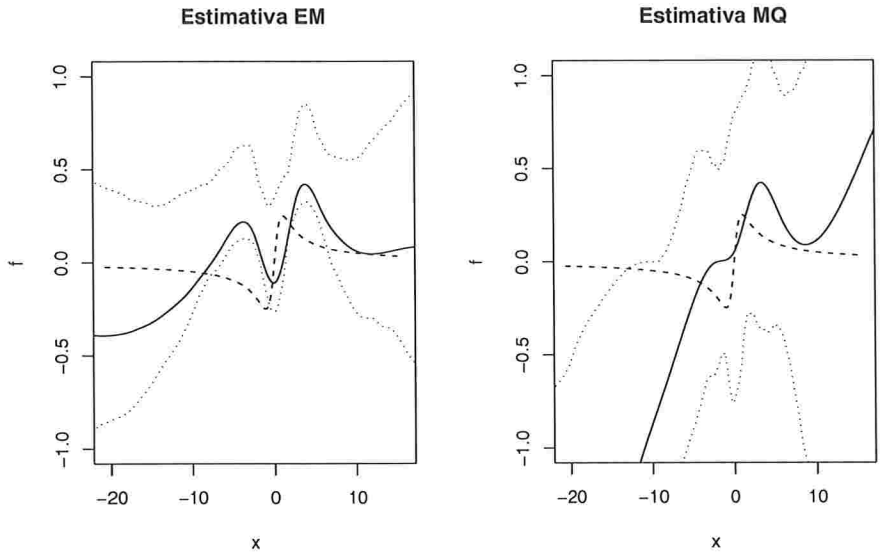


Figura 5.10: Estimativas obtidas para a função alvo através do método sugerido (esquerda) e via mínimos quadrados (direita) e suas respectivas bandas de confiança a 95% obtidas por *bootstrap*. Em ambas, assumimos $\nu = 4$.

assim que o método é robusto a perturbações na escolha da distribuição do ruído.

Capítulo 6

Análise Bayesiana

Nos capítulos anteriores, estimamos os parâmetros associados aos modelos considerados via máxima verossimilhança. Neste capítulo, estudaremos como analisá-los através da metodologia *bayesiana*.

Como antes, consideremos modelos do tipo

$$y_t = f(x_t) + \delta\epsilon_t, \tag{6.1}$$

para $t = 1, ..., T$, onde x_t pode ser determinística, aleatória ou mesmo igual a y_{t-1} e onde os erros ϵ_t são modelados de acordo com uma mistura de normais através da escala. Ou seja,

$$p(\epsilon_t) = \int_0^\infty \frac{1}{\psi_{\boldsymbol{\theta}}(\sigma_t)} \phi\left(\frac{\epsilon_t}{\psi_{\boldsymbol{\theta}}(\sigma_t)}\right) h(\sigma_t) d\sigma_t, \tag{6.2}$$

onde h é uma densidade de probabilidade definida em $[0, \infty)$ e $\psi_{\boldsymbol{\theta}}$ é uma função conhecida a priori. Em particular, (6.2) implica que

$$p(y_t|x_t) = \int_0^\infty \frac{1}{\delta\psi_{\boldsymbol{\theta}}(\sigma_t)} \phi\left(\frac{y_t - f(x_t)}{\delta\psi_{\boldsymbol{\theta}}(\sigma_t)}\right) h(\sigma_t) d\sigma_t.$$

É interessante notar também que, para $\sigma_t \sim h$, temos $\epsilon_t|\sigma_t \sim \mathcal{N}(0, \psi_{\boldsymbol{\theta}}(\sigma_t)^2)$. Finalmente, notamos que, embora devamos condicionar a distribuição de y_t em x_t, escreveremos de agora em diante $p(y_t)$ ao invés de $p(y_t|x_t)$ para simplificar a notação, ficando a dependência em x_t clara pelo contexto.

6.1 Aproximação de f via P-Splines

Para aproximar a função alvo, utilizaremos uma versão bayesiana de *P-splines*. Lembre que, como descrito no capítulo 2, *P-splines* são o resultado da combinação de *B-splines* como funções base com penalizações sobre coeficientes adjacentes da aproximação da função alvo via *B-splines*. De qualquer modo, independentemente da penalização aplicada, aproximaremos $f(\cdot)$ por $\sum_{j=1}^{M} a_j B_j(\cdot)$, onde os B_j são *B-splines* de ordem k. Observe que, usando *B-splines*, o modelo (6.1) pode ser escrito na forma matricial como

$$\mathbf{y} = B\mathbf{a} + \delta\boldsymbol{\epsilon}$$

onde

$$B = \begin{bmatrix} B_1(x_1) & \cdots & B_M(x_1) \\ \vdots & & \vdots \\ B_1(x_T) & \cdots & B_M(x_T) \end{bmatrix} \tag{6.3}$$

e $\mathbf{a} \equiv (a_1, ..., a_M)'$

A diferença entre os casos bayesiano e clássico, onde os coeficientes estimados são obtidos por mínimos quadrados penalizados, é que no enfoque bayesiano tratamos os coeficientes referidos acima como variáveis aleatórias às quais associamos uma distribuição a priori. As penalizações impostas são, então, replicadas impondo-se algumas restrições sobre estas variáveis aleatórias. Os detalhes desta operação estão descritos a seguir.

6.1.1 Variante Bayesiana da Aproximação via P-Splines

No enfoque bayesiano, as penalidades em (2.1) são substituidas pelos seus análogos estocásticos. Por exemplo, quando as primeiras diferenças Δa_j são penalizadas, consideramos um passeio aleatório simples, e quando as segundas diferenças são penalizadas, consideramos um passeio aleatório de segunda ordem. Mais precisamente, para os operadores de primeira e segunda diferença temos, respectivamente, que

$$a_j = a_{j-1} + u_j$$

e

$$a_j = 2a_{j-1} - a_{j-2} + u_j,$$

onde $\{u_j\}$ é um ruído branco. Normalmente, assume-se que $u_j|\tau^2 \sim \mathcal{N}(0, \tau^2)$ e que a_1, ou a_1 e a_2, possui priori não informativa. Como veremos mais adiante, o parâmetro τ^2 controla a quantidade de suavidade associada à função estimada e exerce papel análogo ao parâmetro de suavização, λ, no caso clássico.

Em particular, note que, se considerarmos um passeio aleatório de primeira ordem,

$$\begin{aligned} p(\mathbf{a}|\tau^2) &= p(a_M, ..., a_1|\tau^2) \\ &= p(a_M|a_{n-1}, \tau^2) \cdots p(a_2|a_1, \tau^2) \cdot p(a_1|\tau^2) \\ &\propto \exp\left(-\frac{(a_M - a_{n-1})^2}{2\tau^2}\right) \cdots \exp\left(-\frac{(a_2 - a_1)^2}{2\tau^2}\right) \\ &= \exp\left(-\frac{1}{2\tau^2}\sum_{j=1}^{M}(a_j - a_{j-1})^2\right) \\ &= \exp\left(-\frac{1}{2\tau^2}\mathbf{a}'K\mathbf{a}\right), \end{aligned}$$

onde

$$K = \begin{bmatrix} 1 & -1 & 0 & \cdots & 0 & 0 & 0 \\ -1 & 2 & -1 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -1 & 2 & -1 \\ 0 & 0 & 0 & \cdots & 0 & -1 & 1 \end{bmatrix}. \tag{6.4}$$

e $p(a_1|\tau^2) \propto$ constante. Obviamente, um resultado análogo vale para o caso de um passeio aleatório de segunda ordem, com a diferença de que a matriz K, agora, é dada por

$$K = \begin{bmatrix} 1 & -2 & 1 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 & 0 & 0 \\ -2 & 5 & -4 & 1 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & -4 & 6 & -4 & 1 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & -4 & 6 & -4 & 1 & 0 & \cdots & 0 & 0 & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots & 1 & -4 & 6 & -4 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots & 0 & 1 & -4 & 6 & -4 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 1 & -4 & 5 & -2 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & 1 & -2 & 1 \end{bmatrix}. \tag{6.5}$$

De modo geral, por se tratarem de expressões quadráticas sobre os coeficientes $a_1, ..., a_M$, teremos sempre que

$$p(\mathbf{a}|\tau^2) \propto \exp\left(-\frac{1}{2\tau^2}\mathbf{a}'K\mathbf{a}\right) \tag{6.6}$$

para alguma matriz simétrica K.

6.1.2 Conexão Entre o Enfoque Bayesiano e o Clássico

Como ϵ_t segue uma distribuição definida pela mistura na escala de distribuições gaussianas, para cada $t = 1, ..., T$, o mesmo ocorre com y_t para cada $t = 1, ..., T$, de modo que

$$p(y_t) = \int_0^\infty \frac{1}{\delta\psi(\sigma_t)}\phi\left(\frac{y - f(x_t)}{\delta\sigma_t}\right) h(\sigma_t)d\sigma_t.$$

Novamente, uma maneira de se modelar os dados nestas circunstâncias é assumir que

$$\begin{aligned} y_t|\sigma_t &\sim \mathcal{N}\left(f(x_t), \delta^2\psi(\sigma_t)^2\right), \\ \sigma_t &\sim h, \end{aligned}$$

de modo que, ao aproximar f via *B-splines*, teremos

$$\begin{aligned} y_t|\sigma_t &\sim \mathcal{N}\left(\sum_{j=1}^{M} a_j B_j(x_t), \delta^2\psi(\sigma_t)^2\right), \\ \sigma_t &\sim h. \end{aligned} \tag{6.7}$$

Usando as equações (6.6) e (6.7), temos que a densidade a posteriori dos coeficientes $a_1, ..., a_M$ satisfaz

$$\begin{aligned} p(\mathbf{a}|\mathbf{y}) &\propto p(\mathbf{y}|\mathbf{a})p(\mathbf{a}) \\ &= \int_0^\infty \cdots \int_0^\infty p(\mathbf{y}|\mathbf{a}, \boldsymbol{\sigma}) h(\boldsymbol{\sigma}) d\boldsymbol{\sigma} \cdot p(\mathbf{a}) \\ &= \left(\int_0^\infty \cdots \int_0^\infty \prod_{t=1}^{T} \frac{1}{\psi(\sigma_t)\delta} \phi \left(\frac{y_t - \sum_{j=1}^{M} a_j B_j(x_t)}{\delta \psi(\sigma_t)} \right) h(\sigma_t) d\sigma_1 \cdots \sigma_T \right) \\ &\quad \cdot \exp\left(-\frac{1}{2\tau^2} \mathbf{a}' K \mathbf{a} \right) \end{aligned}$$

de modo que, pelo teorema de Fubini,

$$p(\mathbf{a}|\mathbf{y}) \propto \left(\prod_{t=1}^{T} \int_0^\infty \phi \left(\frac{y_t - \sum_{j=1}^{M} a_j B_j(x_t)}{\delta \psi(\sigma_t)} \right) h(\sigma_t) d\sigma_t \right) \cdot \exp\left(-\frac{1}{2\tau^2} \mathbf{a}' K \mathbf{a} \right) \tag{6.8}$$

e, conseqüentemente,

$$\log p(\mathbf{a}|\mathbf{y}) \approx \sum_{t=1}^{T} \log \int_0^\infty \phi \left(\frac{y_t - \sum_{j=1}^{M} a_j B_j(x_t)}{\delta \psi(\sigma_t)} \right) h(\sigma_t) d\sigma_t - \frac{1}{2\tau^2} \mathbf{a}' K \mathbf{a}.$$

Portanto, é fácil ver que a moda a posteriori de $a_1, ..., a_M$ coincide com o estimador clássico destes coeficientes para $\lambda \equiv \frac{1}{2\tau^2}$.

6.2 Distribuições Condicionais dos Parâmetros

Em um enfoque bayesiano, devemos estabelecer as prioris dos "parâmetros" e variáveis latentes associados ao modelo. Tais "parâmetros", ou variáveis aleatórias, são: τ^2, $\boldsymbol{\sigma} = (\sigma_1, ..., \sigma_T)$. Nas subseções a seguir obtemos as distribuições condicionais completas a posteriori dos parâmetros com o intuito de aplicar o amostrador de Gibbs.

6.2.1 Distribuição Condicional Completa de a

Para determinar a distribuição de $\mathbf{a}$ condicionada nas observações, $\mathbf{y}$, e nos demais parâmetros, observe que $p(\mathbf{a}|\mathbf{y}, \boldsymbol{\sigma}, \delta^2, \tau^2) \propto p(\mathbf{y}|\mathbf{a}, \boldsymbol{\sigma}, \delta^2) p(\mathbf{a}|\tau^2)$. Logo, como $\mathbf{y}|\mathbf{a}, \boldsymbol{\sigma}, \delta^2 \sim \mathcal{N}(B\mathbf{a}, \delta^2 W)$, onde

$$W = \text{diag}(\psi(\sigma_1)^2, ..., \psi(\sigma_T)^2),$$

seque que

$$\begin{aligned} p(\mathbf{a}|\mathbf{y}, \boldsymbol{\sigma}, \delta^2, \tau^2) &\propto \exp\left\{ -\frac{1}{2\delta^2} (\mathbf{y} - B\mathbf{a})' W^{-1} (\mathbf{y} - B\mathbf{a}) \right\} \cdot \exp\left\{ -\frac{1}{2\tau^2} \mathbf{a}' K \mathbf{a} \right\} \\ &= \exp\left\{ -\frac{1}{2\delta^2} (\mathbf{y} - B\mathbf{a})' W^{-1} (\mathbf{y} - B\mathbf{a}) - \frac{1}{2\tau^2} \mathbf{a}' K \mathbf{a} \right\}. \end{aligned}$$

Definindo $\overline{W} \equiv (\delta^2 W)^{-1}$ e $\overline{K} \equiv \tau^{-2} K$, teremos

$$\begin{aligned} A &= -\frac{1}{2}\left[(\mathbf{y} - B\mathbf{a})'\overline{W}(\mathbf{y} - B\mathbf{a}) + \mathbf{a}'\overline{K}\mathbf{a}\right] \\ &= -\frac{1}{2}\left[\mathbf{y}'\overline{W}\mathbf{y} - \mathbf{a}'B'\overline{W}\mathbf{y} - \mathbf{y}'\overline{W}B\mathbf{a} + \mathbf{a}'B'\overline{W}B\mathbf{a} + \mathbf{a}'\overline{K}\mathbf{a}\right] \\ &\approx -\frac{1}{2}\left[-2\mathbf{a}'B'\overline{W}\mathbf{y} + \mathbf{a}'(B'\overline{W}B + \overline{K})\mathbf{a}\right] \end{aligned}$$

onde '$\approx$' na expressão acima significa igualdade a menos de uma constante. Finalmente, definindo $S^{-1} \equiv B'\overline{W}B + \overline{K}$,

$$A \approx -\frac{1}{2}(\mathbf{a} - SB'\overline{W}\mathbf{y})'S^{-1}(\mathbf{a} - SB'\overline{W}\mathbf{y})$$

de modo que

$$p(\mathbf{a}|\mathbf{y}, \boldsymbol{\sigma}, \delta^2, \tau^2) \propto \exp\left\{-\frac{1}{2}(\mathbf{a} - SB'\overline{W}\mathbf{y})'S^{-1}(\mathbf{a} - SB'\overline{W}\mathbf{y})\right\}.$$

Ou seja, a distribuição a posteriori de $\mathbf{a}$ é dada por

$$\mathbf{a}|\mathbf{y}, \boldsymbol{\sigma}, \delta^2, \tau^2 \sim \mathcal{N}_M(SB'\overline{W}\mathbf{y}, S).$$

6.2.2 Distribuição Condicional Completa de δ^2 e de τ^2

Assuma que, a priori, $\delta^2 \sim \text{GI}(\gamma_0, \alpha_0)$ e $\tau^2 \sim \text{GI}(\gamma_1, \alpha_1)$. Ou seja, assuma que se $\overline{\delta} \equiv 1/\delta^2$ e $\overline{\tau} \equiv 1/\tau^2$, então, $\overline{\delta} \sim \Gamma(\alpha_0, \gamma_0)$ e $\overline{\tau} \sim \Gamma(\alpha_1, \gamma_1)$ e que, portanto, suas densidades a priori são dadas por

$$p(\overline{\delta}|\gamma_0, \alpha_0) = \frac{\overline{\delta}^{\alpha_0 - 1}}{\Gamma(\alpha_0)\gamma_0^{\alpha_0}} \exp\left\{-\frac{\overline{\delta}}{\gamma_0}\right\}$$

e

$$p(\overline{\tau}|\gamma_1, \alpha_1) = \frac{\overline{\tau}^{\alpha_1 - 1}}{\Gamma(\alpha_1)\gamma_1^{\alpha_1}} \exp\left\{-\frac{\overline{\tau}}{\gamma_1}\right\}.$$

Para calcular as distribuições condicionais completas, é conveniente fazer a mudança de variáveis $\rho_i = \frac{1}{\gamma_i}$, para $i = 0, 1$. Feito isto, as densidades acima assumem as formas abaixo,

$$p(\overline{\delta}|\rho_0, \alpha_0) \propto \overline{\delta}^{\alpha_0 - 1} \exp\{-\rho_0\overline{\delta}\}$$

e

$$p(\overline{\tau}|\rho_1, \alpha_1) \propto \overline{\tau}^{\alpha_1 - 1} \exp\{-\rho_1\overline{\tau}\}.$$

Para determinar a distribuição de $\delta|\mathbf{y}, \boldsymbol{\sigma}, \mathbf{a}, \tau^2$, note que

$$\begin{aligned} p(\overline{\delta}|\mathbf{y}, \boldsymbol{\sigma}, \mathbf{a}, \tau) &\propto p(\mathbf{y}|\overline{\delta}, \boldsymbol{\sigma}, \mathbf{a}, \tau^2)p(\overline{\delta}) \\ &\propto p(\mathbf{y}|\overline{\delta}, \boldsymbol{\sigma}, \mathbf{a})p(\overline{\delta}) \\ &\propto \overline{\delta}^{T/2} \exp\left\{-\frac{\overline{\delta}}{2}(\mathbf{y} - B\mathbf{a})'W^{-1}(\mathbf{y} - B\mathbf{a})\right\} \overline{\delta}^{\alpha_0 - 1} \exp\{-\rho_0\overline{\delta}\} \end{aligned}$$

de modo que

$$p(\overline{\delta}|\mathbf{y},\boldsymbol{\sigma},\mathbf{a},\tau^2) \propto \overline{\delta}^{\alpha_0'-1}\exp\{-\rho_0'\overline{\delta}\}$$

onde $\alpha_0' = \alpha_0 + \frac{T}{2}$ e $\rho_0' = \rho_0 + \frac{(\mathbf{y}-B\mathbf{a})'W^{-1}(\mathbf{y}-B\mathbf{a})}{2}$. Em outras palavras,

$$\delta^2|\mathbf{y},\boldsymbol{\sigma},\mathbf{a},\tau \sim \text{GI}(\alpha_0',\gamma_0'),$$

onde $\gamma_0' = 1/\rho_0'$.

Analogamente,

$$\begin{aligned} p(\overline{\tau}|\mathbf{y},\boldsymbol{\sigma},\mathbf{a},\delta) &\propto p(\mathbf{a}|\overline{\tau})p(\overline{\tau}) \\ &\propto \overline{\tau}^{M/2}\exp\left\{-\frac{\overline{\tau}}{2}\mathbf{a}'K\mathbf{a}\right\}\overline{\tau}^{\alpha_1-1}\exp\{-\rho_1\overline{\tau}\}. \end{aligned}$$

Logo,

$$p(\overline{\tau}|\mathbf{y},\boldsymbol{\sigma},\mathbf{a},\delta) \propto \overline{\tau}^{\alpha_1'-1}\exp\{-\rho_1'\overline{\tau}\}$$

onde $\alpha_1' = \alpha_1 + \frac{M}{2}$ e $\rho_1' = \rho_1 + \frac{\mathbf{a}'K\mathbf{a}}{2}$. Ou seja,

$$\tau|\mathbf{y},\boldsymbol{\sigma},\mathbf{a},\delta \sim \text{GI}(\alpha_1',\gamma_1'),$$

onde $\gamma_1' = 1/\rho_1'$.

6.2.3 Distribuição Condicional Completa de $\boldsymbol{\sigma}$

Ao contrário dos demais parâmetros, a distribuição condicional completa de $\boldsymbol{\sigma}$ depende da especificação de h e, portanto, não pode ser determinada de modo genérico. Além disso, apenas em alguns casos particulares, poderemos expressá-la em uma forma analiticamente fechada. De qualquer modo, assumindo como priori para σ_i, $i = 1,...,T$,

$$p(\sigma_i) = h(\sigma_i)$$

teremos

$$\begin{aligned} p(\boldsymbol{\sigma}|\mathbf{y},\mathbf{a},\delta,\tau) &\propto p(\mathbf{y}|\boldsymbol{\sigma},\mathbf{a},\delta)p(\boldsymbol{\sigma}) \\ &= \prod_{t=1}^{T} p(y_i|\sigma_t,\mathbf{a},\delta)p(\sigma_t) \\ &\propto \prod_{t=1}^{T}\frac{1}{\psi(\sigma_t)}\exp\left\{-\frac{r_t^2}{2\delta^2\psi(\sigma_t)^2}\right\}h(\sigma_t), \end{aligned}$$

onde $r_t = y_t - \sum_{j=1}^{M} a_j B_j(x_j)$ é o resíduo para a t-ésima observação. Note também que a densidade condicional completa de $\boldsymbol{\sigma}$ independe de τ.

No caso particular em que $\psi(\sigma_t) = 1/\sqrt{\sigma_t}$, temos que

$$\begin{aligned} p(\boldsymbol{\sigma}|\mathbf{y}, \mathbf{a}, \delta) &\propto \prod_{t=1}^{T} \sigma_t^{1/2} \exp\left\{-\frac{r_t^2}{2\delta^2}\sigma_t\right\} h(\sigma_t) \\ &= \prod_{t=1}^{T} p_g\left(\sigma_t \middle| \frac{2\delta^2}{r_t^2}, \frac{3}{2}\right) h(\sigma_t), \end{aligned} \tag{6.9}$$

onde p_g é a função densidade de probabilidade associada à distribuição gama. Mais ainda, se assumirmos que o ruído segue uma distribuição t de Student com ν graus de liberdade, a densidade h será dada por

$$h(s) = \frac{\left(\frac{\nu}{2}\right)^{\frac{\nu}{2}}}{\Gamma\left(\frac{\nu}{2}\right)} s^{\frac{\nu}{2}-1} e^{-\frac{\nu}{2}s}.$$

Ou, em outras palavras, h será a densidade associada à distribuição gama com parâmetros $\alpha = \nu/2$ e $\gamma = 2/\nu$ e isto resultará no fato que

$$\boldsymbol{\sigma}|\mathbf{y}, \mathbf{a}, \delta \sim \Gamma\left(\frac{\nu+1}{2}, 2\left(\frac{r_t^2}{\delta^2} + \nu\right)^{-1}\right).$$

6.2.4 Amostrador de Gibbs

Segue imediatamente dos resultados obtidos anteriormente que o algoritmo de Gibbs para os parâmetros $\mathbf{a}$, δ, τ e $\boldsymbol{\sigma}$ é dado por

Conforme observamos acima, a densidade condicional completa a posteriori de $\boldsymbol{\sigma}$ nem sempre é conhecida. Logo, em alguns casos pode ser necessário usar o algoritmo mais genérico de Metropolis-Hastings para amostrar a partir de (6.10). No entanto, no caso particular em que o ruído segue uma distribuição t de Student com ν graus de liberdade, teremos

$$\boldsymbol{\sigma}^{(k+1)} \sim \Gamma\left(\frac{\nu+1}{2}, 2(\frac{(r_t^{(k+1)})^2}{(\delta^2)^{(k+1)}} + \nu)^{-1}\right).$$

Sobre a Amostragem de $\mathbf{a}^{(k)}$

De acordo com o algoritmo de Gibbs derivado acima, a distribuição condicional completa de $\mathbf{a}^{(k+1)}$, para cada k, é uma gaussiana multivariada com média $S^{(k)}B'\overline{W}^{(k)}\mathbf{y}$ e matriz de covariâncias $S^{(k)}$. Então, para amostrar $\mathbf{a}^{(k+1)}$, recorremos ao seguinte procedimento:

1. amostrar $\tilde{a}_i^{(k+1)} \sim \mathcal{N}(0,1)$, para $i = 1, ..., n$. Note que $\tilde{\mathbf{a}}^{(k+1)} \sim \mathcal{N}_M(\mathbf{0}, I_M)$, onde I_M é a matriz identidade n-dimensional;
2. obter a matriz triangular inferior A tal que $AA' = S^{(k)}$ de acordo com a decomposição de Choleski;
3. obter $\mathbf{a}^{(k)}$ de acordo com
$$\mathbf{a}^{(k)} = A\tilde{\mathbf{a}}^{(k+1)} + S^{(k)}B'\overline{W}^{(k)}\mathbf{y}.$$

Algoritmo 6.2.1 Amostrador de Gibbs

• **No passo** $k+1$:

1. $\mathbf{a}^{(k+1)} \sim \mathcal{N}_M(S^{(k)}B'\overline{W}^{(k)}\mathbf{y}, S^{(k)})$, onde

- $\overline{W}^{(k)} = (\delta^2)^{(k)}\text{diag}(\psi(\sigma_1^{(k)})^2, ..., \psi(\sigma_T^{(k)})^2)$,
- $\overline{K}^{(k)} = \frac{1}{(\tau^2)^{(k)}}K$,
- $S^{(k)} = (B'\overline{W}^{(k)}B + \overline{K}^{(k)})^{-1}$;

2. $(\delta^2)^{(k+1)} \sim \text{GI}(\alpha_0', \gamma_0')$, onde

- $\alpha_0' = \alpha_0 + \frac{T}{2}$,
- $\gamma_0' = \frac{1}{\rho_0'}$ e $\rho_0' = \rho_0 + \frac{(\mathbf{y}-B\mathbf{a}^{(k+1)})'W^{-1}(\mathbf{y}-B\mathbf{a}^{(k+1)})}{2}$;

3. $(\tau^2)^{(k+1)} \sim \text{GI}(\alpha_1', \gamma_1')$, onde

- $\alpha_1' = \alpha_1 + \frac{M}{2}$,
- $\gamma_1' = \frac{1}{\rho_1'}$ e $\rho_1' = \rho_1 + \frac{(\mathbf{a}^{(k+1)})'K\mathbf{a}^{(k+1)}}{2}$;

4. $\boldsymbol{\sigma}^{(k+1)} \sim \prod_{t=1}^{T} p_g\left(\sigma_t \middle| \frac{2(\delta^{(k+1)})^2}{(r_t^{(k+1)})^2}, \frac{3}{2}\right) h(\sigma_t)$ onde

- $r_t^{(k+1)} = y_t - \sum_{j=1}^{M} a_j^{(k+1)} B_j(x_t)$. (6.10)

Sobre o uso do Algoritmo de Metropolis-Hastings

O passo 4 do algoritmo 6.2.1 consiste na simulação de $\boldsymbol{\sigma}^{(k+1)}$ de acordo com a densidade conjunta

$$\prod_{t=1}^{T} p_g\left(\sigma_t \middle| \frac{2(\delta^{(k+1)})^2}{(r_t^{(k+1)})^2}, \frac{3}{2}\right) h(\sigma_t).$$

Note que, dada a forma da densidade conjunta, os elementos de são amostrados independentemente, cada um deles de acordo com a densidade

$$p_g\left(\sigma_t \middle| \frac{2(\delta^{(k+1)})^2}{(r_t^{(k+1)})^2}, \frac{3}{2}\right) h(\sigma_t).$$

a qual, como observado acima, dependendo da forma de h pode não ser representável como de forma análitica como a densidade de uma distribuição conhecida. Logo, para amostrar estas variáveis, necessitamos usar o algoritmo de Metropolis-Hastings.

Observe que a densidade h é, por hipótese, conhecida e que ela tem o mesmo suporte que a densidade de $\sigma_t^{(k+1)}$, logo, podemos tomá-la como densidade proposta. Em particular, a densidade proposta sugerida independe do estado anterior, ie, de $\sigma_t^{(k)}$. Deste modo, no $(k+1)$-ésimo passo do algoritmo, se amostrarmos y de acordo com h, então, a probabilidade de aceitação é dada por

$$\alpha = \min\left\{1, \frac{p_g\left(y \middle| \frac{2(\delta^{(k+1)})^2}{(r_t^{(k+1)})^2}, \frac{3}{2}\right)}{p_g\left(\sigma_t^{(k)} \middle| \frac{2(\delta^{(k+1)})^2}{(r_t^{(k+1)})^2}, \frac{3}{2}\right)}\right\}, \tag{6.11}$$

de modo que o passo 4 do algoritmo 6.2.1 fica como representado no algoritmo 6.2.2.

Algoritmo 6.2.2 Amostragem via Metropolis-Hastings dentro do algoritmo de Gibbs

---

• **No passo** $k+1$:

4.1. Amostrar $s \sim h$;

4.2. Amostrar $u \sim \mathcal{U}[0,1]$;

4.3. Tomar α de acordo com (6.11);

4.4. Se $\alpha < u$,

$$\sigma_t^{(k+1)} = \sigma_t^{(k)},$$

senão

$$\sigma_t^{(k+1)} = s.$$

Uma alternativa ao algoritmo de Metropolis-Hastings para gerar $\sigma_t^{(k+1)}$ é o algoritmo 6.2.3, de aceitação/rejeição. A constante $M_t^{(k+1)}$ no algoritmo deve ser escolhida de modo que

$$p_g\left(\sigma_t \middle| \frac{2(\delta^{(k+1)})^2}{(r_t^{(k+1)})^2}, \frac{3}{2}\right) h(\sigma_t) \leq M_t^{(k+1)} h(\sigma_t),$$

Algoritmo 6.2.3 Amostragem pelo método de aceitação-rejeição dentro do algoritmo de Gibbs

• **No passo** $k+1$:

4.1. Amostrar $s \sim h$;

4.2. Amostrar $u \sim \mathcal{U}[0,1]$;

4.3. Aceitar $\sigma_t^{(k+1)} = s$ se

$$u \leq \frac{p_g\left(s \middle| \frac{2(\delta^{(k+1)})^2}{(r_t^{(k+1)})^2}\right)}{M_t^{(k+1)}}$$

senão retornar ao passo 4.1.;

isto é, tal que

$$p_g\left(\sigma_t \middle| \frac{2(\delta^{(k+1)})^2}{(r_t^{(k+1)})^2}, \frac{3}{2}\right) \leq M_t^{(k+1)}.$$

É fácil ver que $p_g(s|\alpha,\beta)$ atinge seu máximo em $s = \beta(\alpha-1)$ de modo que podemos tomar

$$M_t^{(k+1)} = p_g\left(\frac{3(\delta^{(k+1)})^2}{(r_t^{(k+1)})^2} - \frac{3}{2} \middle| \frac{2(\delta^{(k+1)})^2}{(r_t^{(k+1)})^2}, \frac{3}{2}\right).$$

Os algoritmos 6.2.2 e 6.2.3 são meios automáticos de se amostrar $\sigma_t^{(k+1)}$ para qualquer densidade de mistura h, inclusive para distribuições estáveis. No entanto, ele pode não ser muito eficiente se a densidade h estiver concentrada sobre uma das caudas de p_g como mostra a figura 6.1. Note que, em quaisquer uma das situações ilustradas, o valor amostrado de acordo com h será quase sempre rejeitado. Portanto, devemos considerar a possibilidade de construir distribuições propostas de acordo com o caso, ie, de acordo com a densidade h escolhida.

6.2.5 Simulações

Estudo 1

Neste primeiro estudo de simulaçã consideramos a função alvo

$$f(x) = \text{sen}(2x)$$

definida no intervalo $[0,1]$, veja figura 5.1. Neste primeiro estudo de simulação, consideramos um ruído distribuido de acordo com uma distribuição t de Student com 3 graus de liberdade e parâmetro de escala $\delta = 1$ e geramos artificialmente uma amostra de tamanho 500. Aproximamos f via P-splines e, para tanto, usamos uma base de B-*splines* de grau 3 e 6 graus de liberdade e variamos a matriz de penalização K, de modo que analisamos os resultados tanto para o caso em que os coeficientes $\mathbf{a}$ são modelados de acordo com um passeio aleatório de primeira ordem quanto para o caso em os mesmos são modelados de acordo com um passeio aleatório de segunda ordem. Para obter as estimativas de acordo com as distribuições a posteriori dos parâmetros do modelo, geramos 5000 amostras de cada parâmetro de acordo com o algoritmo de Gibbs

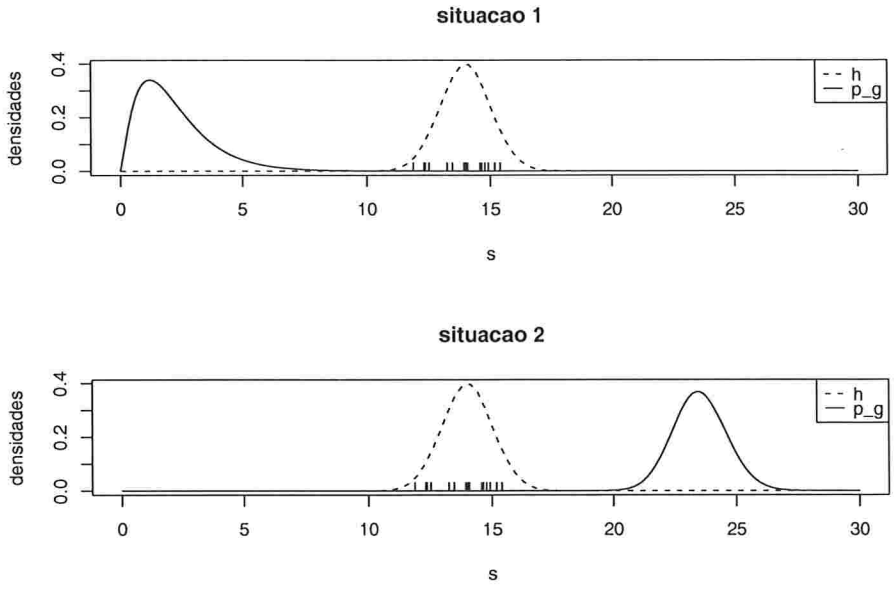


Figura 6.1: Ilustração sobre a amostragem da posteriori através da densidade a priori h. Note que dependendo da posição relativa de h, as propostas geradas serão invariavelmente rejeitadas.

descrito acima. A função alvo f foi, então, estimada por $\widehat{f} = \sum_{j=1}^{20} \widehat{a}_j B_j$, onde $\widehat{a}_j$ é a média da amostra obtida através do algoritmo de Gibbs para a_j. O resultado desta operação pode ser visualizada na figura 6.2. Analogamente, a estimativa obtida para o parâmetro de escala[1], δ, foi $\widehat{\delta} = 0,9618$ e para o parâmetro de suavização[2], τ, foi $\widehat{\tau} = 0,5826$. A figura 6.3 ilustra a convergência da amostra gerada para δ e τ. É interessante notar que as amostras obtidas para o parâmetro de escala convergem rapidamente (na média) para o valor real $\delta = 1$. Quanto ao parâmetro de suavização, embora ele não tenha sido fixado a priori, a convergência na média é rápida, mas não tão rápida quanto a ocorrida para o parâmetro de suavização. Além disso, pode-se observar que a variância das amostras obtidas neste caso é significativamente maior que a variância obtida para o parâmetro de escala, veja tabela 6.1.

Tabela 6.1: Estatísticas para as amostras obtidas via algoritmo de Gibbs para os parâmetros de escala e de suavização assumindo um passeio aleatório de ordem 1.

Parâmetro	Média	Mediana	Variância
δ	0,9618	0,9596	0.0024
τ	0,5826	0,4532	0,2732

[1] média da raiz quadrada das amostras obtidas para δ através do algoritmo de Gibbs

[2] média da raiz quadrada das amostras obtidas para τ através do algoritmo de Gibbs

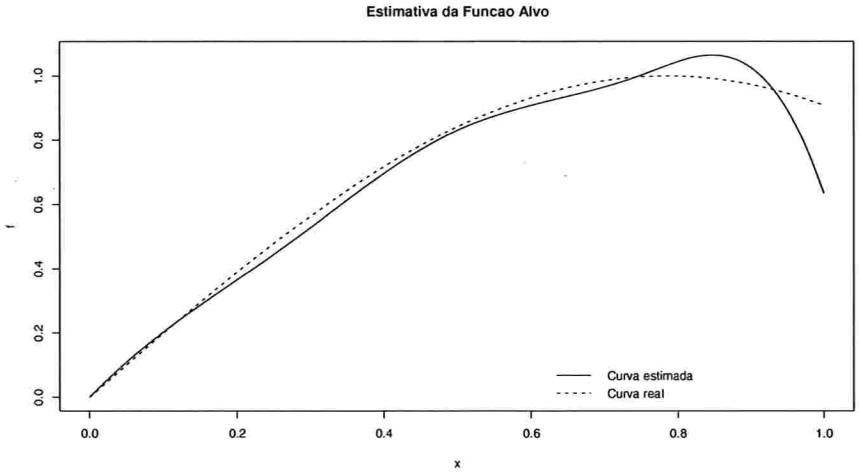


Figura 6.2: Estimativa da função alvo f para o primeiro estudo de simulação quando a matriz de penalização K é dada por (6.4), isto é, assumindo um passeio aleatório de primeira ordem para os coeficientes $\mathbf{a}$.

A mesma análise (assumindo os mesmos parâmetros da análise anterior) acima foi feita assumindo um passeio aleatório de ordem 2 para os coeficientes $\mathbf{a}$. A estimativa obtida para a função alvo pode ser vista na figura 6.4. Note que, embora o ajuste para os demais parâmetros não mude muito entre as análises, a curva obtida assumindo-se um passeio aleatório de ordem 2, como era esperado, é mais suave e se ajusta melhor que a curva obtida assumindo-se um passeio aleatŕio simples. Analogamente, a estimativa obtida para o parâmetro de escala, δ, foi $\widehat{\delta} = 0,9621$ e para o parâmetro de suavização, τ, foi $\widehat{\tau} = 0,5474$. A figura 6.5 ilustra a convergência da amostra gerada para δ e τ. Note que, assim como no caso anterior, as amostras obtidas para o parâmetro de escala convergem rapidamente (na média) para o valor real $\delta = 1$. No entanto, o parâmetro de suavição, parece convergir mais lentamente que no caso do passeio aleatório simples. Ainda assim, podemos dizer que a convergência na média é rápida, mas, novamente, não tõ rápida quanto a ocorrida para o parâmetro de suavização. Repete-se também o fenômeno que a variância das amostras obtidas neste caso é significativamente maior que a variância obtida para o parâmetro de escala, veja tabela 6.2. Em geral,

Tabela 6.2: Estatísticas para as amostras obtidas via algoritmo de Gibbs para os parâmetros de escala e de suavização assumindo um passeio aleatório de ordem 2 no estudo de simulação #1.

Parâmetro	**Média**	**Mediana**	**Variância**
δ	0,9621	0,9598	0.0024
τ	0,5474	0,4155	0,2378

as estatísticas usando passeios aleatórios de ordem 1 e 2, não variam muito de um caso para o outro, exceto por uma ligeira redução na variância da amostra obtida para o parâmetro de suavização quando usamos um passeio aleatório de ordem 2.

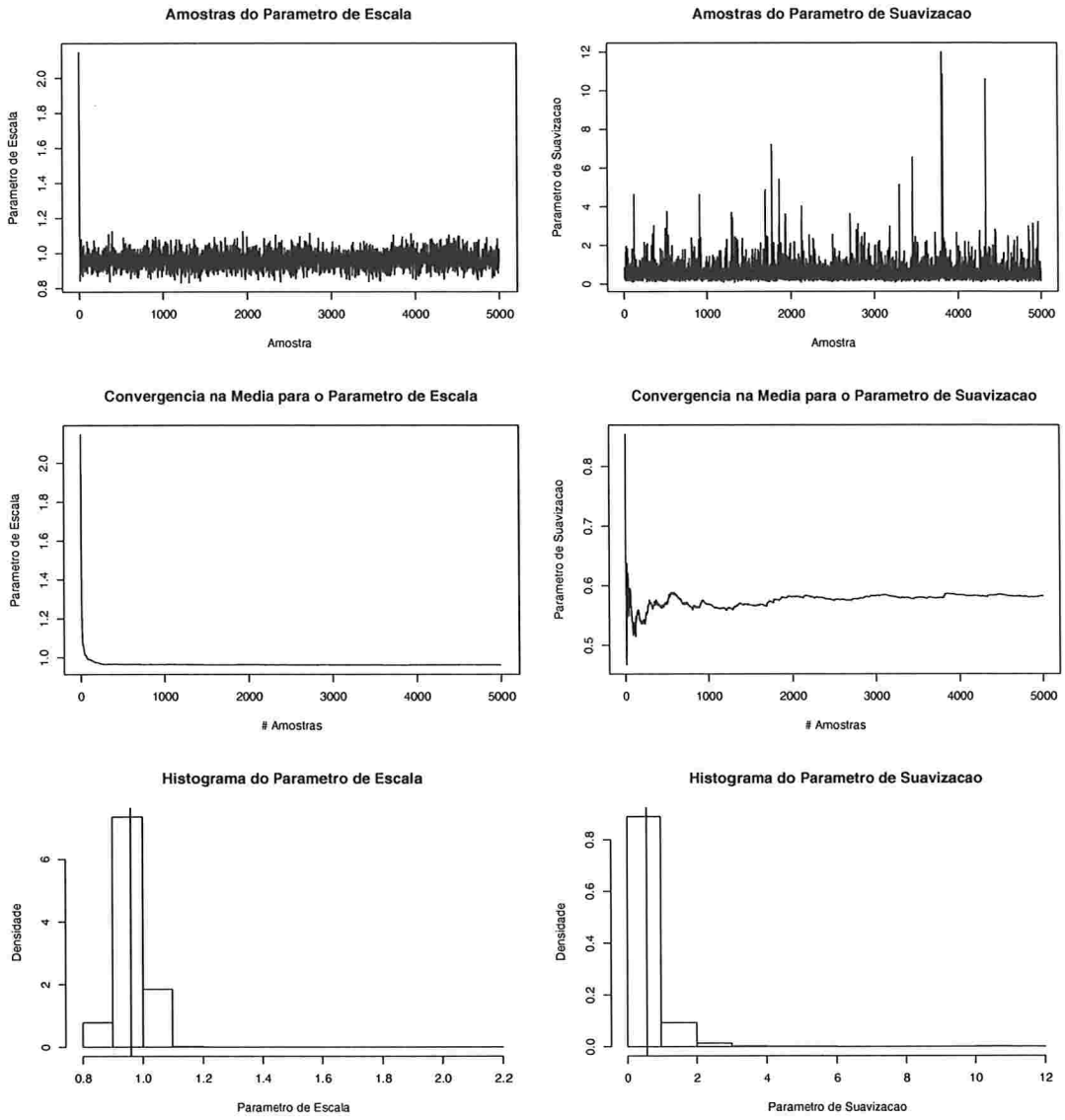


Figura 6.3: Convergência da amostragem para os parâmetros de escala e de suavização para o primeiro estudo de simulação assumindo que os coeficientes a seguem um passeio aleatório de primeira ordem. Os gráficos da esquerda correspondem ao para parâmetro de escala e os da direita ao parâmetro de suavização. Os dois gráficos superiores correspondem à amostragem destes parâmetros através do algoritmo de Gibbs, enquanto que os gráficos intermediários ilustram a convergência na média destas amostragens. Os histogramas na parte inferior da figura ilustram a dispersão da amostragem obtida e as barras verticais neles o valor médio de cada amostra.

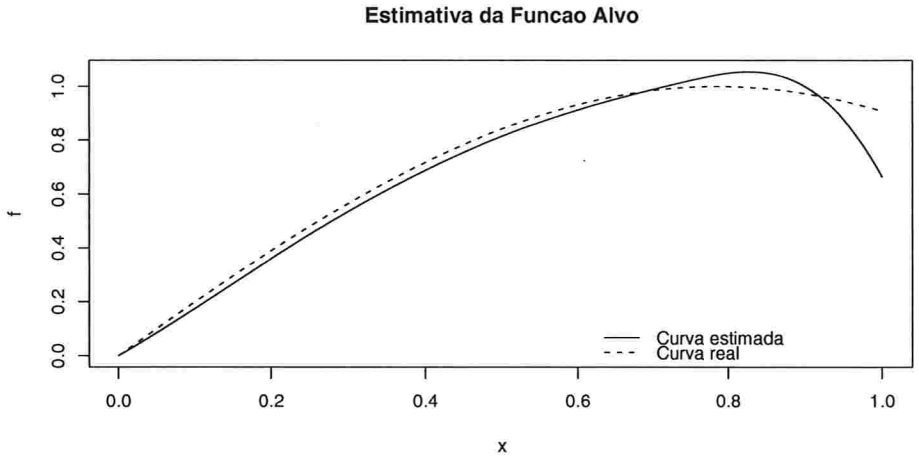


Figura 6.4: Estimativa da função alvo f para o primeiro estudo de simulação quando a matriz de penalização K é dada por (6.5), isto é, assumindo um passeio aleatório de segunda ordem para os coeficientes $\mathbf{a}$.

Estudo 2

Neste segundo de simulação seguimos o mesmo padrão do estudo anterior, mas assumimos agora que o ruído segue uma distribuição de Cauchy e que o parâmetro de escala é dado por $\delta = 1,5$. Assim como no estudo 1, realizamos dois ajustes, o primeiro assumindo que os coeficientes $\mathbf{a}$ seguem um passeio aleatório simples e o segundo assumindo que os mesmos coeficientes seguem um passeio aleatório de segunda ordem. Dado que a função alvo usada neste estudo apresenta uma variabilidade maior que no primeiro caso, tomamos uma base B-spline de ordem 3 com 16 graus de liberdade. A figura 6.6 mostra o quão bem a curva estimada se ajusta à função alvo apesar da presença de diversos valores extremos, como pode-se ver no gráfico de dispersão. Com relação aos demais parâmetros do modelo, temos que a estimativa obtida para o parâmetro de escala é igual a $1,4096$, enquanto que para o parâmetro de suavização, τ, é igual a $16,5208$. A tabela 6.3 apresenta mais algumas estatísticas associadas a estes parâmetros. Já a figura 6.7 mostra como as amostras de Gibbs

Tabela 6.3: Estatísticas para as amostras obtidas via algoritmo de Gibbs para os parâmetros de escala e de suavização assumindo um passeio aleatório simples no estudo de simulação #2.

Parâmetro	Média	Mediana	Variância
δ	1,4096	1,4016	0,0754
τ	16,5208	14,8974	61,4668

para ambos os parâmetros convergem rapidamente. Assim como em todos os casos anteriores, a variância da amostra obtida para o parâmetro de suavização é muito superior àquela obtida para o parâmetro de escala.

Assumindo os mesmos parâmetros do estudo de simulação anterior, exceto pelo fato de que agora modelamos, a priori, os coeficientes $\mathbf{a}$ de acordo com um passeio aleatório de ordem 2, realizamos novas simulações

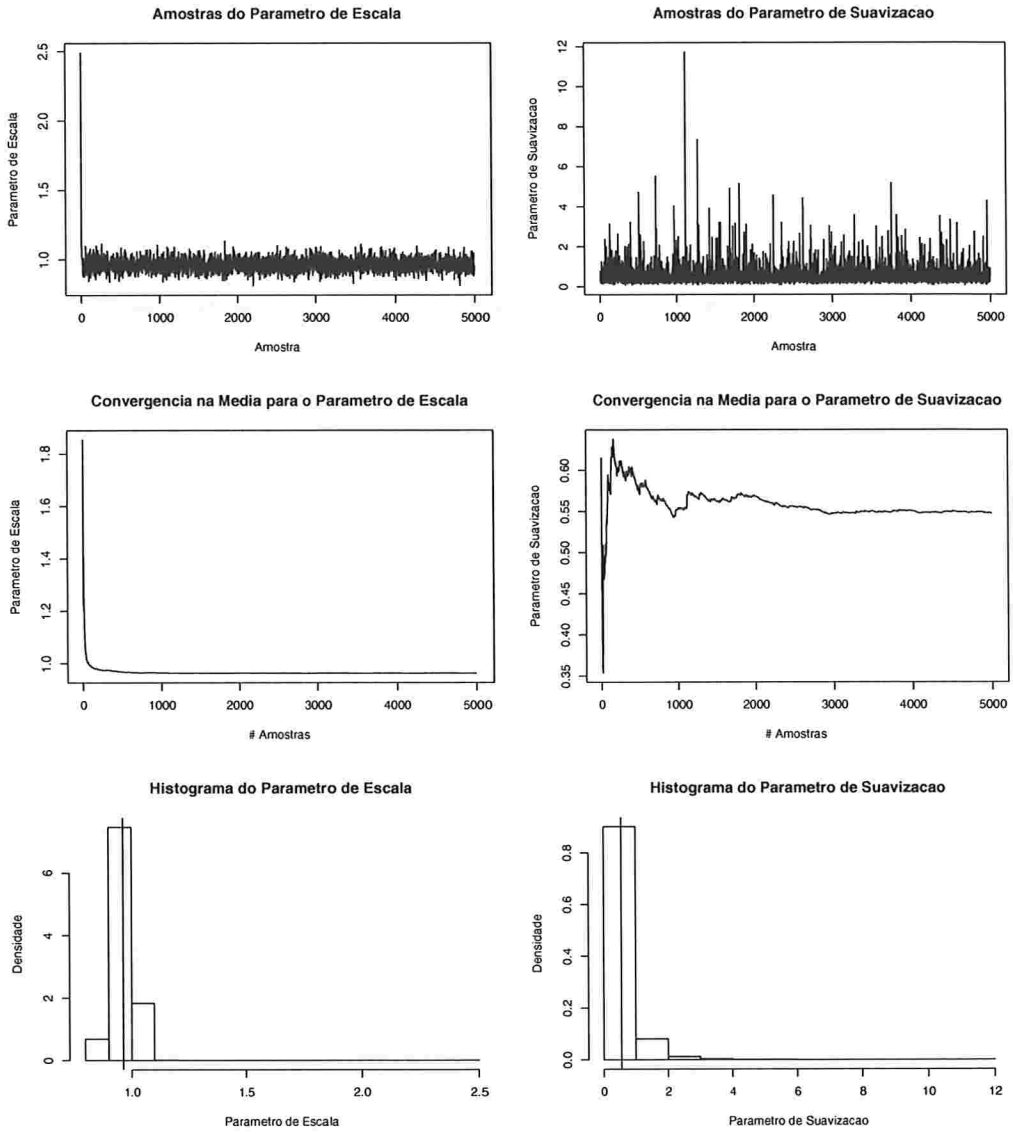


Figura 6.5: Convergência da amostragem para os parâmetros de escala e de suavização para o primeiro estudo de simulação assumindo que os coeficientes **a** seguem um passeio aleatório de segunda ordem. Os gráficos da esquerda correspondem ao para parâmetro de escala e os da direita ao parâmetro de suavização. Os dois gráficos superiores correspondem à amostragem destes parâmetros através do algoritmo de Gibbs, enquanto que os gráficos intermediários ilustram a convergência na média destas amostragens. Os histogramas na parte inferior da figura ilustram a dispersão da amostragem obtida e as barras verticais neles o valor médio de cada amostra.

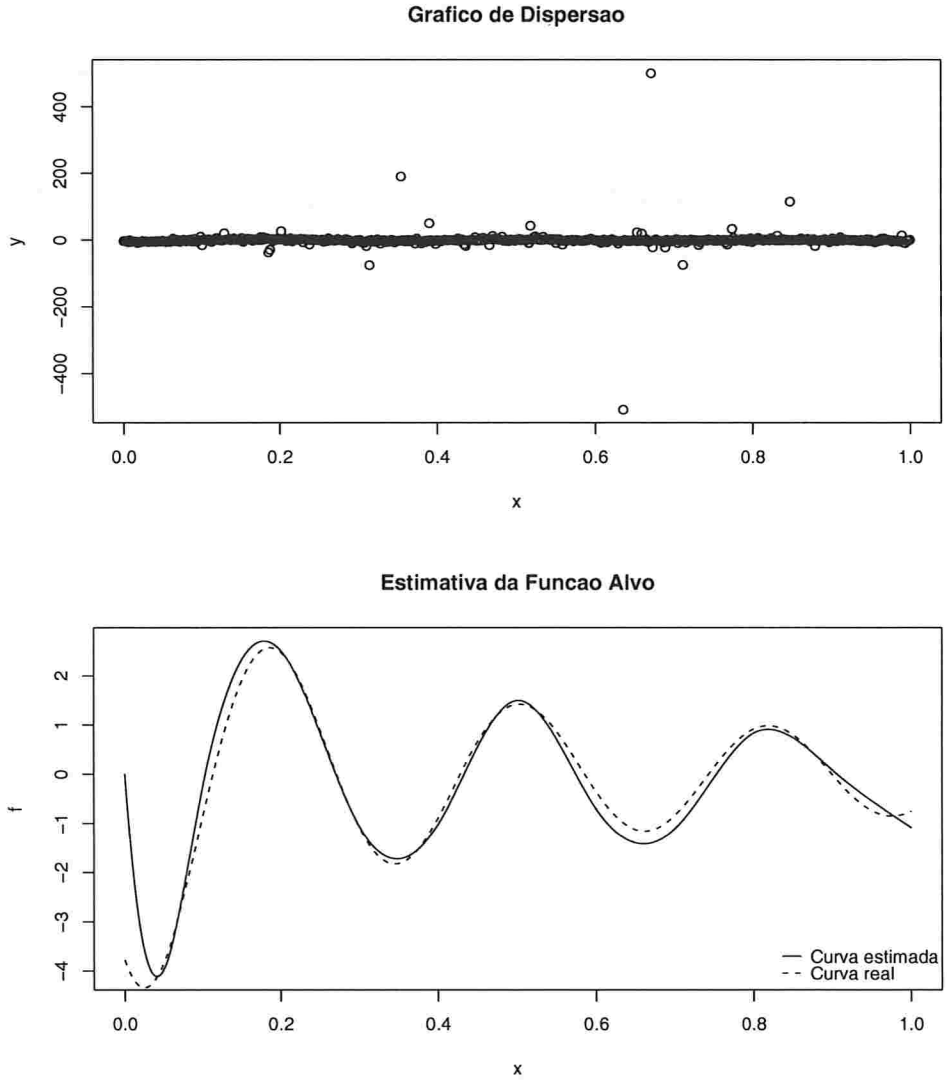


Figura 6.6: **Gráfico superior:** gráfico de dispersão; **Gráfico inferior:** estimativa da função alvo f para o segundo estudo de simulação quando a matriz de penalização K é dada por (6.4), isto é, assumindo um passeio aleatório de primeira ordem para os coeficientes $\mathbf{a}$.

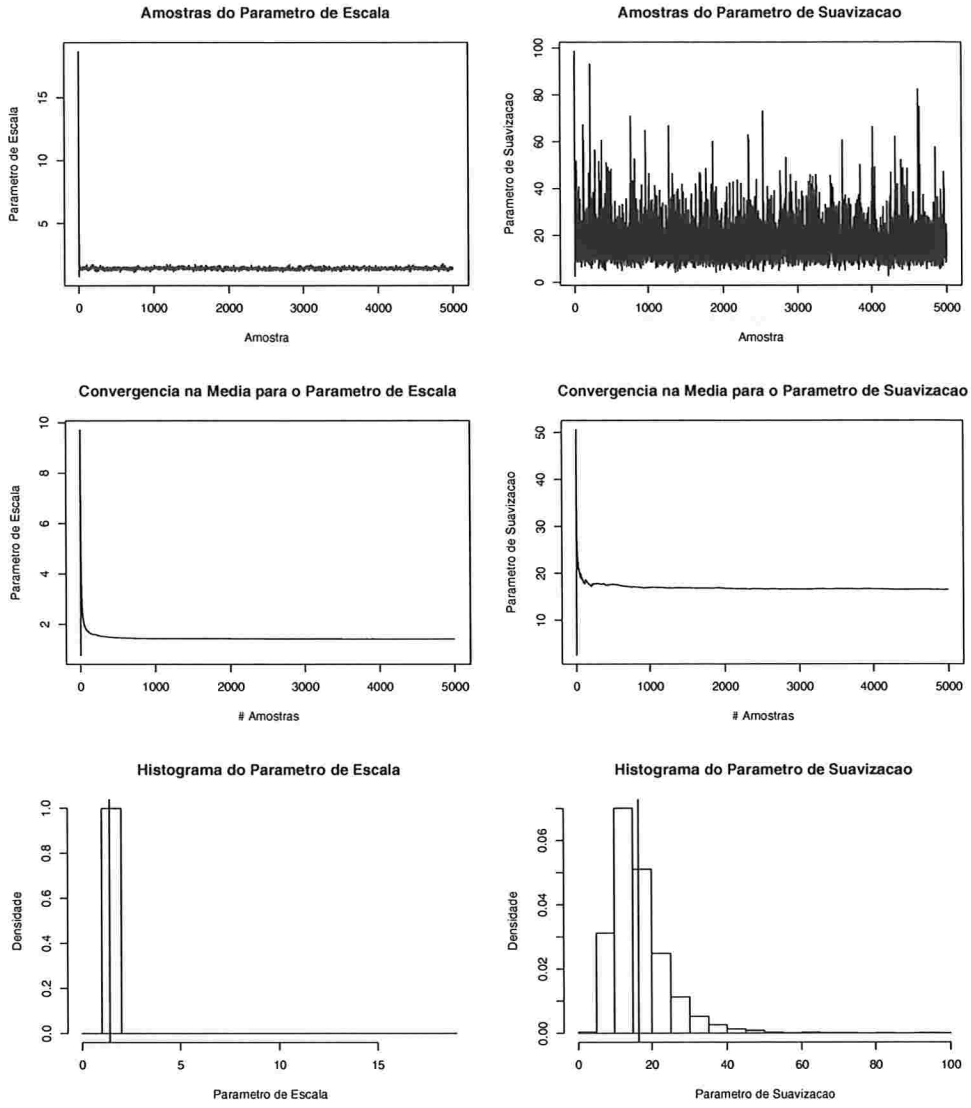


Figura 6.7: Convergência da amostragem para os parâmetros de escala e de suavização para o segundo estudo de simulação assumindo que os coeficientes a seguem um passeio aleatório de primeira ordem. Os gráficos da esquerda correspondem ao para parâmetro de escala e os da direita ao parâmetro de suavização. Os dois gráficos superiores correspondem à amostragem destes parâmetros através do algoritmo de Gibbs, enquanto que os gráficos intermediários ilustram a convergência na média destas amostragens. Os histogramas na parte inferior da figura ilustram a dispersão da amostragem obtida e as barras verticais neles o valor médio de cada amostra.

e obtivemos novas estimativas a posteriori para a função alvo e para os parâmetros de escala e de suavização. A figura 6.8 mostra o quão bem a curva estimada se ajusta à função alvo apesar da presença de diversos valores extremos, como pode-se ver no gráfico de dispersão. Com relação aos demais parâmetros do modelo,

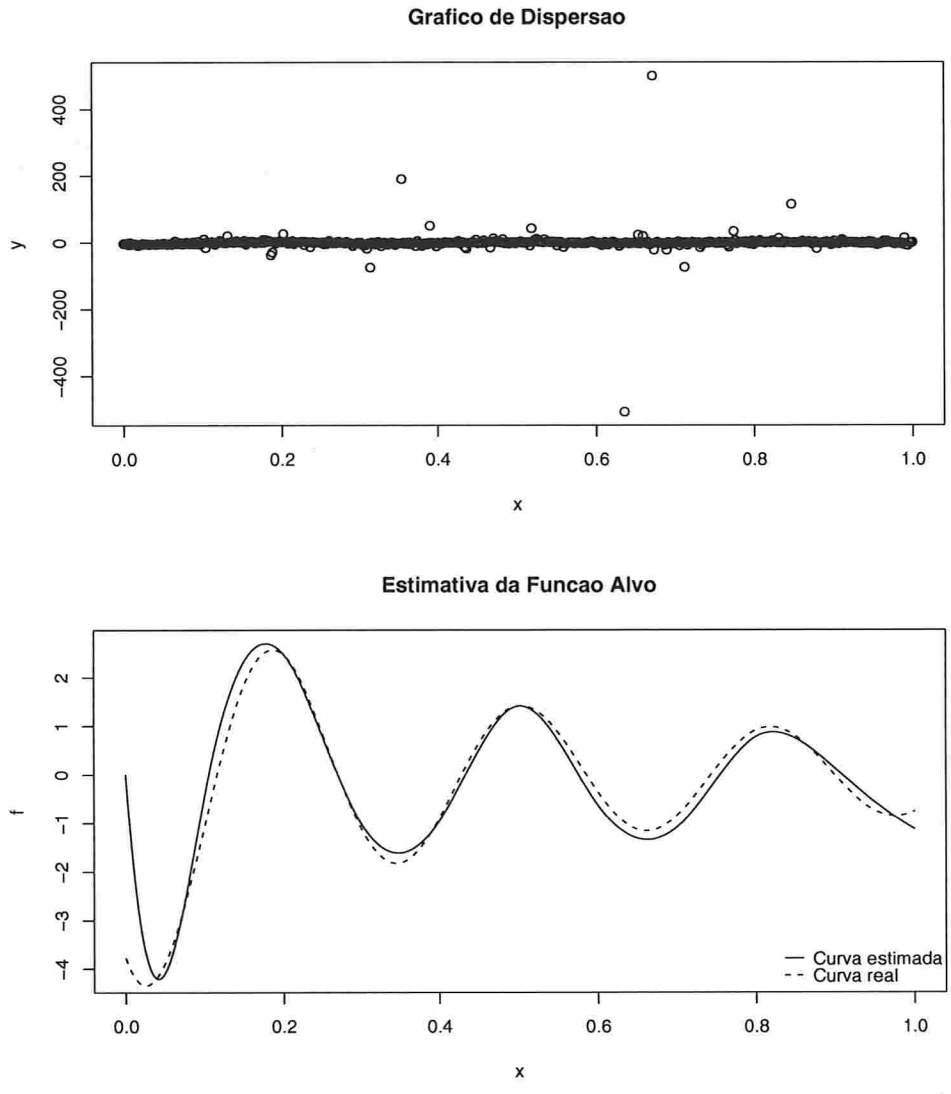


Figura 6.8: **Gráfico superior:** gráfico de dispersão; **Gráfico inferior:** estimativa da função alvo f para o segundo estudo de simulação quando a matriz de penalização K é dada por (6.5), isto é, assumindo um passeio aleatório de segunda ordem para os coeficientes $\mathbf{a}$.

temos que a estimativa obtida para o parâmetro de escala é igual a $1,4123$, enquanto que para o parâmetro de suavização, τ, é igual a $28,2078$. A tabela 6.4 apresenta mais algumas estatísticas associadas a estes

parâmetros. Já a figura 6.9 mostra como as amostras de Gibbs para ambos os parâmetros convergem ra-

Tabela 6.4: Estatísticas para as amostras obtidas via algoritmo de Gibbs para os parâmetros de escala e de suavização assumindo um passeio aleatório de ordem 2 no estudo de simulação #2.

Parâmetro	Média	Mediana	Variância
δ	1,4123	1,4018	0,1601
τ	28,2078	24,9987	187,7998

pidamente. Assim como em todos os casos anteriores, a variância da amostra obtida para o parâmetro de suavização é muito superior àquela obtida para o parâmetro de escala.

6.3 Extensão para Modelos Parcialmente Lineares

Nesta seção aplicaremos o enfoque bayesiano discutido acima no caso em que temos um modelo parcialmente linear

$$y_t = f(x_t) + \boldsymbol{\beta}' \mathbf{u}_t + \delta \epsilon_t, \tag{6.12}$$

onde $\mathbf{u}_t$ é um vetor de dimensão L formado pelas variáveis explicativas (componente linear do modelo) e, como antes, ϵ_t segue uma mistura na escala de gaussianas para todo t. Como veremos abaixo, tal enfoque facilmente ser estendido para esta classe mais de modelos.

Assim como nas seções anteriores, vamos aproximar f por

$$f(x) = \sum_{j=1}^{M} a_j B_j(x)$$

onde $(B_1, ..., B_M)$ é uma base de funções, possivelmente B-splines. Neste caso, o modelo (6.12) pode ser escrito na forma matricial

$$\mathbf{y} = B\mathbf{a} + U\boldsymbol{\beta} + \delta\boldsymbol{\epsilon}$$

onde $U = [\mathbf{u}_1, ..., \mathbf{u}_L]'$ e B é dado por (6.3). Além disso, manteremos as hipóteses sobre o vetor de coeficientes $\mathbf{a}$, ie, de que suas componentes seguem um passeio aleatório e que, portanto, satisfazem, a priori, (6.6) para alguma matriz simétrica K. Deste modo, o vetor com a totalidade de parâmetros e hiper-parâmetros é dado por $(\mathbf{a}', \boldsymbol{\beta}', \delta, \tau, \boldsymbol{\sigma}')'$. Novamente, o cálculo explicito das distribuições e estatísticas a posteriori não é uma tarefa fácil sob as condições dadas e, portanto, usaremos o algoritmo de Gibbs para obter amostras de acordo com as distribuições a posteriori.

6.3.1 Distribuição Condicional Completa de a

Podemos aproveitar os cálculos na seção 6.2.1 para obter a distribuição condicional completa de $\mathbf{a}$ simplesmente substituindo $\mathbf{y}$ nos cálculos daquela seção por $\mathbf{z}_a = \mathbf{y} - U\boldsymbol{\beta}$ de modo que

$$\mathbf{a}|\mathbf{y}; \boldsymbol{\beta}, \delta^2, \tau^2, \boldsymbol{\sigma} \sim \mathcal{N}_M(SB'\overline{W}\mathbf{z}_a, S)$$

onde $S^{-1} = B'\overline{W}B + \overline{K}$, $\overline{W} = \frac{1}{\delta^2}W^{-1}$ e $\overline{K} = \frac{1}{\tau^2}K$.

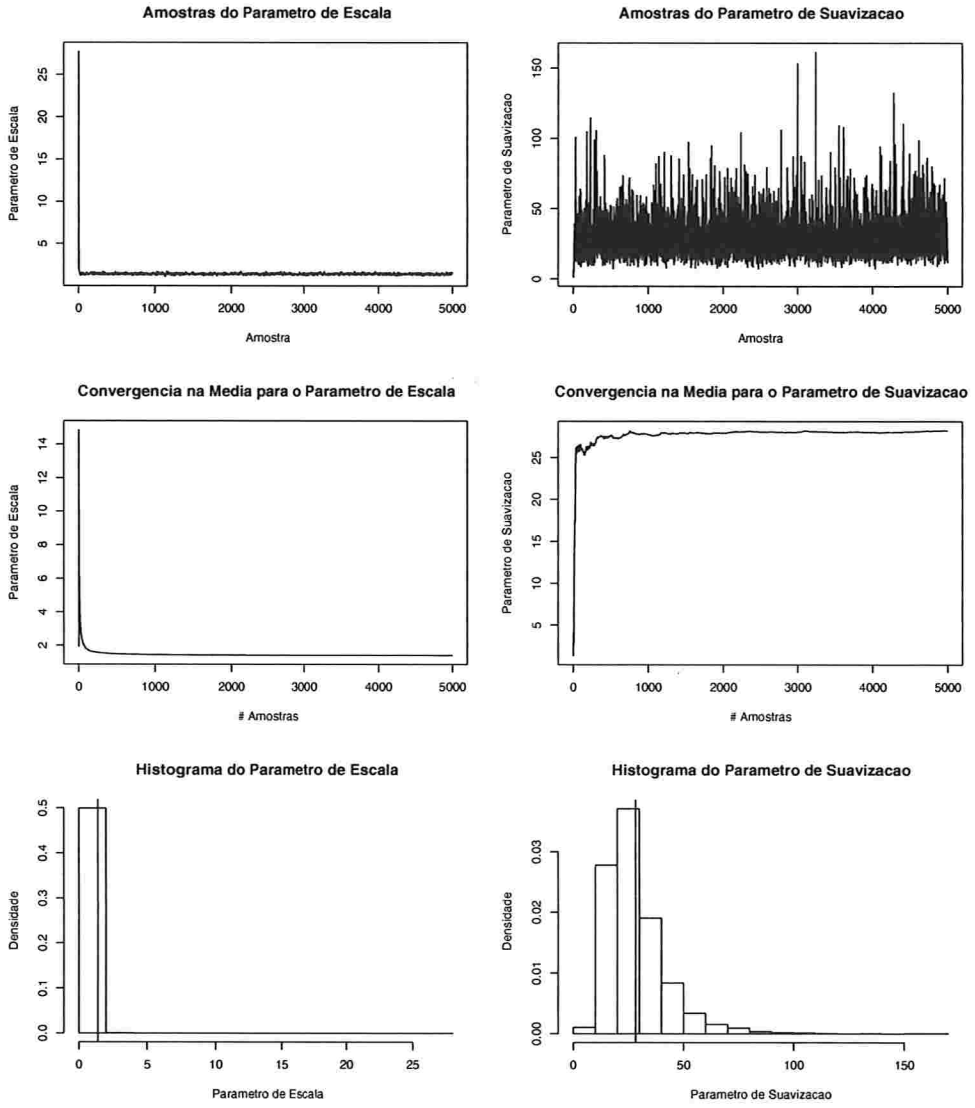


Figura 6.9: Convergência da amostragem para os parâmetros de escala e de suavização para o segundo estudo de simulação assumindo que os coeficientes a seguem um passeio aleatório de segunda ordem. Os gráficos da esquerda correspondem ao para parâmetro de escala e os da direita ao parâmetro de suavização. Os dois gráficos superiores correspondem à amostragem destes parâmetros através do algoritmo de Gibbs, enquanto que os gráficos intermediários ilustram a convergência na média destas amostragens. Os histogramas na parte inferior da figura ilustram a dispersão da amostragem obtida e as barras verticais neles o valor médio de cada amostra.

6.3.2 Distribuição Condicional Completa de δ^2 e de τ^2

O cálculo das distribuições condicionais completas de δ^2 e de τ^2 no caso de modelos parcialmente lineares é análogo aos do caso univariado, seção 6.2.2. Na verdade, no caso de τ^2, é idêntico e a única alteração que devemos considerar é na atualização do hiper-parâmetro ρ_0 associado à δ^2, de modo que, agora,

$$\rho_0' = \rho_0 + \frac{(\mathbf{y} - B\mathbf{a} - U\boldsymbol{\beta})' W^{-1} (\mathbf{y} - B\mathbf{a} - U\boldsymbol{\beta})}{2}. \tag{6.13}$$

Logo,

$$\delta^2 | \mathbf{y}; \mathbf{a}, \boldsymbol{\beta}, \tau^2, \boldsymbol{\sigma} \sim \text{GI}(\alpha_0', \gamma_0'),$$

onde $\gamma_0' = 1/\rho_0'$, e

$$\tau | \mathbf{y}; \mathbf{a}, \boldsymbol{\beta}, \delta^2, \boldsymbol{\sigma} \sim \text{GI}(\alpha_1', \gamma_1'),$$

onde $\gamma_1' = 1/\rho_1'$ e $\rho_1' = \rho_1 + \frac{\mathbf{a}' K \mathbf{a}}{2}$.

6.3.3 Distribuição Condicional Completa de $\boldsymbol{\sigma}$

Usando exatamente a mesma argumentação que na seção 6.2.3, teremos que a densidade condicional completa de $\boldsymbol{\sigma}$ é dada por (6.9), ie,

$$p(\boldsymbol{\sigma} | \mathbf{y}, \mathbf{a}, \delta) \propto \prod_{t=1}^{T} p_g \left(\sigma_t \middle| \frac{2\delta^2}{r_t^2}, \frac{3}{2} \right) h(\sigma_t)$$

porém com $r_t = y_t - \sum_{j=1}^{M} a_j B_(x_t) - \boldsymbol{\beta}' \mathbf{u}_t$.

6.3.4 Distribuição Condicional Completa de $\boldsymbol{\beta}$

Assuma que $\boldsymbol{\beta} \sim \mathcal{N}(\mathbf{m}_\beta, H_\beta)$. Neste caso,

$$\begin{aligned} p(\boldsymbol{\beta} | \mathbf{y}; \mathbf{a}, \boldsymbol{\sigma}, \delta^2, \tau^2) &\propto p(\mathbf{y} | \mathbf{a}, \boldsymbol{\beta}, \boldsymbol{\sigma}, \delta^2) p(\boldsymbol{\beta} | \mathbf{m}_\beta, H_\beta) \\ &\propto \exp\left\{ -\frac{1}{2\delta^2} (\mathbf{y} - B\mathbf{a} - U\boldsymbol{\beta})' W^{-1} (\mathbf{y} - B\mathbf{a} - U\boldsymbol{\beta}) \right\} \\ &\quad \times \exp\left\{ -\frac{1}{2} (\boldsymbol{\beta} - m_\beta)' H_\beta^{-1} (\boldsymbol{\beta} - m_\beta) \right\} \end{aligned}$$

e definindo $\mathbf{z}_\beta = \mathbf{y} - B\mathbf{a}$, temos

$$\begin{aligned} p(\boldsymbol{\beta} | \mathbf{y}; \mathbf{a}, \boldsymbol{\sigma}, \delta^2, \tau^2) &\propto \exp\left\{ -\frac{1}{2\delta^2} (\mathbf{z}_\beta - U\boldsymbol{\beta})' W^{-1} (\mathbf{z}_\beta - U\boldsymbol{\beta}) \right\} \\ &\quad \times \exp\left\{ -\frac{1}{2} (\boldsymbol{\beta} - m_\beta)' H_\beta^{-1} (\boldsymbol{\beta} - m_\beta) \right\}. \end{aligned}$$

Agora, definindo $A = (\mathbf{z}_\beta - U\boldsymbol{\beta})' \overline{W} (\mathbf{z}_\beta - U\boldsymbol{\beta}) + (\boldsymbol{\beta} - m_\beta)' H_\beta^{-1} (\boldsymbol{\beta} - m_\beta)$, pode-se mostrar que

$$A \approx (\boldsymbol{\beta} - C_\beta^{-1} (H_\beta^{-1} \mathbf{m}_\beta + U' \overline{W} \mathbf{z}_\beta))' C_\beta (\boldsymbol{\beta} - C_\beta^{-1} (H_\beta^{-1} \mathbf{m}_\beta + U' \overline{W} \mathbf{z}_\beta))$$

onde $\approx$ significa igualdade a menos de uma constante e C_β é uma matriz de covariâncias dada por

$$C_\beta = H_\beta^{-1} + U'\overline{W}U.$$

Logo, escrevendo $\mu_\beta = C_\beta^{-1}(H_\beta^{-1}\mathbf{m}_\beta - U'\overline{W}\mathbf{z}_\beta)$, temos que

$$\beta|\mathbf{y};\mathbf{a},\sigma,\delta^2,\tau^2 \sim \mathcal{N}_L(\mu_\beta, C_\beta).$$

6.3.5 Amostrador de Gibbs para o Modelo Parcialmente Linear

Reunindo os resultados obtidos nas seções anteriores, obtemos o algoritmo de Gibbs para o problema que estamos analisando, algoritmo 6.3.1.

6.3.6 Simulação

Para ilustrar a metodologia acima consideramos novamente a função alvo

$$f(x) = \frac{\text{sen}(20(x+0.2))}{x+0.2}$$

definida no intervalo $[0,1]$, veja figura 5.1 para a componente não linear, enquanto que para a componente linear assumimos o vetor de coeficientes

$$\beta = (3,45; -2; -4,56)'.$$

Quanto ao ruído e ao parâmetro de escala, δ, supomos o primeiro distribuido de acordo com uma distribuição de Cauchy e o segundo igual a $1,5$. Finalmente, geramos artificialmente uma amostra de tamanho 500. A função alvo f foi aproximada via P-splines e, para tanto, usamos uma base de B-*splines* de grau 3 e 12 graus de liberdade e variamos a matriz de penalização K, de modo que analisamos os resultados tanto para o caso em que os coeficientes $\mathbf{a}$ fossem modelados de acordo com um passeio aleatório de primeira ordem quanto para o caso em os mesmos fossem modelados de acordo com um passeio aleatório de segunda ordem. Para obter as estimativas de acordo com as distribuições a posteriori dos parâmetros do modelo, geramos 10000 amostras de cada parâmetro de acordo com o algoritmo de Gibbs descrito acima. A função alvo f foi, então, estimada por $\widehat{f} = \sum_{j=1}^{12} \widehat{a}_j B_j$, onde $\widehat{a}_j$ é a média da amostra obtida através do algoritmo de Gibbs para a_j. O resultado desta operação pode ser visualizada na figura 6.10 (figura inferior). Analogamente, a estimativa obtida para o parâmetro de escala[3], δ, foi $\widehat{\delta} = 1,5554$ e para o parâmetro de suavização[4], τ, foi $\widehat{\tau} = 3,9114$. A figura 6.11 ilustra a convergência da amostra gerada para δ e τ. É interessante notar que, assim como no caso univariado, as amostras obtidas para o parâmetro de escala convergem rapidamente (na média) para o valor real $\delta = 1,5$. Quanto ao parâmetro de suavição, embora ele não tenha sido fixado a priori, a convergência na média é novamente rápida, mas, diferentemente do caso univariado, não apresenta uma oscilação muito grande. No entanto, ainda pode-se observar que a variância das amostras obtidas neste caso é significativamente maior que a variância obtida para o parâmetro de escala, veja tabela 6.5. Resta avaliar como se comporta a amostra obtida para os coeficientes das componentes linear e não-linear do modelo, β e $\mathbf{a}$. A figura 6.12 ilustra a dispersão e a convergência na média da amostra associada a β. Como podemos notar a convergência é muito rápida e as amostras obtidas permanecem concentradas ao redor do valor real dos parâmetros. Já os boxplots

[3] média da raiz quadrada das amostras obtidas para δ através do algoritmo de Gibbs

[4] média da raiz quadrada das amostras obtidas para τ através do algoritmo de Gibbs

Algoritmo 6.3.1 Amostrador de Gibbs — Modelo Parcialmente Linear

• **No passo** $k+1$:

1. $\mathbf{a}^{(k+1)} \sim \mathcal{N}_M(S^{(k)}B'\overline{W}^{(k)}(\mathbf{y}-U\boldsymbol{\beta}^{(k)}), S^{(k)})$, onde

- $\overline{W}^{(k)} = \frac{1}{(\delta^2)^{(k)}}\text{diag}(\psi(\sigma_1^{(k)})^2, ..., \psi(\sigma_T^{(k)})^2)$,
- $\overline{K}^{(k)} = \frac{1}{(\tau^2)^{(k)}}K$,
- $S^{(k)} = (B'\overline{W}^{(k)}B + \overline{K}^{(k)})^{-1}$;

2. $\boldsymbol{\beta}^{(k+1)} \sim \mathcal{N}_L(\boldsymbol{\mu}_\beta^{(k)}, C_\beta^{(k)})$, onde

- $\boldsymbol{\mu}_\beta^{(k)} = (C_\beta^{(k)})^{-1}(H_\beta^{-1}\mathbf{m}_\beta - U'\overline{W}^{(k)}(\mathbf{y}-B\mathbf{a}^{(k+1)}))$;
- $C_\beta^{(k)} = H_\beta^{-1} + U'\overline{W}^{(k)}U$

3. $(\delta^2)^{(k+1)} \sim \text{GI}(\alpha_0', \gamma_0')$, onde

- $\alpha_0' = \alpha_0 + \frac{T}{2}$,
- $\gamma_0' = \frac{1}{\rho_0'}$ e $\rho_0' = \rho_0 + \frac{(\mathbf{y}-B\mathbf{a}^{(k+1)}-U\boldsymbol{\beta}^{(k+1)})'W^{-1}(\mathbf{y}-B\mathbf{a}^{(k+1)}-U\boldsymbol{\beta}^{(k+1)})}{2}$;

4. $(\tau^2)^{(k+1)} \sim \text{GI}(\alpha_1', \gamma_1')$, onde

- $\alpha_1' = \alpha_1 + \frac{M}{2}$,
- $\gamma_1' = \frac{1}{\rho_1'}$ e $\rho_1' = \rho_1 + \frac{(\mathbf{a}^{(k+1)})'K\mathbf{a}^{(k+1)}}{2}$;

5. $\boldsymbol{\sigma}^{(k+1)} \sim \prod_{t=1}^{T} p_g\left(\sigma_t \middle| \frac{2(\delta^{(k+1)})^2}{(r_t^{(k+1)})^2}, \frac{3}{2}\right) h(\sigma_t)$ onde

- $r_t^{(k+1)} = y_t - \sum_{j=1}^{M} a_j^{(k+1)}B_j(x_t) - \sum_{j=1}^{L}\beta_{t,j}^{(k+1)}u_{t,j}$.

Tabela 6.5: Estatísticas para as amostras obtidas via algoritmo de Gibbs para os parâmetros de escala e de suavização assumindo um passeio aleatório de ordem 1.

Parâmetro	Média	Mediana	Variância
δ	1,5554	1,5483	0.0268
τ	3,9114	3,7698	0,8928

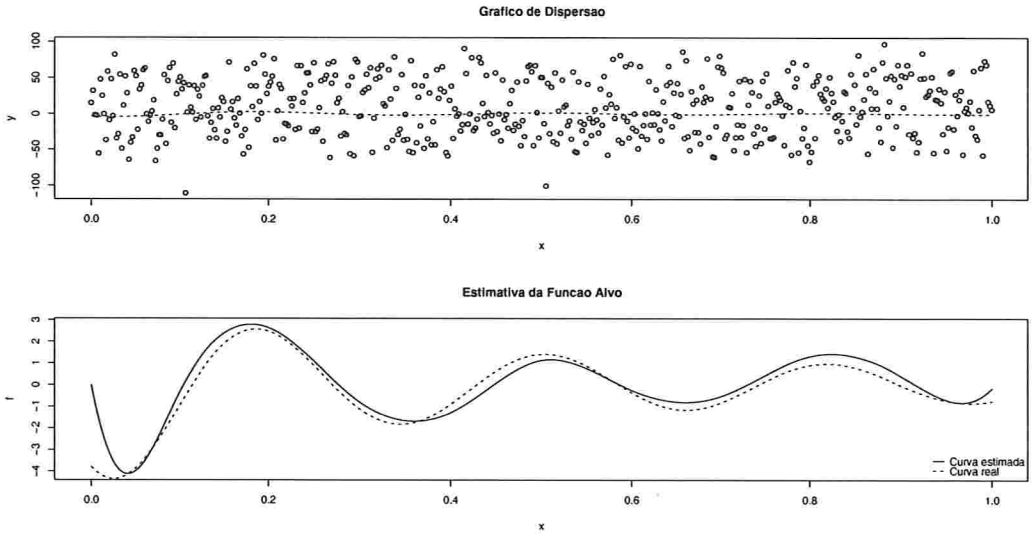


Figura 6.10: Estimativa da função alvo f para o modelo parcialmente linear considerado no estudo de simulação quando a matriz de penalização K é dada por (6.4), isto é, assumindo um passeio aleatório de primeira ordem para os coeficientes $\mathbf{a}$.

6.13 e 6.14 reforçam esta característica em β e indica que o mesmo fenômeno ocorre com $\mathbf{a}$, como pode-se ver pela pequena proximidade entre os quartis inferiores e superiores.

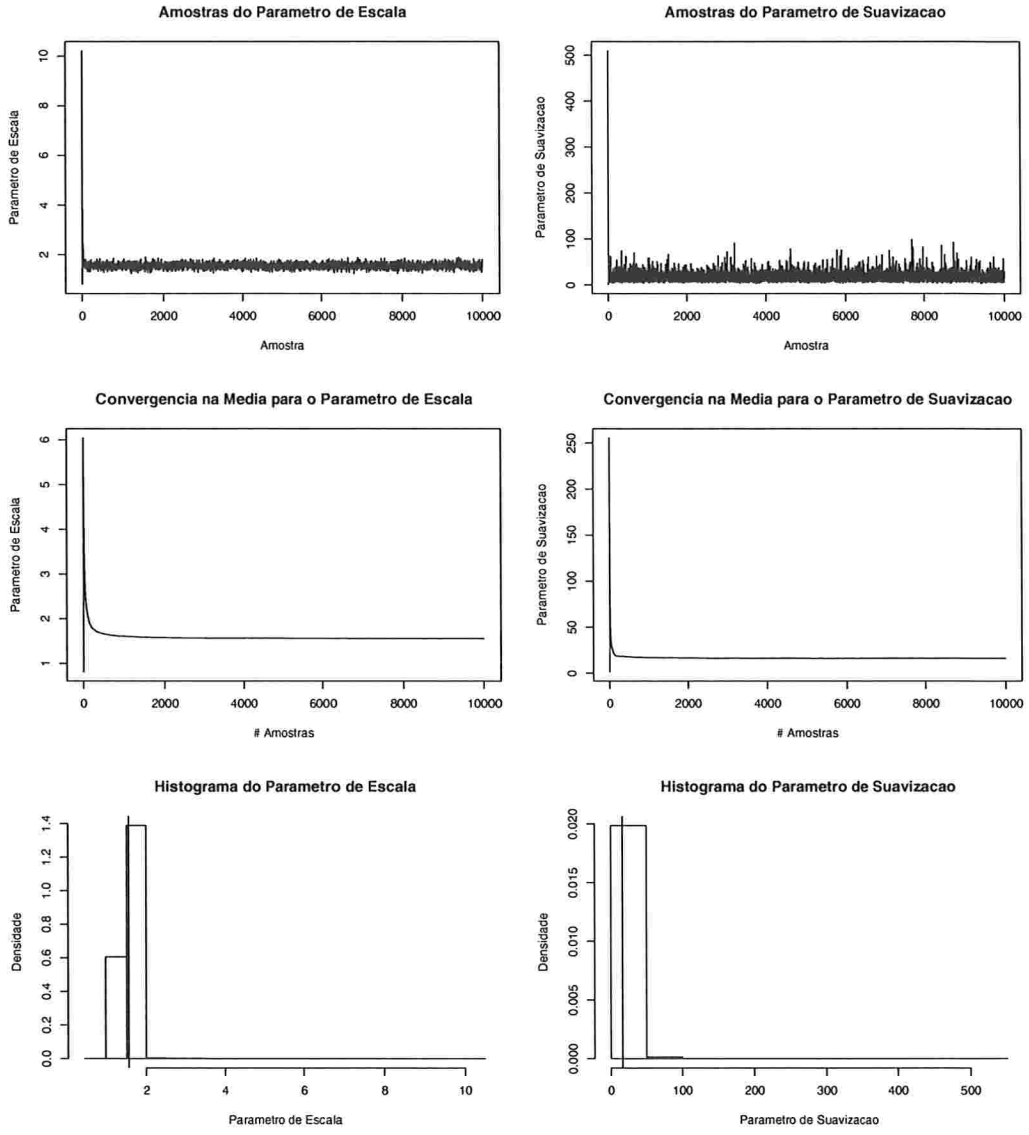


Figura 6.11: Convergência da amostragem para os parâmetros de escala e de suavização para o modelo parcialmente linear considerado no estudo de simulação assumindo que os coeficientes a seguem um passeio aleatório de primeira ordem. Os gráficos da esquerda correspondem ao para parâmetro de escala e os da direita ao parâmetro de suavização. Os dois gráficos superiores correspondem à amostragem destes parâmetros através do algoritmo de Gibbs, enquanto que os gráficos intermediários ilustram a convergência na média destas amostragens. Os histogramas na parte inferior da figura ilustram a dispersão da amostragem obtida e as barras verticais neles o valor médio de cada amostra.

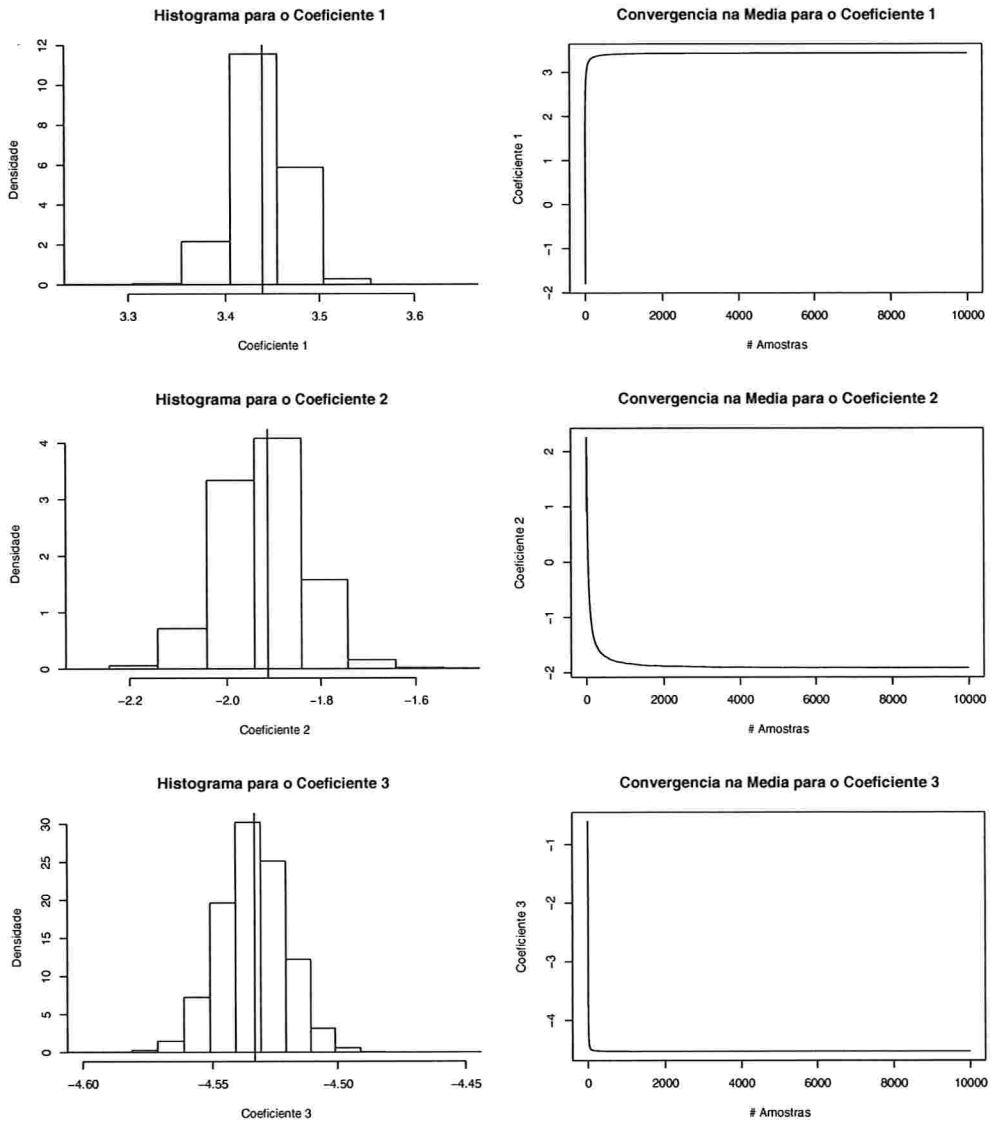


Figura 6.12: Distribuição e convergência da amostragem para os coeficientes da componente linear do modelo parcialmente linear considerado no estudo de simulação assumindo que os coeficientes **a** seguem um passeio aleatório de primeira ordem. A esquerda temos os histogramas associados a estas amostragens e a direita os gráficos indicando a convergência na média dos valores amostrados. A primeira linha corresponde a β_1, a segunda a β_2 e a terceira a β_3.

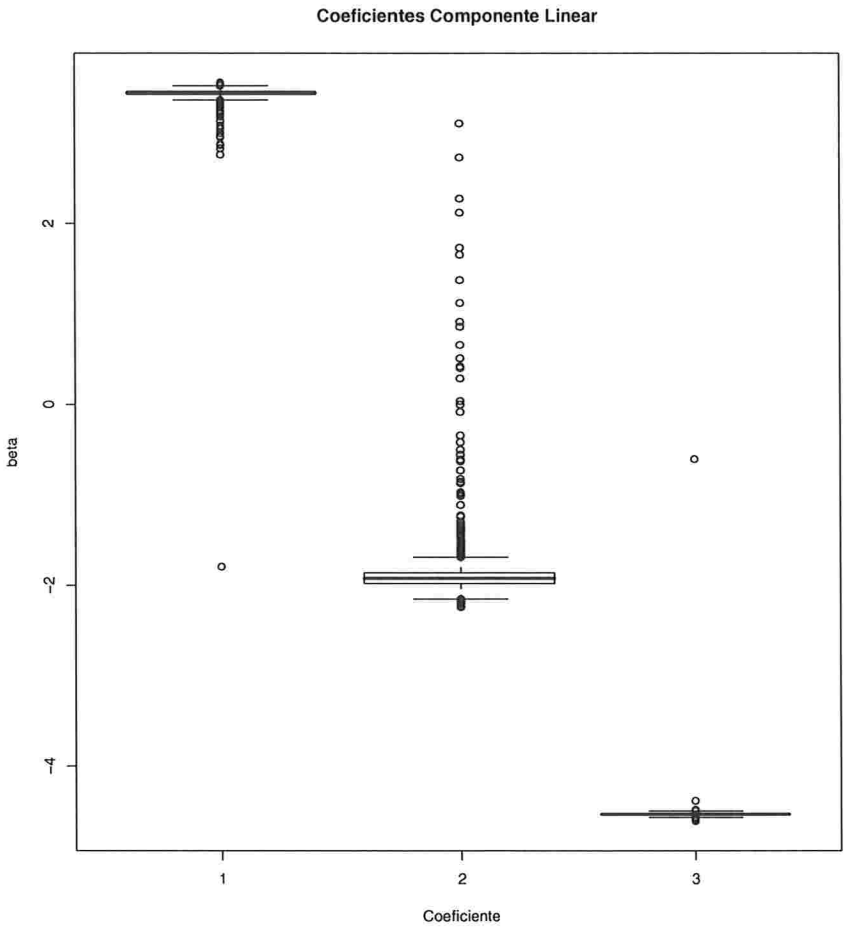


Figura 6.13: Boxplot associado às amostras obtidas via algoritmo de Gibbs para as componentes do vetor de coeficientes β associado à componente linear do modelo.

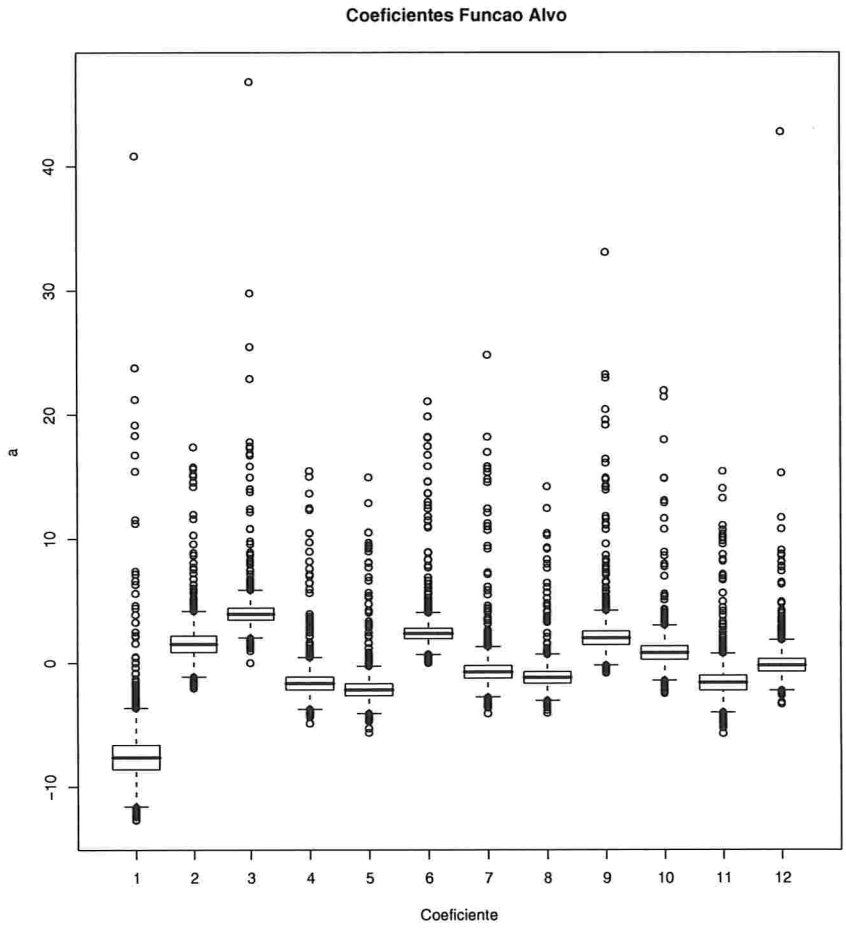


Figura 6.14: Boxplot associado às amostras obtidas via algoritmo de Gibbs para as componentes do vetor de coeficientes a associado à componente não-linear do modelo.

Capítulo 7

Uma Nota Sobre o Uso de Modelos Mistos

7.1 Introdução

Modelos mistos são uma extensão dos modelos de regressão ordinários, os quais têm se mostrado úteis para lidar com medidas repetidas, correlação espacial e não-linearidade através de modelos baseados em *splines*. A forma usual do modelo linear misto com planejamento geral é

$$\mathbf{y} = \mathbf{Xc} + \mathbf{Zu} + \boldsymbol{\epsilon},$$

onde $\mathbf{u}$ e $\boldsymbol{\epsilon}$ são variáveis aleatórias não correlacionadas entre si, com média igual a zero e matrizes de covariância $\mathbf{G}$ e $\mathbf{R}$, respectivamente.

Embora o modelo misto linear seja amplamente utilizado, os estimadores usuais de seus parâmetros são baseados na verossimilhança normal multivariada e, conseqüentemente, não são robustos a *outliers* ou a valores extremos. Em [35], a distribuição t de Student é utilizada para amortecer o impacto de eventuais *outliers* ou valores extremos. Mostraremos neste capítulo que este enfoque pode ser estendido a uma classe mais ampla de distribuições, as misturas na escala de gaussianas, e ainda manter a propriedade desejada de robustez.

7.2 O Modelo

Normalmente, os dados são modelados assumindo-se que os erros $\boldsymbol{\epsilon}$ e os efeitos aleatórios seguem as distribuições gaussianas

$$\begin{aligned} \mathbf{y}|\mathbf{u} &\sim \mathcal{N}(\mathbf{Xc} + \mathbf{Zu}, \mathbf{R}), \\ \mathbf{u} &\sim \mathcal{N}(\mathbf{0}, \mathbf{G}). \end{aligned} \tag{7.1}$$

Supondo-se que os dados podem ser divididos em c classes e considerando-se $\mathbf{R} \equiv \delta_\epsilon^2 \mathbf{I}$,

$$\mathbf{G} \equiv \begin{pmatrix} \delta_{u,1}^2 \mathbf{I}_{q_1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \delta_{u,2}^2 \mathbf{I}_{q_2} & \cdots & \mathbf{0} \\ \vdots & \vdots & & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \delta_{u,c}^2 \mathbf{I}_{q_c} \end{pmatrix}$$

e $\mathbf{u} = (\mathbf{u}_1', ..., \mathbf{u}_c')'$, onde $u_{j,k}$ é a k-ésima entrada de $\mathbf{u}_j$ e, para cada j, $\mathbf{u}_j$ é um vetor de dimensão q_j, temos que o modelo (7.1) pode ser reescrito na forma

$$\begin{aligned} y_t|\mathbf{u} &\sim \mathcal{N}((\mathbf{Xc} + \mathbf{Zu})_t, \delta_\epsilon^2), \\ u_{j,k} &\sim \mathcal{N}(0, \delta_{u,j}^2). \end{aligned} \tag{7.2}$$

para $1 \leq t \leq T$, $1 \leq j \leq c$ e $1 \leq k \leq q_j$, onde os índices t e as duplas (j,k) estão biunivocamente relacionados. Em resumo, deste modo, os dados estão particionados em c classes onde, para cada classe j, a variância das correspondentes variáveis aleatórias $u_{j,k}$ é dada por $\delta_{u,j}^2$.

Uma possibilidade para aumentar a robustez o modelo (7.2) é aquela proposta em [35] por Staudenmayer *et al* de acordo com a qual

$$\begin{aligned} y_t|\mathbf{u} &\sim t((\mathbf{Xc} + \mathbf{Zu})_t, \delta_\epsilon^2, \nu_y), \\ u_{j,k} &\sim t(0, \delta_{u,j}^2, \nu_u), \end{aligned} \tag{7.3}$$

para $1 \leq t \leq T$, $1 \leq j \leq c$ e $1 \leq k \leq q_j$. Finalmente, usando o fato que a distribuição t de Student é uma mistura na escala de normais, temos que o modelo (7.3) pode ser escrito na forma

$$\begin{aligned} \mathbf{y}|\mathbf{u}, \boldsymbol{\sigma}_y &\sim \mathcal{N}\left(\mathbf{Xc} + \mathbf{Zu}, \delta_\epsilon^2 \text{diag}\left(\frac{1}{\boldsymbol{\sigma}_y}\right)\right), \\ \mathbf{u}|\boldsymbol{\sigma}_u &\sim \mathcal{N}(\mathbf{0}, \tilde{\mathbf{G}}), \end{aligned} \tag{7.4}$$

onde

$$\tilde{\mathbf{G}} = \begin{pmatrix} \delta_{u,1}^2 \frac{1}{\boldsymbol{\sigma}_{u,1}} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \delta_{u,2}^2 \frac{1}{\boldsymbol{\sigma}_{u,2}} & \cdots & \mathbf{0} \\ \vdots & \vdots & & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \delta_{u,c}^2 \frac{1}{\boldsymbol{\sigma}_{u,c}} \end{pmatrix},$$

$1/\boldsymbol{\sigma}_y = (1/\sigma_{y,1}, ..., 1/\sigma_{y,T})'$, $\boldsymbol{\sigma}_\mathbf{u} = (\boldsymbol{\sigma}_{u,1}', ..., \boldsymbol{\sigma}_{u,c}')'$, $1/\boldsymbol{\sigma}_{u,j} = (1/\sigma_{u,j,1}, ..., 1/\sigma_{u,j,q_j})'$ e, finalmente,

$$\begin{aligned} \sigma_{y,t} &\sim \chi_{\nu_y}^2/\nu_y \\ \sigma_{u,j,k} &\sim \chi_{\nu_u}^2/\nu_u. \end{aligned}$$

7.3 Conexão com Splines

Nesta seção explicitaremos como se dá a conexão entre modelos mistos e *splines*. Suponha que, para cada $t = 1, ..., T$,

$$y_t = f(x_t) + \epsilon_t \tag{7.5}$$

e que f pode ser representada por um determinado número de elementos de uma base de potências truncadas,

$$f(x_t) = c_0 + c_1 x_t + \sum_{k=1}^{K} u_k (x_t - \kappa_k)_+$$

onde $\{\kappa_1, ..., \kappa_K\}$ é um conjunto arbitrário de nós. Deste modo, usando a notação

$$\mathbf{c} = (c_0, c_1)' \text{ e } \mathbf{u} = (u_1, .., u_K)',$$

para os coeficientes e

$$\mathbf{X} = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_T \end{bmatrix} \text{ e } \mathbf{Z} = \begin{bmatrix} (x_1 - \kappa_1)_+ & \cdots & (x_1 - \kappa_K)_+ \\ \vdots & & \vdots \\ (x_T - \kappa_1)_+ & \cdots & (x_T - \kappa_K)_+ \end{bmatrix}$$

para a base de potências truncadas, o critério de ajuste via splines por mínimos quadrados penalizados pode ser escrito na forma

$$\frac{1}{\delta_\epsilon^2}\|\mathbf{y} - \mathbf{Xc} - \mathbf{Zu}\|^2 + \frac{\lambda^2}{\delta_\epsilon^2}\|\mathbf{u}\|^2. \tag{7.6}$$

Agora, supondo que $\mathbf{u}$ é um vetor aleatório com matriz de covariâncias $\text{Cov}(\mathbf{u}) = \delta_u^2 \mathbf{I}$, onde $\delta_u^2 = \delta_\epsilon^2/\lambda^2$, e média zero, podemos obter estimativas para (7.6) através do critério do melhor preditor linear não-viciado.[1] Além disso, tudo isso junto resulta em uma representação característica de modelos mistos para o modelo de regressão via splines,

$$\mathbf{y} = \mathbf{Xc} + \mathbf{Zu} + \boldsymbol{\epsilon}, \quad \text{Cov}\begin{bmatrix} \mathbf{u} \\ \boldsymbol{\epsilon} \end{bmatrix} = \begin{bmatrix} \delta_u^2 \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \delta_\epsilon^2 \mathbf{I} \end{bmatrix}.$$

Em outras palavras, pode-se assim modelar uma eventual relação não-linear entre a variável resposta e a variável explicativa via *splines* usando-se o aparato computacional, ou *software*, já existente para modelos mistos.

Notamos, finalmente, que a matriz de covariâncias associada a $\mathbf{u}$ pode ser generalizada de modo a incorporar correlações entre elementos de determinadas classes (que compõem a massa de dados considerada) para uma matriz $\mathbf{G}$ e que o fator de penalização acima pode ser generalizado para $\mathbf{u}'\Omega_u\mathbf{u}$. De fato, neste caso, decompodo a matriz $\Omega_u = \mathbf{A}'\mathbf{DA}$, onde $\mathbf{D}$ é a matriz diagonal composta dos autovalores de Ω_u e $\mathbf{A}$ é a matriz ortonormal formada pelos respectivos autovetores, e definindo $\mathbf{Z}^* = \mathbf{ZA}'$ e $\mathbf{v} = \mathbf{Au}$, o problema

$$\frac{1}{\sigma_\epsilon^2}\|\mathbf{y} - \mathbf{Xc} - \mathbf{Zu}\|^2 + \frac{\lambda^2}{\sigma_\epsilon^2}\mathbf{u}'\Omega_u\mathbf{u} \tag{7.7}$$

pode ser reescrito na forma de um modelo misto dado por

$$\mathbf{y} = \mathbf{Xc} + \mathbf{Z}^*\mathbf{v} + \boldsymbol{\epsilon}, \quad \text{Cov}\begin{bmatrix} \mathbf{v} \\ \boldsymbol{\epsilon} \end{bmatrix} = \begin{bmatrix} \mathbf{AGA}' & \mathbf{0} \\ \mathbf{0} & \sigma_\epsilon^2 \mathbf{I} \end{bmatrix}.$$

[1] o método do *melhor preditor não-viciado linear* consiste em estimar a parte fixa $\mathbf{c}$ utilizando-se o modelo

$$\mathbf{y} = \mathbf{Xc} + \boldsymbol{\epsilon}^*,$$

onde $\boldsymbol{\epsilon}^* = \boldsymbol{\epsilon} + \mathbf{Zu}$ de modo que $\widehat{\mathbf{c}} = (\mathbf{X}'\mathbf{VX})^{-1}\mathbf{X}'\mathbf{Vy}$, com $\mathbf{V} = \mathbf{ZGZ}' + \mathbf{R}$, onde $\mathbf{G} = \text{Cov}(\mathbf{u})$ e $\mathbf{R} = \text{Cov}(\boldsymbol{\epsilon})$. Finalmente, para obter uma "estimativa" para $\mathbf{u}$ busca-se o vetor que minimiza o erro quadrático médio dentre os elementos da forma $\mathbf{Ay} + \mathbf{b}$, daí o nome linear. A solução é dada por $\mathbf{GZ}'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{Xc})$.

Em ambos os casos, os efeitos de cada grupo são modelados não-linearmente via *splines*, enquanto que a componente relativa à amostra como um todo permanece linear em relação às variáveis explicativas.

7.4 Misturas na Escala de Gaussianas

Nesta seção estendemos o uso de modelos mistos para a classe de distribuições formadas pela mistura na escala de gaussianas. Dado que esta classe inclui, entre outras, a distribuição t de Student e a normal, ela generaliza os modelos acima.

Consideremos inicialmente o seguinte modelo

$$\mathbf{y} = \mathbf{X}\mathbf{c} + \mathbf{Z}\mathbf{u} + \delta\boldsymbol{\epsilon}, \tag{7.8}$$

onde $\boldsymbol{\epsilon}$ e $\mathbf{u}$ são independentes e satisfazem $\mathbf{u} = \Delta_u \mathbf{u}_0$,

$$u^0_{j,k} \overset{\text{ind}}{\sim} SM_g(0,1,\boldsymbol{\zeta}_g) \text{ e } \epsilon_t \overset{\text{ind}}{\sim} SM_h(0,1,\boldsymbol{\zeta}_h)$$

para $t = 1, ..., T$, $1 \leq j \leq c$ e $1 \leq k \leq q_j$. Os vetores $\boldsymbol{\zeta}_g$ e $\boldsymbol{\zeta}_h$ contém os parâmetros associados às densidades g e h, respectivamente. Além disso, assumindo-se que $\sum_{j=1}^{c} q_j = K$, então Δ_u é uma matriz $K \times K$ dada por

$$\Delta_u = \begin{bmatrix} \delta_{u,1} I_{q_1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \delta_{u,2} I_{q_2} & \cdots & \mathbf{0} \\ \vdots & \vdots & & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \delta_{u,c} I_{q_c} \end{bmatrix}$$

onde I_{q_j} é a matriz diagonal $q_j \times q_j$. Por definição, podemos representar as variáveis aleatórias ϵ_t e $u^0_{j,k}$ por $Z/\sigma_t^{1/2}$ e $Z/\pi_{j,k}^{1/2}$, respectivamente, onde $Z \sim \mathcal{N}(0,1)$, $\sigma_t \overset{\text{ind}}{\sim} h$ e $\pi_{j,k} \overset{\text{ind}}{\sim} g$. Esta representação é extremamente útil para a obtenção de um algoritmo de estimação genérico, via o algoritmo EM, para toda a classe de distribuições obtida através da mistura na escala de gaussianas. No entanto, como as distribuições associadas a $\boldsymbol{\epsilon}$ e $\mathbf{u}$ não são gaussianas, não podemos aplicar o método sugerido na seção anterior dado que não estamos lidando com a soma entre duas variáveis seguindo distribuições normais, mas com a soma entre duas variáveis seguindo distribuições mais complexas e, possivelmente, distintas. Assim como nos demais capítulos e em [35], o algoritmo EM é usado para obter as estimativas do parâmetros do modelo, enquanto que a média da amostra obtida via simulação para $\mathbf{u}$ é usada como preditor do efeito aleatório desta mesma variável, ie, do efeito aleatório.

Conexão com Splines

Para adequar o formalismo acima ao caso do modelo da seção 7.3, basta notar que, como as componentes de $\mathbf{u}$ são não-correlacionados, então devemos simplesmente tomar $\Delta_u = \frac{\delta_\epsilon^2}{\lambda} I_K$.

7.4.1 Densidades Condicionais e Funções de Verossimilhança

Trataremos $\mathbf{u}$ como um vetor de variáveis aleatórias latentes ou não-observáveis e modelaremos $\mathbf{y}$ e $\mathbf{u}$ de acordo com

$$\begin{aligned} y_t|\mathbf{u}, \sigma_t, \boldsymbol{\theta} &\overset{\text{ind}}{\sim} \mathcal{N}\left((\mathbf{Xc} + \mathbf{Zu})_t, \frac{\delta^2}{\sigma_t}\right), \\ u^0_{j,k}|\pi_{j,k}, \boldsymbol{\theta} &\overset{\text{ind}}{\sim} \mathcal{N}\left(0, \frac{1}{\pi_{j,k}}\right), \end{aligned} \tag{7.9}$$

onde $\boldsymbol{\theta}$ representa o vetor de parâmetros do modelo. Em particular, isto resulta em

$$u_{j,k}|\pi_{j,k}, \boldsymbol{\theta} \overset{\text{ind}}{\sim} \mathcal{N}\left(0, \frac{\delta^2_{u,j}}{\pi_{j,k}}\right). \tag{7.10}$$

Proposição 7.4.1. *A menos de uma constante (em relação aos parâmetros $\boldsymbol{\theta}$), a log-verossimilhança associada a $\boldsymbol{\theta}$ é dada por*

$$\begin{aligned} l(\mathbf{c}, \delta, \boldsymbol{\delta}_u, \boldsymbol{\zeta}_h, \boldsymbol{\zeta}_g|\mathbf{y}; \mathbf{u}, \boldsymbol{\sigma}, \boldsymbol{\pi}) = &-\frac{T}{2}\log\delta^2 - \frac{1}{2\delta^2}\sum_{t=1}^{T}\sigma_t(y_t - \mathbf{X}_t'\mathbf{c} - \mathbf{Z}_t'\mathbf{u})^2 \\ &- \frac{1}{2}\sum_{j=1}^{c} q_j \log\delta^2_{u,j} - \frac{1}{2}\sum_{j=1}^{c}\frac{1}{\delta_{u,j}}\sum_{k=1}^{q_j}\pi_{jk}u^2_{jk} \\ &+ \sum_{t=1}^{T}\log h(\sigma_t|\boldsymbol{\zeta}_h) + \sum_{j=1}^{c}\sum_{k=1}^{q_j}\log g(\pi_{jk}|\boldsymbol{\zeta}_g), \end{aligned} \tag{7.11}$$

onde $\mathbf{X}_t$ e $\mathbf{Z}_t$ são a t-ésima linha de $\mathbf{X}$ e $\mathbf{Z}$, respectivamente.

Demonstração. Basta escrever a densidade conjunta de $(\mathbf{y}, \mathbf{u}, \boldsymbol{\sigma}, \boldsymbol{\pi})$ como o produto de densidades condicionais e tomar o logaritmo,

$$\begin{aligned} l(\mathbf{c}, \delta, \boldsymbol{\zeta}_h, \boldsymbol{\zeta}_g|\mathbf{y}; \mathbf{u}, \boldsymbol{\sigma}, \boldsymbol{\pi}) = &\log p(\mathbf{y}|\mathbf{u}, \boldsymbol{\sigma}; \mathbf{c}, \delta) + \log p(\mathbf{u}|\boldsymbol{\pi}; \boldsymbol{\delta}_u) \\ &+ \log p(\boldsymbol{\sigma}|\boldsymbol{\zeta}_h) + \log p(\boldsymbol{\pi}|\boldsymbol{\zeta}_g) \end{aligned}$$

para, então, aplicar (7.9) e (7.10). □

Conexão com Splines

No caso do modelo em 7.3, a expressão (7.11) pode ser simplificada para

$$\begin{aligned} l(\mathbf{c}, \delta, \boldsymbol{\delta}_u, \boldsymbol{\zeta}_h, \boldsymbol{\zeta}_g | \mathbf{y}; \mathbf{u}, \boldsymbol{\sigma}, \boldsymbol{\pi}) = & -\frac{T}{2} \log \delta^2 - \frac{1}{2\delta^2} \sum_{t=1}^{T} \sigma_t (y_t - \mathbf{X}_t' \mathbf{c} - \mathbf{Z}_t' \mathbf{u})^2 \\ & - \frac{K}{2} \log \delta_u^2 - \frac{1}{2\delta_u} \sum_{k=1}^{K} \pi_k u_k^2 \\ & + \sum_{t=1}^{T} \log h(\sigma_t | \boldsymbol{\zeta}_h) + \sum_{k=1}^{K} \log g(\pi_k | \boldsymbol{\zeta}_g). \end{aligned}$$

7.4.2 O Algoritmo

Como antes, o algoritmo de estimação dos parâmetros, cujas estimativas são de máxima verossimilhança, consiste na aplicação do algoritmo EM. Lembramos que, a cada passo do algoritmo, a primeira etapa (etapa E), como nos casos anteriores, consiste no cálculo do valor esperado da log-verossimilhança (7.11) dado $\mathbf{y}$. No entanto, agora, temos a inclusão de mais algumas variáveis latentes ao modelo, a saber, $\mathbf{u}$ e $\boldsymbol{\pi}$, o que torna ainda mais difícil a obtenção de uma expressão análitica para estas esperanças. A solução consiste no uso de métodos de Monte-Carlo. Em resumo, no passo $(k+1)$ do algoritmo, dado $\boldsymbol{\theta}^{(k)}$, sua primeira etapa consiste em calcular

$$Q(\boldsymbol{\theta} | \boldsymbol{\theta}^{(k)}) \equiv E_{\boldsymbol{\theta}^{(k)}} \{ l(\mathbf{c}, \delta, \boldsymbol{\delta}_u, \boldsymbol{\zeta}_h, \boldsymbol{\zeta}_g | \mathbf{y}; \mathbf{u}, \boldsymbol{\sigma}, \boldsymbol{\pi}) | \mathbf{y} \},$$

que, na prática, é aproximada por

$$\frac{1}{M-m} \sum_{j=m}^{M} l(\mathbf{c}, \delta, \boldsymbol{\delta}_u, \boldsymbol{\zeta}_h, \boldsymbol{\zeta}_g | \mathbf{y}; \mathbf{u}^{(j)}, \boldsymbol{\sigma}^{(j)}, \boldsymbol{\pi}^{(j)}), \tag{7.12}$$

onde $\{\mathbf{u}^{(j)}, \boldsymbol{\sigma}^{(j)}, \boldsymbol{\pi}^{(j)}\}_{j=1}^{M}$ é uma amostra pseudo-aleatória de $\mathbf{u}, \boldsymbol{\sigma}, \boldsymbol{\pi}$ obtida de acordo com as distribuições condicionais destas variáveis dado $\mathbf{y}$ e onde valor m indica o ponto a partir do qual começamos a considerar a amostra gerada.

No caso em que os parâmetros em $\boldsymbol{\zeta}_g$ e $\boldsymbol{\zeta}_h$ são conhecidos, temos

Proposição 7.4.2. *Definidas as matrizes*

$$W^{(k)} = diag \left(E_{\boldsymbol{\theta}^{(k)}} \{\sigma_1 | \mathbf{y}\}, ..., E_{\boldsymbol{\theta}^{(k)}} \{\sigma_T | \mathbf{y}\} \right),$$

$$U_j^{(k)} = diag \left(E_{\boldsymbol{\theta}^{(k)}} \{u_{j,1} | \mathbf{y}\}, ..., E_{\boldsymbol{\theta}^{(k)}} \{u_{j,q_j} | \mathbf{y}\} \right)$$

e

$$P_j^{(k)} = diag \left(E_{\boldsymbol{\theta}^{(k)}} \{\pi_{j,1} u_{j,1}^2 | \mathbf{y}\}, ..., E_{\boldsymbol{\theta}^{(k)}} \{\pi_{j,q_j} u_{j,q_j}^2 | \mathbf{y}\} \right)$$

para $j = 1, ..., c$, *temos que*

$$\begin{aligned} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)}) = & -\frac{1}{2\delta^2}(\mathbf{y} - \mathbf{Xc})'W^{(k)}(\mathbf{y} - \mathbf{Xc}) - 2J'_{K,1}(U^{(k)})'\mathbf{Z}'W^{(k)}(\mathbf{y} - \mathbf{Xc}) \\ & + J'_{K,1}(U^{(k)})'W^{(k)}U^{(k)}J_{K,1} - \frac{1}{2}\Delta_u J'_{K,1}P^{(k)} \\ & - \frac{1}{2}\sum_{j=1}^{c} q_j \log \delta^2_{u,j} - \frac{T}{2}\log \delta^2. \end{aligned} \tag{7.13}$$

onde $J_{m,n}$ *é a matriz de 1's com* m *linhas e* n *colunas, e*

$$U^{(k)} = \begin{pmatrix} U_1^{(k)} & 0 & \cdots & 0 \\ 0 & U_2^{(k)} & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & U_c^{(k)} \end{pmatrix}$$

e

$$P^{(k)} = \begin{pmatrix} P_1^{(k)} & 0 & \cdots & 0 \\ 0 & P_2^{(k)} & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & P_c^{(k)} \end{pmatrix}.$$

No entanto, como as esperanças na proposição acima não podem, em geral, ser calculadas analíticamente, a expressão (7.13) para $Q(\cdot|\boldsymbol{\theta}^{(k)})$ é aproximada via Monte-Carlo.

Proposição 7.4.3. *No* $(k+1)$*-ésimo passo do algoritmo, dado* $\boldsymbol{\theta}^{(k)}$*, temos que*

$$\mathbf{c}^{(k+1)} = (\mathbf{X}'\widehat{W}^{(k)}\mathbf{X})^{-1}\mathbf{X}'(\widehat{W}^{(k)}\mathbf{y} - \tilde{\mathbf{u}}_T^{(k)}) \tag{7.14}$$

onde[2]

$$\widehat{W}^{(k)} = \operatorname{diag}\left(\frac{1}{M-m}\sum_{i=m}^{M}\sigma_1^{(i)}, ..., \frac{1}{M-m}\sum_{i=m}^{M}\sigma_T^{(i)}\right) \tag{7.15}$$

e

$$\tilde{\mathbf{u}}_T^{(k)} = \left(\frac{1}{M-m}\sum_{i=m}^{M}\sigma_1^{(i)}\mathbf{Z}_1'\mathbf{u}^{(i)}, ..., \frac{1}{M-m}\sum_{i=m}^{M}\sigma_T^{(i)}\mathbf{Z}_T'\mathbf{u}^{(i)}\right)'. \tag{7.16}$$

Além disso,

$$(\delta^2)^{(k+1)} = \frac{1}{T}\left[\frac{1}{M-m}\sum_{t=1}^{T}\sum_{i=m}^{M}\sigma_t^{(i)}(y_t - \mathbf{X}_t'\mathbf{c}^{(k+1)} - \mathbf{Z}_t'\mathbf{u}^{(i)})^2\right]$$

[2] $\widehat{W}^{(k)}$ é a aproximação por Monte-Carlo de $W^{(k)}$.

Demonstração. Tomando o gradiente de (7.12) com relação a $\mathbf{c}$ e igualando o resultado a zero, obtemos a igualdade

$$\mathbf{c}^{(k+1)} = \left(\frac{1}{M-m} \sum_{t=1}^{T} \sum_{i=m}^{M} \sigma_t^{(i)} \mathbf{X}_t \mathbf{X}_t' \right)^{-1} \left(\frac{1}{M-m} \sum_{t=1}^{T} \sum_{i=m}^{M} \sigma_t^{(i)} (y_t - \mathbf{Z}_t' \mathbf{u}^{(i)}) \mathbf{X}_t \right)$$

de modo que a expressão em (7.14) segue imediatamente da aplicação das expressões (7.15) e (7.16) na expressão acima. □

7.4.3 Sobre a Amostragem de u, σ e π

A amostragem de $\{\mathbf{u}, \boldsymbol{\sigma}, \boldsymbol{\pi}\}$ é realizada através do algoritmo de Gibbs e, dependendo de h e g, sugerimos o uso do algoritmo de Metropolis-Hastings dentro do algoritmo de Gibbs. Em primeiro lugar, note que as condicionais completas associadas a $\{\mathbf{u}, \boldsymbol{\sigma}, \boldsymbol{\pi}\}$ são dadas por

$$\begin{aligned} p(\boldsymbol{\pi}|\mathbf{y}; \boldsymbol{\sigma}, \mathbf{u}) &\propto p(\mathbf{u}|\boldsymbol{\pi}) g(\boldsymbol{\pi}), \\ p(\boldsymbol{\sigma}|\mathbf{y}; \boldsymbol{\pi}, \mathbf{u}) &\propto p(\mathbf{y}|\mathbf{u}, \boldsymbol{\sigma}) h(\boldsymbol{\sigma}), \\ p(\mathbf{u}|\mathbf{y}; \boldsymbol{\sigma}, \boldsymbol{\pi}) &\propto p(\mathbf{y}|\mathbf{u}, \boldsymbol{\sigma}) p(\mathbf{u}|\boldsymbol{\pi}) \end{aligned}$$

e que, independentemente de h e g, a distribuição condicional completa a posteriori de $\mathbf{u}$ é dada por uma normal multivariada como mostra a proposição abaixo.

Proposição 7.4.4. *A distribuição condicional completa a posteriori de* $\mathbf{u}$ *é uma distribuição gaussiana multivariada com média igual a* $(\mathbf{Z}'\Sigma\mathbf{Z} + \delta^2 \Pi \Delta_u^{-2})^{-1} \mathbf{Z}'\Sigma(\mathbf{y} - \mathbf{Xc})$ *e matriz de covariâncias* $\mathrm{Cov}(\mathbf{u}) = (\delta^{-2}\mathbf{Z}'\Sigma\mathbf{Z} + \Pi\Delta_u^{-2})^{-1}$. *Em outros termos, temos que*

$$\mathbf{u}|\mathbf{y}; \boldsymbol{\sigma}, \boldsymbol{\pi} \sim \mathcal{N}((\mathbf{Z}'\Sigma\mathbf{Z} + \delta^2\Pi\Delta_u^{-2})^{-1}\mathbf{Z}'\Sigma(\mathbf{y} - \mathbf{Xc}), (\delta^{-2}\mathbf{Z}'\Sigma\mathbf{Z} + \Pi\Delta_u^{-2})^{-1}).$$

Demonstração. Por definição,

$$p(\mathbf{y}|\mathbf{u}, \boldsymbol{\sigma}) p(\mathbf{u}|\boldsymbol{\pi}) \propto \exp\left\{ -\frac{1}{2\delta^2} (\mathbf{y} - \mathbf{Xc} - \mathbf{Zu})' \Sigma (\mathbf{y} - \mathbf{Xc} - \mathbf{Zu}) \right\} \cdot \exp\left\{ -\frac{1}{2} \mathbf{u}' \Pi \Delta_u^{-2} \mathbf{u} \right\}$$

onde $\Sigma = \mathrm{diag}(\sigma_1, ..., \sigma_T)$ e $\Pi = \mathrm{diag}(\pi_1, ..., \pi_K)$. Agora, definindo $\mathbf{r} \equiv \mathbf{y} - \mathbf{Xc}$, $\tilde{\Sigma} \equiv \frac{1}{\delta^2}\Sigma$ e $\tilde{\Pi} \equiv \Pi\Delta_u^{-2}$, temos que

$$\begin{aligned} p(\mathbf{u}|\mathbf{y}; \boldsymbol{\sigma}, \boldsymbol{\pi}) &\propto \exp\left\{ -\frac{1}{2} [(\mathbf{r} - \mathbf{Zu})' \tilde{\Sigma} (\mathbf{r} - \mathbf{Zu}) + \mathbf{u}' \tilde{\Pi} \mathbf{u}] \right\} \\ &\propto \exp\left\{ -\frac{1}{2} [-2\mathbf{r}' \tilde{\Sigma} \mathbf{Zu} + \mathbf{u}' \mathbf{Z}' \tilde{\Sigma} \mathbf{Zu} + \mathbf{u}' \tilde{\Pi} \mathbf{u}] \right\} \\ &\propto \exp\left\{ -\frac{1}{2} [-2\mathbf{r}' \tilde{\Sigma} \mathbf{Zu} + \mathbf{u}' (\mathbf{Z}' \tilde{\Sigma} \mathbf{Z} + \tilde{\Pi}) \mathbf{u}] \right\} \end{aligned}$$

de modo que, tomando $\mathbf{A} \equiv \mathbf{Z}'\tilde{\Sigma}\mathbf{Z} + \tilde{\Pi}$ e $\mathbf{a} \equiv \mathbf{A}^{-1}\mathbf{Z}'\tilde{\Sigma}\mathbf{r}$ e usando o fato que $\mathbf{A}$ é simétrica, temos

$$p(\mathbf{u}|\mathbf{y}; \boldsymbol{\sigma}, \boldsymbol{\pi}) \propto \exp\{(\mathbf{u} - \mathbf{a})' \mathbf{A} (\mathbf{u} - \mathbf{a})\}.$$

□

Os resultados acima estão resumidos no algoritmo 7.4.1.

Algoritmo 7.4.1 MCEM para estimação dos parâmetros de (7.8)

Dados $\theta^{(k)}$

1. Amostrar $\{\mathbf{u}^{(i)}, \boldsymbol{\sigma}^{(i)}, \boldsymbol{\pi}^{(i)}\}_{i=1}^{M}$ via amostrador de Gibbs:

i. amostrar $\boldsymbol{\pi}^{(k+1)} \sim p(\mathbf{u}^{(k)}|\boldsymbol{\pi})g(\boldsymbol{\pi})$;
ii. amostrar $\mathbf{u}^{(k+1)} \sim p(\mathbf{y}|\mathbf{u}, \boldsymbol{\sigma}^{(k)})p(\mathbf{u}|\boldsymbol{\pi}^{(k+1)})$;
iii. amostrar $\boldsymbol{\sigma}^{(k+1)} \sim p(\mathbf{y}|\mathbf{u}^{(k+1)}, \boldsymbol{\sigma})h(\boldsymbol{\sigma})$;

2. Obter a $(k+1)$-ésima estimativa de $\mathbf{c}$ através de

$$\mathbf{c}^{(k+1)} = (\mathbf{X}'\widehat{\mathbf{W}}^{(k)}\mathbf{X})^{-1}\mathbf{X}'(\widehat{\mathbf{W}}^{(k)}\mathbf{y} - \tilde{\mathbf{u}}_T^{(k)});$$

3. Obter $\delta^{(k+1)}$ através de:

$$(\delta^2)^{(k+1)} = \frac{1}{T}\left[\frac{1}{M-m}\sum_{t=1}^{T}\sum_{i=m}^{M}\sigma_t^{(i)}(y_t - \mathbf{X}_t'\mathbf{c}^{(k+1)} - \mathbf{Z}_t'\mathbf{u}^{(i)})^2\right];$$

4. Obter $\boldsymbol{\delta}_u^{(k+1)}$ através da maximização de

$$-\frac{1}{2}\sum_{j=1}^{c} q_j \log \delta_{u,j}^2 - \frac{1}{2}\sum_{j=1}^{c}\frac{1}{\delta_{u,j}}\sum_{k=1}^{q_j}\left[\frac{1}{M-m}\sum_{i=m}^{M}\pi_{jk}^{(i)}(u_{jk}^{(i)})^2\right];$$

5. Obter $\boldsymbol{\zeta}_h^{(k+1)}$ e $\boldsymbol{\zeta}_g^{(k+1)}$ através da maximização de

$$\sum_{t=1}^{T}\left[\frac{1}{M-m}\sum_{i=m}^{M}\log h(\sigma_t|\boldsymbol{\zeta}_h)\right] \text{ e } \sum_{j=1}^{c}\sum_{k=1}^{q_j}\left[\frac{1}{M-m}\sum_{i=m}^{M}\log g(\pi_{jk}|\boldsymbol{\zeta}_g)\right]$$

respectivamente.

Observação 7.4.1. *A depender de h e g, não será possível determinar explicitamente as distribuições das amostragens [i] e [iii] no primeiro passo do algoritmo 7.4.1 e, portanto, elas devem ser feitas via Metropolis-Hastings.*

Capítulo 8

Aplicação

8.1 Nível d'Água no Rio Moselle

Nesta aplicação, examinaremos os níveis da água no rio Moselle em Zeltingen entre os meses de janeiro de 1986 a janeiro de 1996. Observamos que, a partir de 1988, as medidas representam o máximo diário, enquanto que anteriormente era realizada apenas uma medida por dia. Este fato tende a superestimar o nível usual, ou nível médio, da água jogando para "cima" e certamente deve tornar estimativas via mínimos quadrados deste nível médio mais oscilantes do que ele realmente é. Nas análises subseqüentes consideraremos apenas os segundo bloco de observações, isto é, aquele a partir de 1988. Antes de qualquer outra análise, observamos

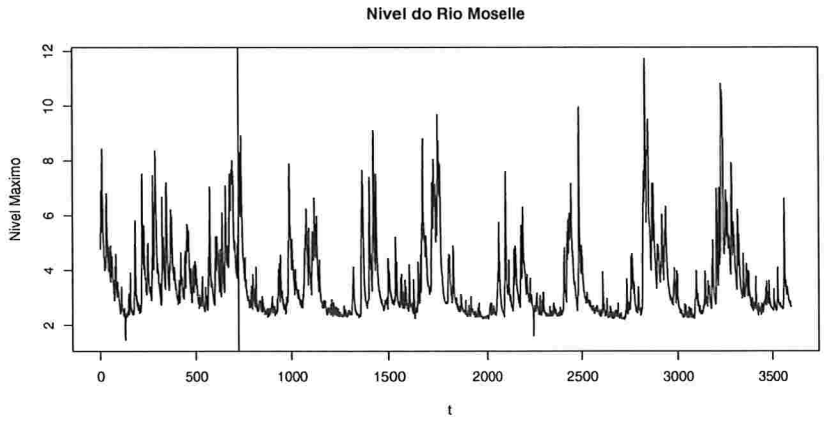


Figura 8.1: Nível da água no rio Moselle entre os anos 1986 e 1996. A linha vertical indica a mudança na metodologia de amostragem.

que a inspeção direta do gráfico na figura 8.1 sugere que os dados têm uma sazonalidade de aproximadamente um ano. De fato, o periodograma desta série, figura 8.2, revela que a freqüência associada ao valor máximo

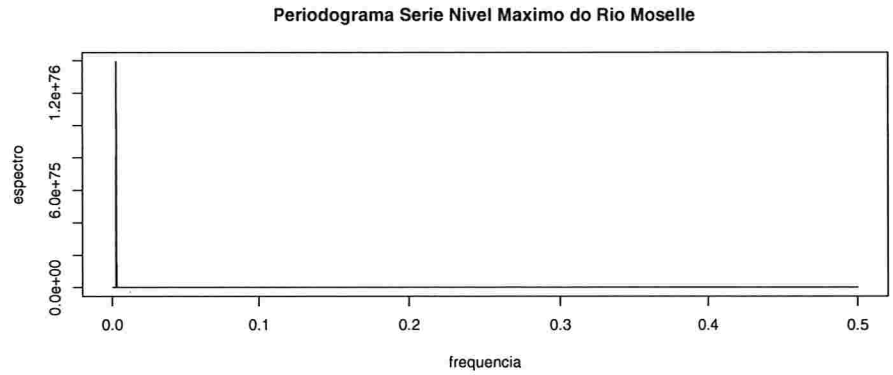


Figura 8.2: Peridograma da série representada pelo máximo do nível da água no rio Moselle entre os anos 1986 e 1996.

do espectro é 375 dias. Mais ainda, este periodograma (aliado à função de autocorrelação empírica) indica a presença de memória longa na série considerada. Felizmente, ao tomarmos a primeira diferença da série original, estas características indesejadas, sazonalidade e memória longa, são fortemente atenuadas e, portanto, na seção 8.1.1, modelaremos a variação diária do nível máximo do rio usando a técnica sugerida na tese e, apesar da sazonalidade relatada anteriormente, trataremos os dados diretamente na esperança de que esta sazonalidade seja absorvida pela função alvo estimada (não-linear).

Iniciamos a análise dos dados suavizando a curva subjacente aos dados assumindo ruído gaussiano e distribuido conforme uma t de Student. Calculando o valor AIC para diversos graus de liberdade, ν, e diversos valores de M pudemos checar que, fixado M, o AIC tende a ser menor para menores valores de ν. Portanto, optamos por tomar $\nu = 2$ e quanto a M, escolhemos $M = 50$. Os resultados destes procedimentos estão representados na figura 8.3. Os parâmetros de escala estimados estão na tabela 8.1. Para avaliar a qualidade

Tabela 8.1: Estimativas do parâmetro de escala, δ para os dados do nível d'água no rio Moselle obtidos para diferentes métodos e os seus respectivos intervalos de confiança com um nível de 95%.

	$\widehat{\delta^2}$	**IC95%**
MQ	0,988	[0,979;1,003]
t_2	0,129	[0,113;0,142]

dos ajustes fizemos os gráficos Q×Q dos resíduos e confrontamos o histograma dos resíduos ajustados contra as densidades das distribuições assumidas para o ruído em cada um dos ajustes, os quais podem ser vistos na figura 8.4. Dos gráficos Q×Q, pode-se notar que a distribuição dos resíduos para o ajuste via mínimos quadrados tem uma cauda significativamente mais pesada do que aquela assumida a priori, ie, da gaussiana. Esta incongruência entre as distribuições é ressaltada pela diferença entre o histograma e a densidade para a normal padrão exibida na figura 8.4 (d). Note que o ajuste via mínimos quadrados ilustra um rio com um nível d'água mais oscilante do que o ajuste robusto. Isto deve-se principalmente ao efeito dos valores extremos sobre a estimativa final.

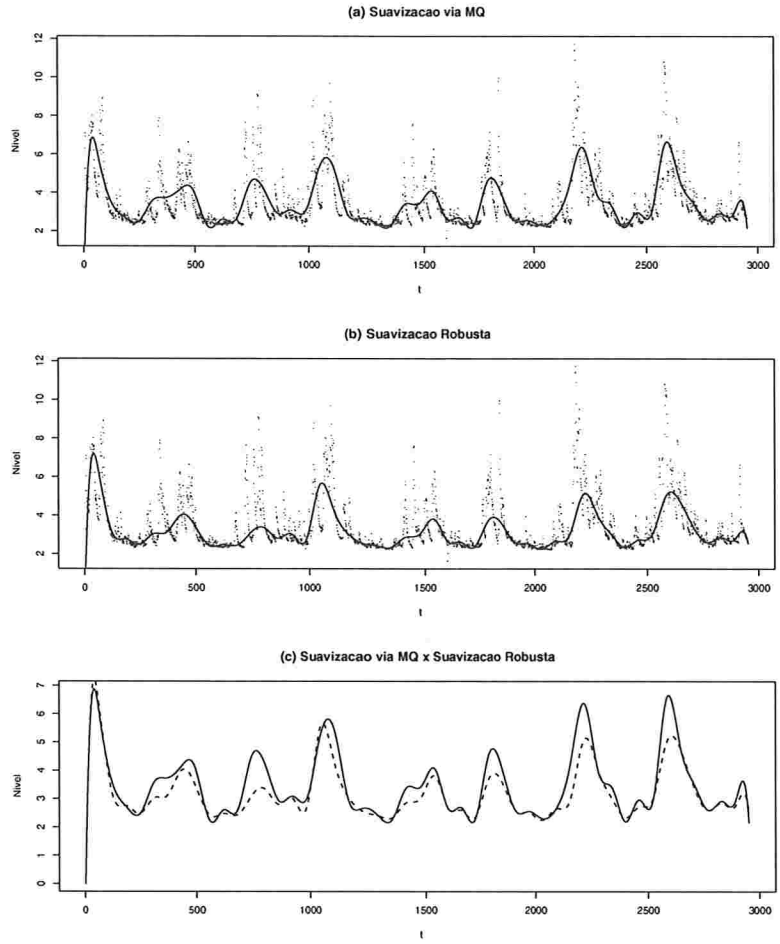


Figura 8.3: Suavização do nível da água no Moselle usando duas técnicas distintas. Nos dois gráficos superiores, a linha cheia representa a estimativa obtida, as linhas tracejadas representam bandas de confiança a um nível de 95% obtidas via *bootstrap* e os pontos representam as observações obtidas. Em (b), temos a estimativa obtida pelo método robusto cujas ponderações foram calculadas a partir de uma distribuição t de Student com 2 graus de liberdade. No gráfico intermediário, a suavização do nível do rio foi feita via M.Q.. Finalmente, o terceiro gráfico serve apenas para confrontar as duas estimativas.

8.1.1 Ajuste do Modelo Autorregressivo

A série para o nível d'água no rio Moselle apresenta uma clara sazonalidade (observada anteriormente) e uma heterocedasticidade que pode ser observada por períodos de pequena volatilidade associados às baixas nos níveis de água seguidos de períodos de grande volatilidade, desta vez associados aos periodos de altas no nível de água. Além disso, a série é positiva e o ruído aparenta uma certa assimetria. Logo, para eliminar estes efeitos da série, ie, para eliminar sazonalidade, estabilizar as variações de volatilidade e, finalmente, tornar

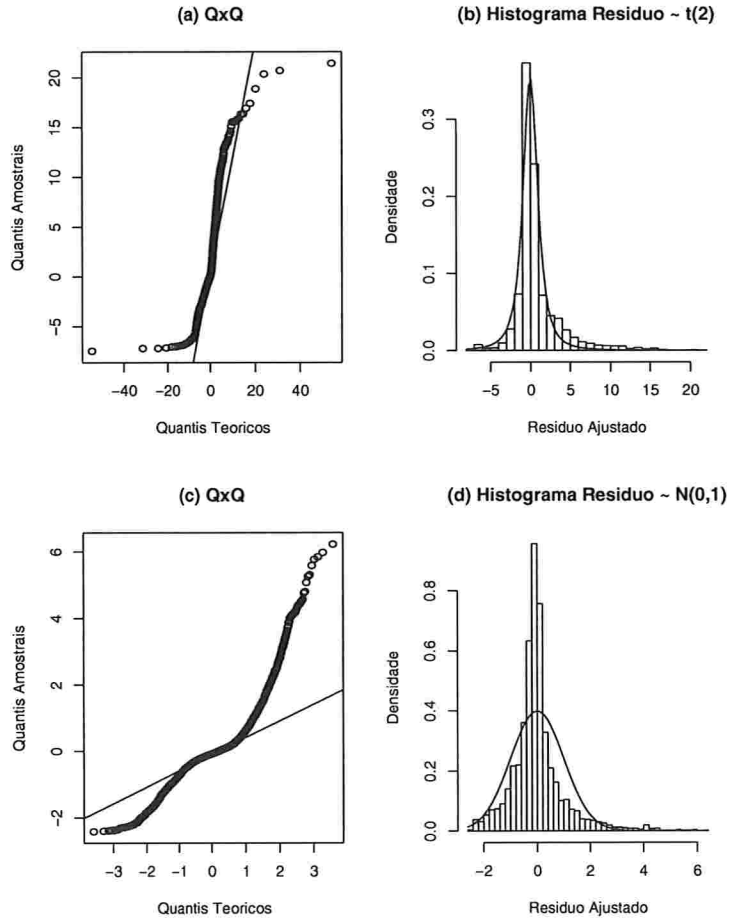


Figura 8.4: (a) Gráfico Q×Q para o ajuste assumindo ruídos distribuidos de acordo com uma t de Student com $\nu = 2$. (b) Histograma para os resíduos ajustados contra o densidade da t_2. (c) Gráfico Q×Q para o ajuste assumindo ruídos distribuidos de acordo com uma gaussiana $\mathcal{N}(0,1)$. (d) Histograma para os resíduos ajustados contra o densidade da $\mathcal{N}(0,1)$

a série simétrica em torno da média (como exigem os modelos estudados na tese) aplicamos o operador de diferença na série de modo a obtermos a nova série definida por

$$d_t = y_t - y_{t-1} = \Delta y_t,$$

onde y_t representa o nível d'água no instante t. O resultado foi a série representada na figura 8.5. As linhas horizontais pontilhadas exteriores delimitam 90% das observações, enquanto que as linhas intermediárias, também pontilhadas, representam o primeiro e o terceiro quartis associados às observações. É evidente que a série, de modo aproximado, oscila simetricamente em torno do zero, mas, ainda mais interessante, é que a presença ostensiva de *outliers* ou valores extremos de grande magnitude relativa é forte evidência de que esta

série não é um processo gaussiano e tem caudas pesadas. A figura 8.6 ilustra a função de autocorrelação para

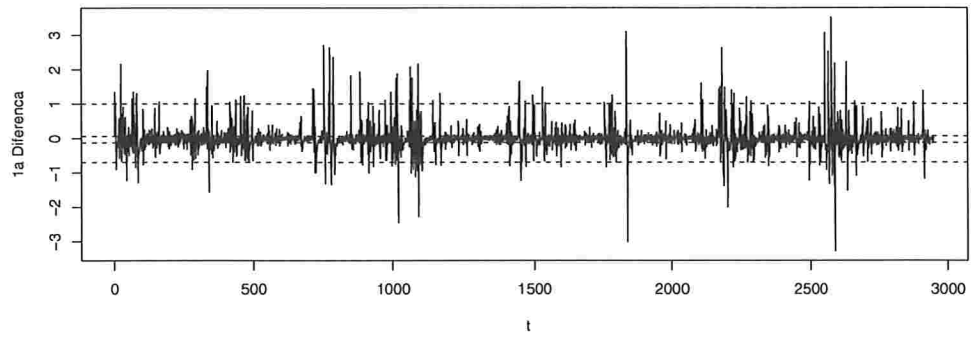


Figura 8.5: Série $d_t = \Delta y_t$ obtida a partir da diferenciação da série do rio Moselle. A série aparenta estacionariedade apesar da grande presença de *outliers*. As linhas tracejadas intermediárias representam, respectivamente, o primeiro e o terceiro quartil e as linhas tracejadas exteriores os quantis 2,5% e 97,5%, respectivamente.

$\{d_t\}$ e seu decaimento aparentemente suave indica a possibilidade desta série seguir um processo autorregressivo.

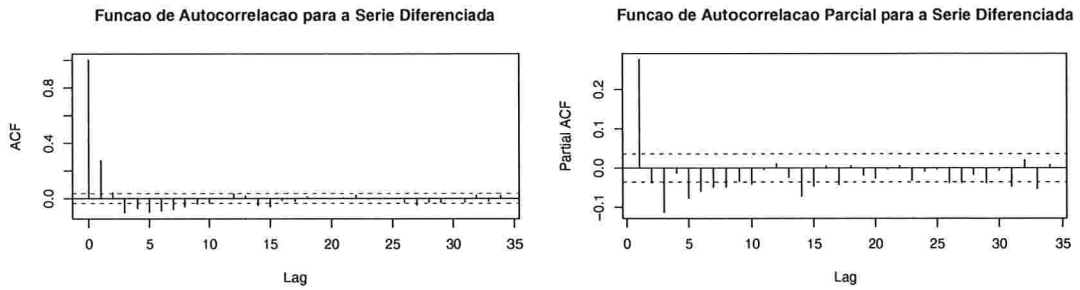


Figura 8.6: Funções de autocorrelação e de autocorrelação parcial amostrais para $\{d_t\}$.

Primeiro Ajuste — Modelo Autorregressivo Não-Linear de Ordem 1 via M.Q.

Embora tenhamos evidências de que o processo estudado não seja gaussiano, nosso primeiro ajuste será via mínimos quadrados. Assumiremos que

$$d_t = f(d_{t-1}) + \delta\epsilon_t \tag{8.1}$$

onde f é uma função suave e tentaremos aproximá-la usando bases B-splines. Para escolher o tamanho da base, testamos modelos com M (número de componentes na base) variando entre 5 e 20 e concluimos que

$M = 15$ é o que melhor se ajusta aos dados com relação ao critério AIC, veja figura 8.7. O resultado do ajuste

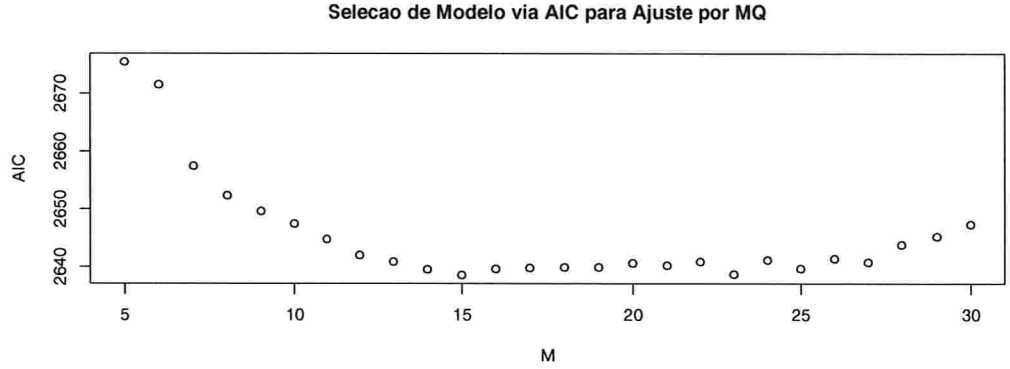


Figura 8.7: AIC's obtidos para os modelos ajustados com base em $\{d_t\}$ e fazendo M variar de 5 a 30. Como pode-se notar no gráfico, $M = 15$ corresponde ao menor AIC.

para este valor de M pode ser visualizado na figura 8.8. Ela indica que a série transformada $\{d_t\}$ é não-linear

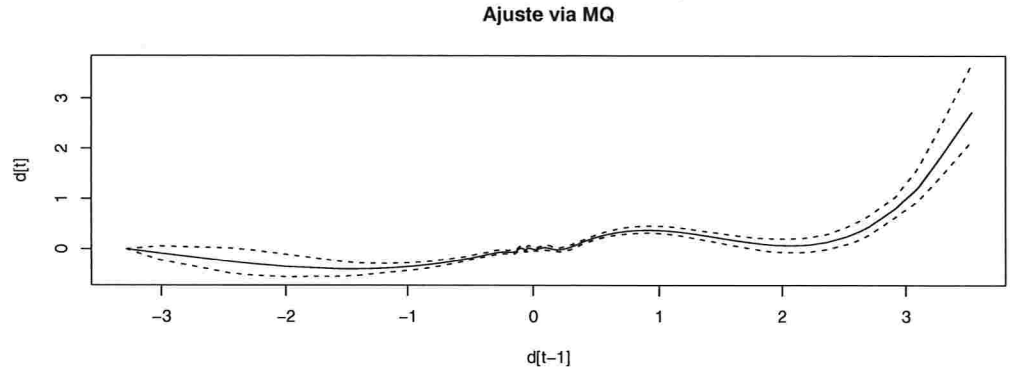


Figura 8.8: Ajuste de f com base em $\{d_t\}$ (rio Moselle) via B-splines tomando e $M = 15$ e suas respectivas bandas de confiança a 95% obtidas via *bootstrap*.

e que a função f tem um comportamento senoidal. Para avaliar a qualidade da estimativa, calculamos via *bootstrap* as bandas de confiança a um nível de 95% representadas pelas linhas tracejadas na figura 8.8 e também checamos o gráfico QQ-Plot dos resíduos ajustados,

$$\widehat{r}_t = \frac{d_t - \widehat{f}(d_{t-1})}{\widehat{\delta}},$$

contra os quantis da normal padrão, figura 8.9. A diferença entre a curva amostral e a curva teórica indica

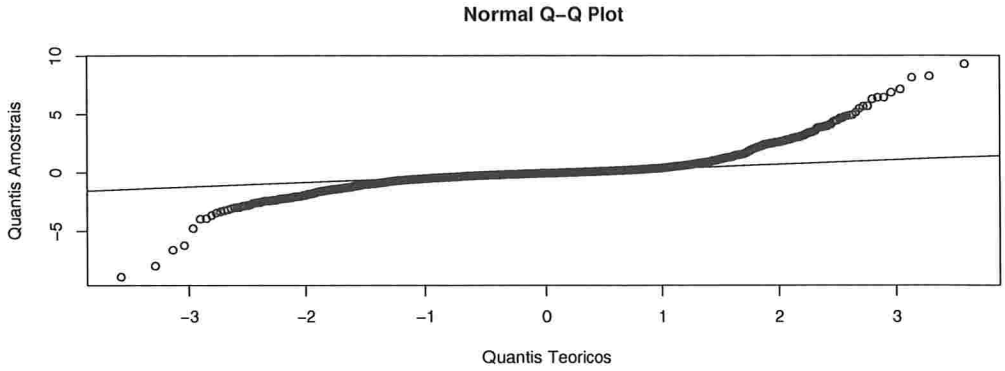


Figura 8.9: QQ-Plot normal baseado nos dados $\{d_t\}$ obtidos tomando-se a primeira diferença dos níveis do rio Moselle. A linha reta representa os valores teóricos para a normal padrão. O resultado indica claramente a presença de um ruído com caudas pesadas.

fortemente que o ruído segue uma distribuição com caudas pesadas e, desta forma, corrobora as impressões obtidas anteriormente. Para modelar o ruído, devemos, então, assumir para o mesmo uma distribuição com caudas mais pesadas do que a distribuição normal, a qual é implicitamente assumida quando usamos mínimos quadrados.

Ajuste Robusto

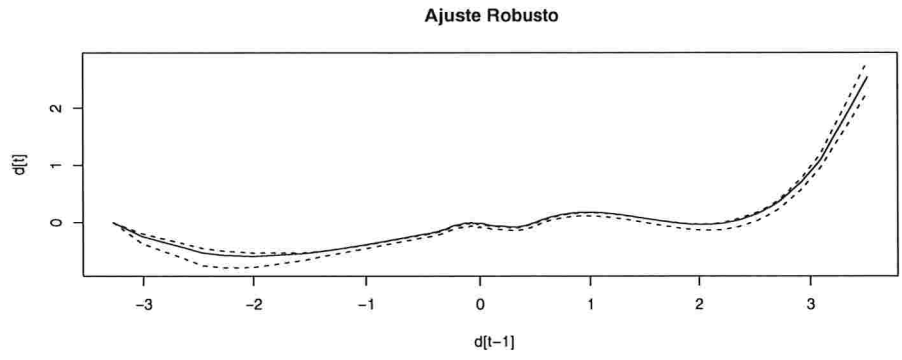


Figura 8.10: Ajuste de f com base em $\{d_t\}$ (rio Moselle) via B-splines tomando $M = 22$ e suas respectivas bandas de confiança a 95% obtidas via *bootstrap*. O ajuste foi feito assumindo que o ruído segue uma distribuição t de Student com 1,25 graus de liberdade.

Assumindo que o ruído segue uma t de Student, ajustamos os dados para uma série de valores distintos de

ν (os graus de liberdade associados a distribuição t) e M. Os resultados obtidos estão contidos na tabela 8.2 e lá podemos observar que o critério de AIC tende a escolher entre modelos com menores graus de liberdade, aproximando-se muito da distribuição de Cauchy. De fato, entre os modelos avaliados, a combinação com menor AIC foi aquela com $(\nu, M) = (1,25; 22)$. A figura 8.10 ilustra o resultado obtido usando-se estes parâmetros para o ajuste. Quando comparamos com o ajuste via mínimos quadrados, notamos que o ajuste robusto é mais preciso no sentido que tem bandas de confiança mais estreitas e que apresenta um comportamento levemente diferente nas proximidades de zero, com uma menor inclinação da curva e, conseqüentemente, indicando um intervalo de relativa estabilidade. Através da figura 8.11 é possível comparar ambos os ajustes em níveis de resolução mais altos. Com relação à qualidade do ajuste, na figura 8.12 temos o histograma dos resíduos ajustados, ie, divididos pelo parâmetro de escala, δ, cuja estimativa é dada por $\widehat{\delta^2} = 0,0097$, com $IC(95\%) = [0,0087; 0,0106]$ Observando o resultado exposto na figura 8.12, notamos que superdimensionamos o peso das caudas do ruído. Investigando outros modelos com ν próximo a $1,25$, chegamos a um outro candidato para o qual os resíduos ajustam-se melhor à distribuição assumida. Neste modelo, $\nu = 2$ e $M = 15$. Resultados análogos aos obtidos no ajuste anterior podem ser vistos nas figuras 8.13, 8.14 e 8.15.

Tabela 8.2: AIC's calculados para diferentes valores de ν e M associados usados em diversos ajustes do modelo 8.1 para os dados do nível d'água no rio Moselle. Dentre todos os modelos experimentados, aquele para o qual obtivemos o menor AIC foi o modelo correspondente aos parâmetros $\nu = 1,25$ e $M = 22$.

ν	M	AIC	ν	M	AIC	ν	M	AIC
1.00	12	21.4091	1.50	22	−3.8410	3.00	17	372.6283
1.00	13	20.0556	1.50	23	−3.2768	3.00	18	372.3267
1.00	14	19.5327	1.50	24	−3.0987	3.00	19	372.7184
1.00	15	16.6084	1.50	25	−1.7114	3.00	20	373.4162
1.00	16	14.7496	1.50	26	0.3357	3.00	21	373.4332
1.00	17	11.7796	1.75	12	53.0662	3.00	22	373.5685
1.00	18	10.2061	1.75	13	51.5761	3.00	23	373.9313
1.00	19	8.8968	1.75	14	49.9522	3.00	24	373.7985
1.00	20	8.8120	1.75	15	48.1519	3.00	25	375.1474
1.00	21	9.1617	1.75	16	46.9309	3.00	26	376.3851
1.00	22	7.0899	1.75	17	45.5104	4.00	12	610.9847
1.00	23	7.4423	1.75	18	44.7351	4.00	13	609.4610
1.00	24	7.5390	1.75	19	44.5613	4.00	14	608.1854
1.00	25	8.7370	1.75	20	45.0519	4.00	15	607.4935
1.00	26	11.1758	1.75	21	45.3440	4.00	16	607.6056
1.25	12	−14.6601	1.75	22	44.9174	4.00	17	607.6906
1.25	13	−16.0771	1.75	23	45.4804	4.00	18	607.6552
1.25	14	−17.1133	1.75	24	45.6024	4.00	19	608.2257
1.25	15	−19.5239	1.75	25	47.0337	4.00	20	609.0062
1.25	16	−21.0978	1.75	26	48.9057	4.00	21	609.0234
1.25	17	−23.4531	2.00	12	113.2976	4.00	22	609.2862
1.25	18	−24.6473	2.00	13	111.7107	4.00	23	609.5569
1.25	19	−25.4327	2.00	14	110.0436	4.00	24	609.4093
1.25	20	−25.2288	2.00	15	108.4315	4.00	25	610.6239
1.25	21	−24.8017	2.00	16	107.3715	4.00	26	611.6173
1.25	**22**	**−26.0600**	2.00	17	106.2847	5.00	12	802.5753
1.25	23	−25.5476	2.00	18	105.6363	5.00	13	801.1611
1.25	24	−25.3555	2.00	19	105.6438	5.00	14	800.1169
1.25	25	−24.0465	2.00	20	106.2064	5.00	15	799.6846
1.25	26	−21.8088	2.00	21	106.4147	5.00	16	800.1134
1.50	12	5.6647	2.00	22	106.1979	5.00	17	800.4457
1.50	13	4.2118	2.00	23	106.7317	5.00	18	800.6126
1.50	14	2.8057	2.00	24	106.7869	5.00	19	801.2888
1.50	15	0.7498	2.00	25	108.2323	5.00	20	802.1375
1.50	16	−0.6140	2.00	26	109.9449	5.00	21	802.2287
1.50	17	−2.4716	3.00	12	377.5186	5.00	22	802.5828
1.50	18	−3.4189	3.00	13	375.8611	5.00	23	802.8394
1.50	19	−3.8452	3.00	14	374.4191	5.00	24	802.7730
1.50	20	−3.4664	3.00	15	373.3199	5.00	25	803.8878
1.50	21	−3.0909	3.00	16	372.9494	5.00	26	804.7732

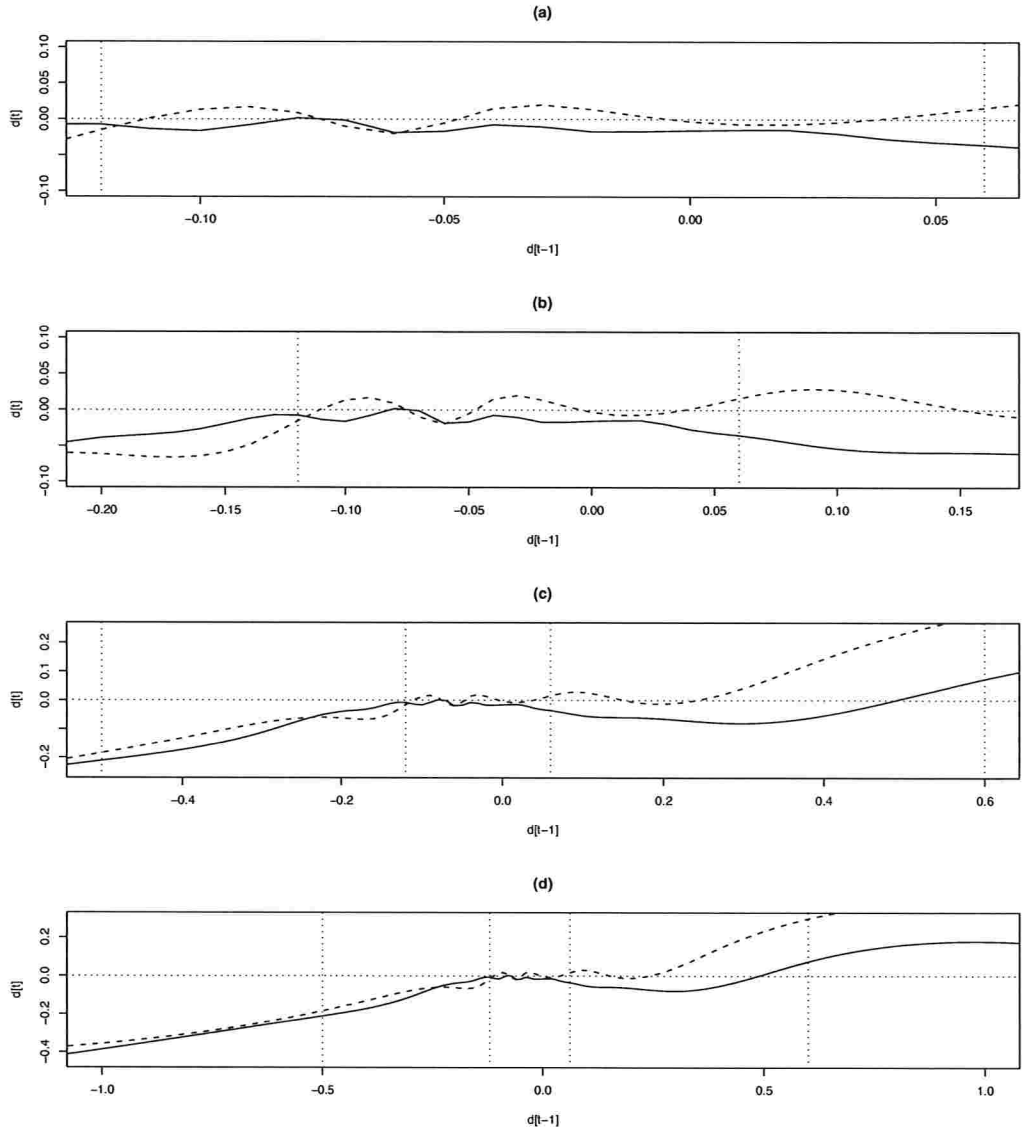


Figura 8.11: Comparação entre os ajuste via M.Q. e robusto em diversos níveis de resolucção. As linhas verticais pontilhadas representam os quantis 5%, 25%, 50% e 95%, respectivamente. O ajuste via M.Q. é representado pela curva tracejada e o ajuste assumindo um ruído distribuido de acordo com uma t de Student com 1,25 graus de liberdade, pela curva cheia.

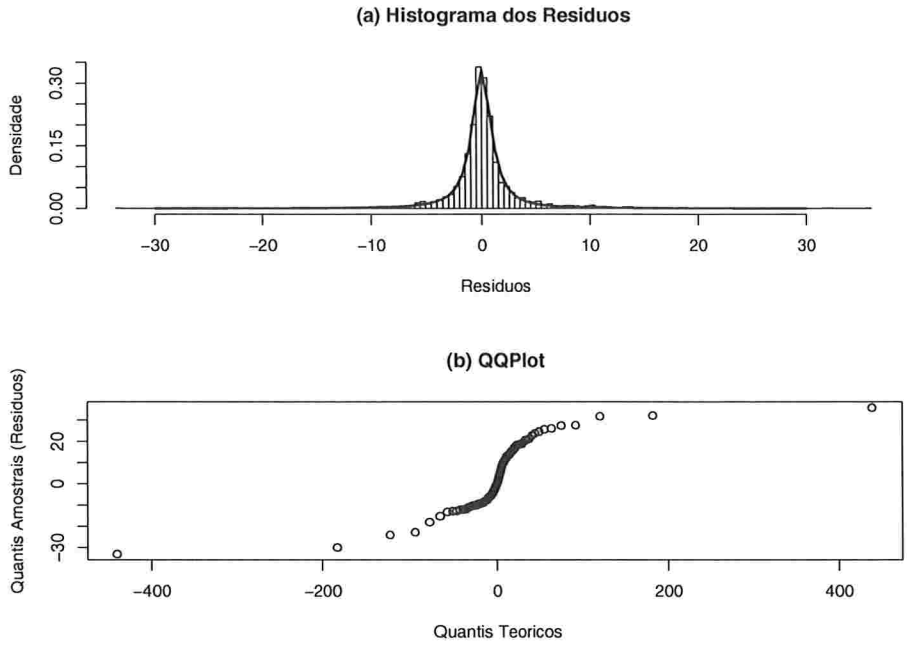


Figura 8.12: (a) Histograma dos resíduos ajustados versus a densidade da t de Student com $\nu = 1,25$. (b) QQ-Plot dos quantis amostrais amostrais associados aos resíduos ajustados versus os quantis teóricos da t de Student com $\nu = 1,25$.

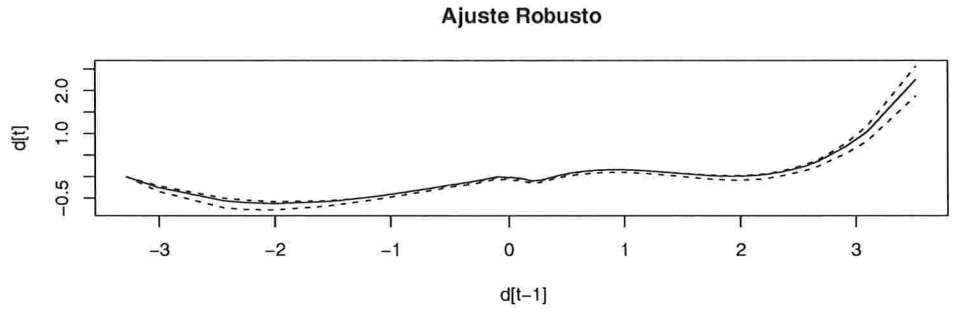


Figura 8.13: Ajuste de f com base em $\{d_t\}$ (rio Moselle) via B-splines tomando $M = 15$ e suas respectivas bandas de confiança a 95% obtidas via *bootstrap*. O ajuste foi feito assumindo que o ruído segue uma distribuição t de Student com 2 graus de liberdade.

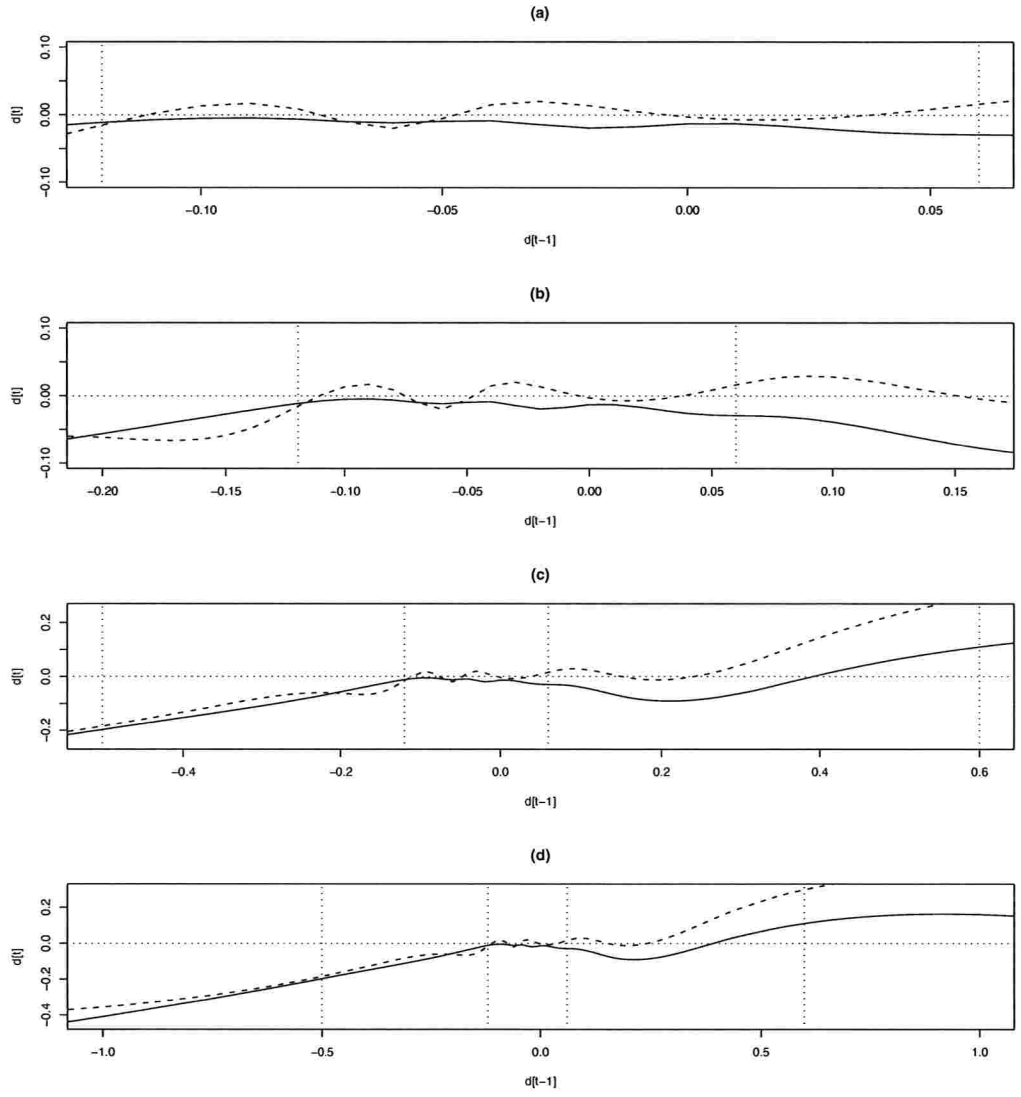


Figura 8.14: Comparação entre os ajuste via M.Q. e robusto em diversos níveis de resolucção. As linhas verticais pontilhadas representam os quantis 5%, 25%, 50% e 95%, respectivamente. O ajuste via M.Q. é representado pela curva tracejada e o ajuste assumindo um ruído distribuido de acordo com uma t de Student com 2 graus de liberdade e $M = 15$ funções base, pela curva cheia.

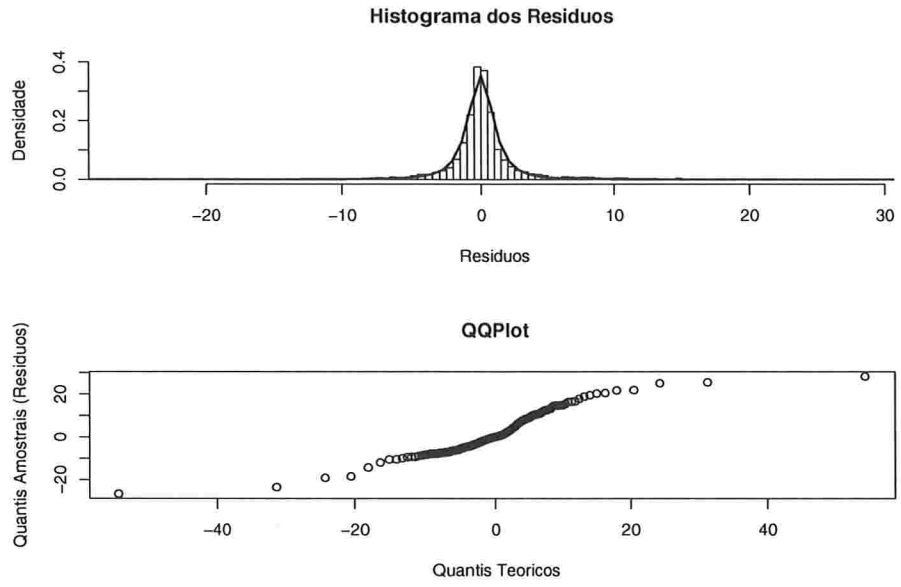


Figura 8.15: (a) Histograma dos resíduos ajustados versus a densidade da t de Student com $\nu = 2$ e $M = 15$. (b) QQ-Plot dos quantis amostrais amostrais associados aos resíduos ajustados versus os quantis teóricos da t de Student com $\nu = 2$ e $M = 15$.

Capítulo 9

Conclusão

Neste trabalho, mostramos que o uso de distribuições definidas por misturas de normais através do parâmetro de escala para modelos de séries temporais lineares autorregressivos resultam em estimadores consistentes para os coeficientes do modelo e para o parâmetro de escala, e estendemos seu uso para modelos lineares penalizados onde calculamos a taxa de convergência para estes modelos. Também estendemos o uso destas distribuições para os casos em que a função média era desconhecida, caso típico dos modelos não-paramétricos. A maioria da literatura sobre modelos não-paramétricos e semi-paramétricos trata o problema de aproximação de curvas independentemente de hipóteses sobre a verossimilhança do ruído. No entanto, acreditamos ter mostrado aqui, através de diversos estudos de simulação, que métodos baseados em verossimilhança aliados a outras técnicas de aproximação de funções, tais como splines ou ondaletas, podem ser muito úteis na estimação de curvas, tanto em modelos de regressão quanto em modelos de séries temporais. Mostramos que tais métodos não apenas são úteis nos casos em que o ruído segue uma distribuição com caudas pesadas, como por exemplo a distribuição de Cauchy, que tem variância infinita, como podem unificar métodos consagrados na literatura (mínimos quadrados, estimadores L^1, M-estimador de Huber) em um único algoritmo. Em particular, no caso de ondaletas discutimos a inadequação da transformada discreta de ondaletas aplicada a sinais sob a influência de ruídos com caudas pesadas e mostramos via simulação que podemos usar misturas de normais para aproximar sinais via ondaletas sob tais circunstâncias. Tanto para os modelos univariados, quanto para os modelos parcialmente lineare (multivariados) e tanto para splines quanto para ondaletas, o método sugerido mostrou-se eficaz na presença de valores extremos, mesmo quando a hipótese assumida sobre a distribuição do ruído não correspondia à realidade.

Além disso, dada a natureza das distribuições consideradas, mostramos que existe uma conexão muito forte entre o método "clássico" de estimação e método bayesiano, para o qual derivamos o algoritmo de Gibbs no caso em que a função é aproximada via splines ou ondaletas. Assim como nos demais caso, mostramos via estudo de simulação que o método sugerido consegue captar bem o sinal e estimá-lo (usando o valor esperado ou mediana a posteriori) adequadamente. Generalizamos também o modelo proposto em [35] para modelos mistos para toda a classe de distribuições definidas por misturas de normais através da escala de modo a permitir que tanto o ruído quanto a componente aleatória sejam distribuidos de acordo com um elemento (não necessariamente o mesmo para cada um deles) desta família.

Todos os casos tratados aqui assumiram que o ruído seguia uma distribuição simétrica e que os modelos

eram homocedásticos. Extensões naturais deste trabalho, portanto, incluem o tratamento destes modelos nos casos em que a distribuição do ruído apresenta alguma assimetria e nos casos em que eles apresentam heterocedasticidade. Como notamos anteriormente, algum trabalho já foi feito no caso de modelos paramétricos de volatilidade estocástica e processos tipo GARCH, porém, podemos considerar o uso de misturas assimétricas, como definidas em [7], para a análise destes modelos assim como para a estimação de curvas nos casos em que a função média é desconhecida e para modelos mistos.

Referências Bibliográficas

[1] F. Abramovich, T. Sapatinas, and B. W. Silverman. Wavelet thresholding via a Bayesian approach. *J. R. Statist. Soc. B*, 60(4):725–749, 1998.

[2] R. Adler. *An Introduction to Continuity, Extrema and Related Topics for General Gaussian Processes*. Institute of Mathematical Statistics, 1990.

[3] H. Akaike. Information theory as an extension of the maximum likelihood principle. In B.N. Petrov and F. Csaki, editors, *Second International Symposium on Information Theory*, pages 267–281, Budapest, 1973. Akademiai Kiado.

[4] T. W. Anderson. *The Statistical Analysis or Time Series*. John Wiley & Sons, 1971.

[5] D. R. Andrews and C. L. Mallows. Scale mixtures of normal distributions. *J. R. Statist. Soc. B*, 36:99–102, 1974.

[6] L. Bauwens and M. Lubrano. Bayesian inference on GARCH models using the Gibbs sampler. *Econometrics Journal*, 1:c23–c46, 1998.

[7] M. D. Branco and D. K. Dey. A general class of multivariate skew-elliptical distributions. *J. Multivar. Anal.*, 79(1):99–113, 2001.

[8] K. P. Burnham and D. R. Anderson. *Model Selection and Inference*. Springer-Verlag, New York, 1998.

[9] G. Casella and C. P. Robert. *Monte Carlo Statistical Methods*. Springer-Verlag, New York, second edition, 2004.

[10] J. M. Chambers, C. L. Mallows, and B. W. Stuck. A method for simulating stable random variables. *Journal of American Statistical Association*, 71:340–344, 1976.

[11] S. Chib, F. Nardari, and N. G. Shephard. Markov chain Monte Carlo methods for stochastic volatility models. *Journal of Econometrics*, 108(2):281–316, 2002.

[12] C. de Boor. *A Practical Guide to Splines — Revised Edition*. Springer-Verlag, New York, second edition, 2001.

[13] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39:1–22, 1977.

[14] A. P. Dempster, N. M. Laird, and D. B. Rubin. Iteratively reweighted least squares for linear regression when errors are normal/independent distributed. In P. R. Krishnaiah, editor, *Multivariate Analysis V*, pages 35–57, New York, 1980. Academic Press, Inc.

[15] D. L. Donoho and I. M. Johnstone. Idea spatial adaptation by wavelet shrinkage. *Biometrika*, 81:425–455, 1994.

[16] P. Eilers and B. Marx. Flexible smoothing with b-splines and penalties. *Statistical Science*, 11(2):89–121, 1996.

[17] D. Gamerman. *Markov Chain Monte Carlo — Stochastic Simulation for Bayesian Inference*. CRC Press, Florida, 200.

[18] J. E. Gentle, W. Härdle, and Y. Mori, editors. *Handbook of Computational Statistics*. Springer-Verlag, Berlin, 2004.

[19] S. J. Godsill. Bayesian enhancement of speech and audio signals which can be modelled as ARMA processes. *International Statistical Review*, 65(1):1–21, 1997.

[20] S. J. Godsill and P. J. W. Rayner. Robust treatment of impulsive noise in speech and audio signals. In J. O. Berger, B. Betro, E. Moreno, L. R. Perichi, F. Ruggeri, G. Salinetti, and L. Wasserman, editors, *Bayesian Robustness*, volume 29, pages 331–342, Rimini, Italy, 1995.

[21] S. J. Godsill and P. J. W. Rayner. *Digital Audio Restoration*. Springer, Berlin, 1998.

[22] S. J. Godsill and P. J. W. Rayner. Robust reconstruction and analysis of autoregressive signals in impulsive noise using the Gibbs sampling. *IEEE Trans. on Speech and Audio Processing*, 6(4):352–372, 1998.

[23] P. J. Green. On use of the EM algorithm for penalized likelihood estimation. *J. R. Statist. Soc. B*, 52(3):443–452, 1990.

[24] P. J. Green and B. W. Silverman. *Nonparametric Regression and Generalized Linear Models — A Roughness Penalty Approach*. Chapman & Hall, London, 1994.

[25] P. B. Guest. *Laplace Transforms and an Introduction to Distributions*. Ellis Horwood Limited, 1991.

[26] A. C. Harvey, E. Ruiz, and N. G. Shephard. Multivariate stochastic variance models. *Review of Economic Studies*, 61:247–264, 1994.

[27] E. Jacquier, N. G. Polson, and P. Rossi. Stochastic volatility: Univariate and multivariate extensions. Computing in Economics and Finance 1999 112, Society for Computational Economics, 1999. Available at http://ideas.repec.org/p/sce/scecf9/112.html.

[28] S. N. Lahiri. *Resampling Methods for Dependent Data*. Springer-Verlag, New York, 2003.

[29] K. L. Lange, R. J. A. Little, and J. M. G. Taylor. Robust statistical modeling using the t-distribution. *JASA*, 84:881–896, 1989.

[30] P. A. Morettin. *Ondas e Ondaletas: da Análise de Fourier à Análise de Ondaletas*. EDUSP, 1999.

[31] C. L. Nikias and M. Shao. *Signal Processing with α-Stable Distributions and Applications*. John Wiley & Sons, 1995.

[32] G. Samorodnitsky and M. S. Taqqu. *Stable Non-Gaussian Random Processes — Stochastic Models with Infinite Variance*. Chapman & Hall, New York, 1994.

[33] N. G. Shephard. Local scale models: state space alternative to integrated garch processes. *Journal of Econometrics*, 60:181–202, 1994.

[34] N. G. Shephard. Partial non-Gaussian state space. *Biometrika*, 81:115–131, 1994.

[35] J. Staudenmayer, E. E. Lake, and M. P. Wand. Robustness for general design mixed models using the t-distribution. Submitted to Statistical Modelling, November 2005.

[36] M. West. On scale mixtures of normal distributions. *Biometrika*, 74(6):646–648, 1987.