
Métodos de Monte Carlo Hamiltoniano na inferência
Bayesiana não-paramétrica de valores extremos

Marcelo Hartmann

SERVIÇO DE PÓS-GRADUAÇÃO DO ICMC-USP

Data de Depósito:

Assinatura: _____

Métodos de Monte Carlo Hamiltoniano na inferência Bayesiana não-paramétrica de valores extremos

Marcelo Hartmann

Orientador: Prof. Dr. Ricardo Sanders Ehlers

Dissertação apresentada ao Instituto de Ciências Matemáticas e de Computação - ICMC-USP e ao Departamento de Estatística – DEs-UFSCar, como parte dos requisitos para obtenção do título de Mestre em Estatística – Interinstitucional de Pós-Graduação em Estatística. *VERSÃO REVISADA.*

USP/UFSCar – São Carlos
Março de 2015

**Ficha catalográfica elaborada pelo DePT da
Biblioteca Comunitária da UFSCar**

H333mm Hartmann, Marcelo.
Métodos de Monte Carlo Hamiltoniano na inferência
Bayesiana não-paramétrica de valores extremos / Marcelo
Hartmann. -- São Carlos : UFSCar, 2015.
87 f.

Dissertação (Mestrado) -- Universidade Federal de São
Carlos, 2015.

1. Estatística. 2. Inferência bayesiana. 3. Método de
Monte Carlo. 4. Distribuição valor extremo. I. Título.

CDD: 519.5 (20^a)

MARCELO HARTMANN

MÉTODOS DE MONTE CARLO HAMILTONIANO NA INFERÊNCIA BAYESIANA NÃO-PARAMÉTRICA
DE VALORES EXTREMOS

Dissertação apresentada ao Instituto de Ciências Matemáticas e de Computação - ICMC-USP, como parte dos requisitos para obtenção do título de Mestre em Estatística.

Aprovado em 09 de março de 2015.

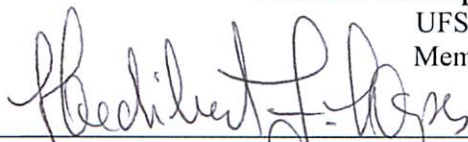
COMISSÃO JULGADORA:



Prof. Dr. Ricardo Sandes Ehlers
ICMC-USP
Presidente



Prof. Dr. Luis Aparecido Milan
UFSCar
Membro



Prof. Dr. Hedibert Freitas Lopes
Insper
Membro

Agradecimentos

O maior objetivo desta dissertação é a discussão que poderá causar. Jamais a fiz por reconhecimento. Faço pelo compartilhamento do conhecimento gerado e a evolução de idéias que precisam ser discutidas para alcançar uma definição precisa de seus conceitos.

Muitas pessoas fazem parte desta dissertação. Certamente a maior parte delas indiretamente. Agradeço de modo incondicional meus pais, Ana e Luiz Roberto. Foi somente através deles que cheguei aqui. Serei eternamente grato por isso.

Agradeço ao meu irmão Junior, à Carla e à Julia. Por me mostrarem que a vida irá continuar não importe o que aconteça. Agradeço a minha irmã Ingrid e ao Markus, por mostrarem que sonhos são ainda melhores quando compartilhados.

Por bandas como, Ashes of Pompeii, The Wonder Years, Explosions in the Sky, This Will Destroy You, God is an Astronaut, Sigur Rós, Pink Floyd, Front Porch Step, Old Gray, Sienna Skies, Staind, Alice in Chains, The Citizen, The Story So Far, Neil Young, Fit For Rivals e muitas outras que ajudaram na sua maneira o decorrer desta dissertação.

Ao meu orientador Ricardo S. Ehlers, por me deixar livre para buscar aquilo que desejei. Pelas breves discussões e pela orientação durante esses anos no mestrado. Mais ainda pela filosofia free-software e o manifesto GNU.

Now we know that technical advance
By itself creates more problems than it solves
Everything changes and so do we
So we have to do. So we have to
Throw the dead into the sea, the future is far away
The end justifies the means, focus on the progress
The ocean brought the silence back, everyone is gone
Nothing will ever change

Gunkanjima, *Ashes of Pompeii*

Resumo

Neste trabalho propomos uma abordagem Bayesiana não-paramétrica para a modelagem de dados com comportamento extremo. Tratamos o parâmetro de localização μ da distribuição generalizada de valor extremo como uma função aleatória e assumimos um processo Gaussiano para tal função (Rasmussen & Williams 2006). Esta situação leva à intratabilidade analítica da distribuição a posteriori de alta dimensão. Para lidar com este problema fazemos uso do método Hamiltoniano de Monte Carlo em variedade Riemanniana que permite a simulação de valores da distribuição a posteriori com forma complexa e estrutura de correlação incomum (Calderhead & Girolami 2011). Além disso, propomos um modelo de série temporal autoregressivo de ordem p , assumindo a distribuição generalizada de valor extremo para o ruído e determinamos a respectiva matriz de informação de Fisher. No decorrer de todo o trabalho, estudamos a qualidade do algoritmo em suas variantes através de simulações computacionais e apresentamos vários exemplos com dados reais e simulados.

Palavras-chaves: Inferência Bayesiana não-paramétrica, processo Gaussiano latente, distribuição generalizada de valor extremo, método de Monte Carlo Hamiltoniano em variedade Riemanniana.

Abstract

In this work we propose a Bayesian nonparametric approach for modeling extreme value data. We treat the location parameter μ of the generalized extreme value distribution as a random function following a Gaussian process model (Rasmussen & Williams 2006). This configuration leads to no closed-form expressions for the high-dimensional posterior distribution. To tackle this problem we use the Riemannian Manifold Hamiltonian Monte Carlo algorithm which allows samples from the posterior distribution with complex form and non-usual correlation structure (Calderhead & Girolami 2011). Moreover, we propose an autoregressive time series model assuming the generalized extreme value distribution for the noise and obtained its Fisher information matrix. Throughout this work we employ some computational simulation studies to assess the performance of the algorithm in its variants and show many examples with simulated and real data-sets.

Keywords: Bayesian nonparametrics, latent Gaussian process, generalized extreme values distribution, Riemannian manifold Hamiltonian Monte Carlo algorithm.

Lista de Figuras

2.1	Movimento do disco na superfície.	17
2.2	Histograma, série temporal e autocorrelação para os dados de nível do mar em Port-Pirie, Australia.	22
2.3	Gráficos da série de tempo e autocorrelação para os valores gerados pelo algoritmo HMC.	23
2.4	Gráficos da série de tempo e autocorrelação para os valores gerados pelo algoritmo MH.	24
3.1	Série temporal e gráficos de autocorrelação e autocorrelação parcial.	42
3.2	Gráficos da série de tempo e autocorrelação para os valores gerados pelo algoritmo RMHMC.	43
3.3	Valores preditos marcados com 'x'. Valores observados ponto cheio. Intervalo de credibilidade 95% - barra horizontal.	44
3.4	Histogramas e gráficos de dispersão respectivamente para μ, θ, σ, ξ	44
4.1	Realizações do processo Gaussiano. Linhas tracejadas são funções amostrais do processo. Linha cheia é a média do processo.	48
4.2	Realizações do processo Gaussiano com função exponencial quadrática. $\phi_1 = 1$ e $\phi_2 = 2$	49
4.3	Realizações do processo Gaussiano com função periódica e parâmetros $\phi_1 = 1$, $\phi_2 = 2$ e $\gamma = 1$	50
4.4	Pontos amostrados.	54
4.5	Histogramas (distribuições marginais) e gráficos de dispersão (distribuições marginais bivariadas) para os hiperparâmetros $\sigma^2, \phi_1, \phi_2, \phi_3, \phi_4, \phi_5$ respectivamente.	57
4.6	Estimativa (média) da valor da função nas abscissas simuladas	58
4.7	predição (média) da função desconhecida.	58

4.8	Pontos amostrados.	62
4.9	Histogramas (distribuições marginais) e gráficos de dispersão para os valores das função, $f(x_1) D$, $f(x_2) D$, $f(x_{99}) D$ e $f(x_{100}) D$	64
4.10	Histogramas (distribuições marginais) e gráficos de dispersão dos hiperparâmetros na reta	65
4.11	Intervalo de previsão $I_* = (10, 22)$. Média da distribuição preditiva $\mathbf{y}_* D$ à esquerda com linha tracejada. Moda da distribuição preditiva $\mathbf{y}_* D$ à direita com linha tracejada. Função verdadeira com linhas sólidas.	66
4.12	Intervalo previsão para os dados faltantes $I_* = (3, 7)$. Média da distribuição predi- tiva $\mathbf{y}_* D$ à esquerda com linha tracejada. Moda da distribuição preditiva $\mathbf{y}_* D$ à direita com linha tracejada. Função verdadeira com linhas sólidas.	66
4.13	Temperatura máxima na estação Rothera, Antártica.	67
4.14	Valores gerados da posteriori para os hiperparâmetros na reta.	69
4.15	Previsões com média e moda a posteriori da distribuição de $\mathbf{y}_* D$. A Linha cheia representa a média a posteriori, linha tracejada representa a moda a posteriori. Os pontos observados são representados pelas cruces. O intervalo de credibilidade 95% é representado pela região cinza.	70

Lista de Tabelas

2.1	Resultados da simulação do estudo 1 e estudo 2 para cada parâmetro da distribuição GEV com tamanhos amostrais variados. A primeira linha de cada parâmetro corresponde as cadeias de 20000 valores. A linha abaixo corresponde as cadeias de 1100 valores.	25
3.1	Resultados da simulação	40
3.2	Aproximações para média, desvio-padrão, mediana e intervalo de credibilidade a posteriori	43
4.1	Resumos de interesse dos hiperparâmetros a posteriori (4.19)	56
4.2	Resumos de interesse dos hiperparâmetros a posteriori (4.38)	64
4.3	Resumos de interesse dos hiperparâmetros a posteriori (4.46)	70

Conteúdo

1	Introdução	7
1.1	Motivação e eventos extremos	7
1.2	Modelos paramétricos e não-paramétricos	9
1.3	Abordagem Bayesiana	10
1.4	Método de Monte Carlo via Cadeias de Markov	12
1.5	O algoritmo de Metropolis-Hastings (MH)	14
2	Monte Carlo Hamiltoniano	16
2.1	Dinâmica Hamiltoniana	17
2.2	O algoritmo de Monte Carlo Hamiltoniano (HMC)	19
2.3	Estudos de simulação	23
2.4	Monte Carlo Hamiltoniano em variedade Riemanniana (RMHMC)	27
3	Teoria de valores extremos	31
3.1	Fundamentos básicos	31
3.2	Autoregressão temporal em valores extremos	33
3.3	Modelo autoregressivo-GEV	35
3.4	Matriz informação de Fisher	36
3.5	Estudo de simulação	37
3.6	Aplicação em dados reais	41
4	Inferência Bayesiana não-paramétrica via processo Gaussiano	45
4.1	Processo Gaussiano	46
4.2	Funções de covariância	47
4.3	Processo Gaussiano como representação a priori para $f(\cdot)$	50
4.4	Inferência Bayesiana não-paramétrica em valores extremos	59

4.5	Exemplo simulado com dados faltantes	61
4.6	Aplicação em dados reais	67
5	Conclusão	71
5.1	Perspectivas Futuras	72
A	Fórmulas Matriciais	80
A.1	Derivadas matriciais	80
A.2	Fórmula de Sherman-Morrison-Woodbury-Schur	80
A.3	Matriz informação de Fisher para o modelo autoregressivo-GEV	80
A.4	Matriz de Informação de Fisher para o processo Gaussiano latente	85

1

Introdução

1.1 Motivação e eventos extremos

Fenômenos naturais como vazões de rios, velocidades de ventos, índices pluviométricos dentre vários outros, estão sujeitos a níveis extremamente baixos ou extremamente altos que podem implicar em grandes perdas financeiras. Mercados financeiros aonde o aporte de grandes somas em investimentos pode ter um impacto na economia de um país precisam ter seus riscos de grandes perdas ou grandes ganhos quantificados. Em análise de risco, estimar perdas futuras através da modelagem de eventos associados a inadimplência é de fundamental importância. No campo dos seguros, o risco potencial de sinistros de alto valor precisa ser quantificado e associado a possíveis eventos catastróficos devido às grandes quantias envolvidas em indenizações.

Duas à três décadas atrás o estudo de processos aleatórios focou-se no comportamento médio e na variabilidade 'normal' de sistemas ambientais, sociais, financeiros, biológicos, etc. Recentemente o estudo de valores extremos é uma das mais importantes disciplinas estatísticas e suas técnicas se tornaram ferramentas fundamentais em várias outras áreas da ciência aplicada, veja Coles (2004), Castillo et al. (2004) ou Ghil et al. (2011) para vários exemplos.

A teoria de valores extremos estuda processos aleatórios que se manifestam em níveis muito altos ou baixos, no entanto, quando trabalha-se em problemas aplicados, eventos extremos são raros (escassos), e invalidam teoremas baseados em tamanho de amostra grande. Na prática isto torna difícil a inferência em modelos de valores extremos. Uma boa alternativa é realizar uma abordagem Bayesiana não-paramétrica

aliviando suposições muito restritivas em aproximações assintóticas.

Neste trabalho, propomos a modelagem de séries de valores extremos modelando o parâmetro de locação da distribuição generalizada de valor extremo como uma função aleatória, i.e., uma vez que a abordagem é Bayesiana associamos um processo estocástico Gaussiano a priori ao invés de assumir modelos paramétricos. Coles (2004), introduz a idéia de inferência em sequências não estacionárias de dados utilizando a distribuição de valor extremo. Mais especificamente, o modelo proposto é,

$$Z_t \sim GEV(\mu(t), \sigma, \xi), \quad (1.1)$$

sendo que $GEV(\mu, \sigma, \xi)$ denota a distribuição de valor extremo generalizada com parâmetros de locação $\mu \in \mathbb{R}$, escala $\sigma > 0$ e forma $\xi \in \mathbb{R}$, $\mu(t)$ é a função média do processo. A suposição usual é de que $\mu(t)$ segue um modelo linear geral, por exemplo $\mu(t) = \beta_0 + \beta_1 t + \beta_2 t^2$ ou mesmo um modelo com mudança de regime, por exemplo $\mu(t) = \mu_1$ para $t \leq t_0$ ou $\mu(t) = \mu_2$ para $t > t_0$.

Ao invés de assumir uma forma funcional específica para a função média propomos utilizar um processo Gaussiano latente, i.e.

$$\mu(t) \sim PG(m(t), k(t_1, t_2)), \quad (1.2)$$

sendo que $m(\cdot)$ e $k(\cdot, \cdot)$ descrevem a média e a estrutura de covariância do processo latente. Deste modo não se impõe uma única forma funcional paramétrica ao processo latente.

Autores como Rasmussem & Williams (2006), assumem uma função de ligação estocástica para problemas de classificação. Jylanki et al. (2011) propõem um modelo de regressão com distribuição observacional t-Student na presença de outliers, modelando a média como uma função estocástica. Calderhead (2012), estuda previsões futuras num sistema de equações diferenciais para processos biológicos de crescimento populacional e Chkrebtii (2013) analisa o sistema caótico de Lorenz, ambos baseados em processo estocástico Gaussiano latente com abordagem Bayesiana. Brahim-Belhouari & Bermak (2004) predizem valores futuros com diferentes funções de covariâncias para séries temporais não-estacionárias.

Nas próximas seções descrevemos as principais ferramentas propostas no trabalho e as dificuldades analíticas existentes na inferência a posteriori sobre modelos com processo estocástico latente. No Capítulo 2, apresentamos o método de Monte Carlo

Hamiltoniano e sua variante, que é atualmente um algoritmo eficiente para geração de valores aleatórios em modelos complexos de alta dimensão. No Capítulo 3, apresentamos a distribuição generalizada de valor extremo e suas principais propriedades. Além disso, discutimos a modelagem paramétrica de dados com características de extremos e apresentamos um modelo de série temporal autoregressivo com distribuição generalizada de valor extremo para o ruído. No capítulo 4, introduzimos o modelo Bayesiano não-paramétrico sobre uma função desconhecida baseado no processo Gaussiano e aplicamos esta abordagem em modelos de valores extremos. No capítulo 5, apresentamos as conclusões e proposta de trabalhos futuros.

1.2 Modelos paramétricos e não-paramétricos

Métodos estatísticos tradicionais são comumente baseados em suposições paramétricas, i.e., assumem uma forma funcional dependente de parâmetros (e.g. modelos de regressão) juntamente a um modelo probabilístico subjacente aos dados tal como, Normal, Poisson, Weibull, Exponencial, Gama, etc. Por exemplo, a análise de regressão em modelos lineares generalizados fundamentada em suposições paramétricas, é a mais utilizada em todas as áreas no meio científico e profissional, e a suposição da forma paramétrica para a variável resposta é chave para o estudo ideal do experimento.

No entanto, em várias situações a pressuposição de qualquer forma funcional para o estudo do comportamento de dados observados não é tarefa fácil. Algumas características de assimetria, caudas pesadas, multimodalidade, sensibilidade a outliers, dentre outras, podem dificultar o bom ajustamento dos dados ao modelo probabilístico proposto.

Modelos Bayesianos não-paramétricos são motivados pela imposição menos restritiva sobre suposições paramétricas. Por exemplo, suponha que $X_i \stackrel{i.i.d}{\sim} F$. A abordagem Bayesiana não-paramétrica assume uma distribuição a priori sobre F , i.e., um processo estocástico representando o processo gerador das funções densidade. Neste caso é comum assumir um processo de Dirichlet (ver por exemplo Ferguson 1973 e Müller & Quintana 2004). Em um problema de regressão não-linear é possível supor uma distribuição a priori sobre funções contínuas, i.e., um processo estocástico Gaussiano, veja Rasmussem & Williams (2006).

Em ambos os exemplos acima, priori e posteriori são processos estocástico e que podem ser colocados como modelos de dimensões infinitas (veja Schervish 2011). Em problemas práticos de inferência, a complexidade deste tipo de modelo se torna evidente uma vez que o número de parâmetros aumenta com o tamanho amostral. Com isso métodos computacionalmente intensivos e eficientes são fundamentais para uma boa performance dos modelos, por exemplo o método de Monte Carlo Hamiltoniano e mais atualmente o método de Monte Carlo Hamiltoniano em variedade Riemanniana que serão apresentados nos próximos capítulos. Para uma discussão profunda sobre estes assuntos, veja Bernardo & Smith (1994), Ghosh & Ramamoorthi (2003), Müller & Quintana (2004), Teh & Orbanz (2010), Calderhead & Girolami (2011) e Neal (2011).

1.3 Abordagem Bayesiana

Quando lidamos com problemas estatísticos ou qualquer tipo de experimentos científicos, trabalhamos com dados observados, mensurações de aparelhos, condições adversas e várias variáveis que não podemos controlar. Deste modo torna-se natural pensar que é impraticável obter certeza absoluta a partir de mensurações. Medidas são realizadas para obter informação quantitativa/qualitativa e utilizadas para descrever, explicar ou elucidar fenômenos inerentes a processos aleatórios.

Uma maneira formal para analisar dados é através da inferência científica. Entende-se como inferência científica o processo de aprendizado de alguma característica de interesse com base em observações de dados (medidas ou amostra) que estão vinculados a tal característica (Bonassi (2009) apud Johnson & Kotz). É importante ressaltar que tal característica, em estatística, é sempre desconhecida na prática.

Se a interdependência entre a amostra e a característica de interesse é dada por uma função de probabilidade, temos o processo chamado de inferência estatística. Neste caso é comum assumir um modelo probabilístico gerador dos dados indexado por um vetor paramétrico θ fixo e desconhecido. A tratabilidade de θ como uma constante baseia-se em resultados assintóticos e não permite que sejam incorporadas informações externas sobre seu comportamento.

O teorema de Bayes trata todas características de interesse desconhecidas como

variáveis aleatórias e permite que informação externa seja incorporada a análise dos dados. Por muitas vezes algum tipo de conhecimento sobre $\boldsymbol{\theta}$ é de grau extremamente representativo sendo necessário incorporar tal conhecimento de forma consistente à análise dos dados.

A inferência é Bayesiana quando a abordagem é via o teorema de Bayes (1763) dado por

$$\pi(\boldsymbol{\theta}|D) = \frac{L(D|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\pi(D)} \quad (1.3)$$

com

$$\pi(D) = \int_{\Theta} L(D|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}$$

em que $L(D|\boldsymbol{\theta}) = f_{\mathbf{X}}(\mathbf{x}|\boldsymbol{\theta})$ é a função de verossimilhança. $D = \mathbf{x}$ é a amostra aleatória observada seguindo a f.d.p. $f_{\mathbf{X}}(\cdot)$, $\boldsymbol{\theta}$ é o vetor paramétrico (características de interesse), $\pi(\boldsymbol{\theta})$ a função densidade de probabilidade a priori de $\boldsymbol{\theta}$. $\pi(\boldsymbol{\theta}|D)$ a função densidade de probabilidade a posteriori de $\boldsymbol{\theta}$ e $\pi(D)$ é a constante normalizadora que também pode ser colocada como a probabilidade marginal dos dados ou ainda a distribuição preditiva a priori. Θ é o conjunto de possíveis valores de $\boldsymbol{\theta}$ (espaço paramétrico).

A expressão acima é apenas a probabilidade condicional de $\boldsymbol{\theta}$ dado os dados. $\pi(\boldsymbol{\theta})$ representa o conhecimento atual que se possui sobre $\boldsymbol{\theta}$ antes de qualquer tipo de informação relacionada às características de interesse. A escolha de $\pi(\boldsymbol{\theta})$ pode ou não ter grande influência sobre $\pi(\boldsymbol{\theta}|D)$ e tem sido alvo de muita discussão no meio científico. Principalmente devido a escolha de distribuições a priori impróprias que geram distribuições a posteriori próprias (veja Kass & Wasserman 1996).

Existem também métodos formais de extração do conhecimento sob a forma de uma distribuição de probabilidade que são conhecidos como métodos de elicitación de distribuições a priori. Para detalhes veja Garthwaith et al. (2005).

Inferências sobre os parâmetros do modelo são obtidas através do cálculo de esperanças a posteriori, i.e., precisamos determinar integrais do tipo,

$$E_{\boldsymbol{\theta}|D}[g(\boldsymbol{\theta})] = \int_{\Theta} g(\boldsymbol{\theta})\pi(\boldsymbol{\theta}|D)d\boldsymbol{\theta}. \quad (1.4)$$

Na maioria das aplicações práticas a dimensão elevada e analiticamente complexa de (1.4) leva ao cálculo de integrais multivariadas que não são de fácil determinação.

Metropolis & Ulam (1949) propuseram o método de Monte Carlo baseado em aproximações estocásticas para integrais multivariadas, motivados pelo cálculo de coeficientes de partículas sub-atômicas em bombas atômicas. O método é baseado na Lei Forte dos Grandes números e garante que, se possuímos uma amostra aleatória de $\pi(\boldsymbol{\theta}|D)$ podemos aproximar o valor de (1.4) na seguinte forma,

$$\begin{aligned} E_{\boldsymbol{\theta}|D}[\widehat{g(\boldsymbol{\theta})}] &= \frac{1}{N} \sum_{i=1}^N g(\boldsymbol{\theta}^{(i)}) \\ &\xrightarrow{q.c.} E_{\boldsymbol{\theta}|D}[g(\boldsymbol{\theta})], \end{aligned} \tag{1.5}$$

em que $g : \Theta \rightarrow \mathbb{R}$.

No entanto, obter valores de $\pi(\boldsymbol{\theta}|D)$ em dimensões elevadas também não é tarefa fácil. Vários métodos estão propostos na literatura e podem ser revisados em Robert & Casella (2004). O método de Monte Carlo via Cadeias de Markov (MCMC) proposto por Metropolis et al. (1953) e generalizado por Hastings (1970) é dominante em aplicações Bayesianas para a simulação de valores da distribuição a posteriori.

Mais recentemente, Duane et al. (1987) propôs o método de Monte Carlo Híbrido, ou melhor denominado Monte Carlo Hamiltoniano para simulação de um sistema molecular através da dinâmica Hamiltoniana, sendo mais tarde reformulado por Neal (1995) em aplicações de redes Bayesianas. Calderhead & Girolami (2011) estenderam o método de Monte Carlo Hamiltoniano usando conceitos de geometria diferencial na distribuição a posteriori.

1.4 Método de Monte Carlo via Cadeias de Markov

Nesta seção apresentamos algumas definições e teoremas sobre cadeias de Markov utilizadas no restante do trabalho. O algoritmo de Metropolis-Hastings (MH) também é apresentado com maiores detalhes pois é um simulador geral de valores aleatórios. Algumas referências sobre processos Markovianos podem ser consultadas em Schinazi (1999) e Robert & Casella (2004).

Definição 1.1 (Processo estocástico) *Um processo estocástico é uma função de dois argumentos $X(t, \omega) : T \times \Omega \rightarrow R$, tal que para $t^* \in T$ fixo, $X(t^*, \omega)$ é uma*

variável aleatória e para $\omega^* \in \Omega$ fixo, $X(t, \omega^*)$ é uma realização do processo (uma função determinística). T é um subconjunto dos reais.

Dada uma coleção de tempos $\mathbf{t} = [t_1, \dots, t_n]$, $\mathbf{X} = [X_{t_1}, \dots, X_{t_n}]$ segue uma função densidade de probabilidade conjunta,

$$\mathbf{X} \sim p(\mathbf{x}), \quad \forall t \in T \text{ e } \forall n \in N$$

em que, $\mathbf{x} = [x_{t_1}, \dots, x_{t_n}]$ e $X(t_j) = X_{t_j}$.

Definição 1.2 (Cadeia de Markov) *Um processo estocástico é dito ser uma cadeia de Markov se,*

$$p(x_{t_j} | x_{t_{j-1}}, \dots, x_{t_1}) = p(x_{t_j} | x_{t_{j-1}}), \quad \forall t_j \in T \text{ e } \forall j \in N \quad (1.6)$$

ou seja, a função densidade de probabilidade conjunta pode ser escrita como,

$$p(\mathbf{x}) = p(x_{t_1}) \prod_{i=2}^n p(x_{t_i} | x_{t_{i-1}}) \quad (1.7)$$

Para o uso adequado do método de Monte Carlo, precisamos que a cadeia Markoviana gerada satisfaça certas condições. Essas condições fazem com que independente do valor inicial da cadeia, no limite de $t \rightarrow \infty$, X_t será amostrado de uma distribuição invariante.

Definição 1.3 (Probabilidade de transição) *Para todo $x \in S$ e $A \subseteq S$ definimos $T^n(x, A)$ como a probabilidade condicional da cadeia se encontrar numa região A depois de n passos dado o início em x , i.e., $T^n(x, A) = p(X_n \in A | X_0 = x)$ e $T^n(x, y)$ a função densidade de probabilidade condicional, i.e., $T^n(x, y) = p_{X_n}(y | X_0 = x)$.*

Definição 1.4 (Distribuição invariante) *A cadeia de Markov com probabilidade de transição $T(x, A)$ possui uma distribuição invariante $\pi(x)$ se, para todo conjunto $A \subset S$,*

$$\int_A \pi(x) dx = \int T(x, A) \pi(x) dx \quad (1.8)$$

Definição 1.5 (Equação de balanço detalhada) *Dizemos que a probabilidade de transição $T(x, y)$ satisfaz a equação da balanço detalhada com respeito a densidade $\pi(x)$ se, para todo x e $y \in S$ vale,*

$$T(x, y) \pi(x) = T(y, x) \pi(y) \quad (1.9)$$

Ou seja, depois de um tempo suficientemente grande, a taxa com que a cadeia passa de x para y é mesma que passa de y para x .

1.5 O algoritmo de Metropolis-Hastings (MH)

Em situações que a amostragem direta de valores da distribuição a posteriori não é tarefa fácil, Hastings (1970) desenvolveu um algoritmo geral para simulação de valores aleatórios de uma distribuição alvo por meio de uma distribuição auxiliar, tal que, gerar valores desta distribuição é tarefa fácil.

Suponha que gostaríamos de gerar valores de uma distribuição qualquer $p(\mathbf{x})$. Assuma uma distribuição auxiliar $q(\mathbf{x})$ no mesmo domínio de $p(\mathbf{x})$, tal que gerar de $q(\mathbf{x})$ é tarefa fácil. O algoritmo de Metropolis-Hastings é,

(i) Tome $\mathbf{X}_0 = \mathbf{x}^{(0)}$ gerado da distribuição auxiliar $q(\mathbf{x})$.

(ii) Para $n = 1, 2, \dots, N$

- Faça $\mathbf{X}_{n-1} = \mathbf{x}^{(n-1)}$
- Gere \mathbf{y} à partir de $q(\mathbf{y}|\mathbf{x}^{(n-1)})$ e $u \sim U[0, 1]$
- Calcule $\alpha(\mathbf{x}^{(n-1)}, \mathbf{y}) = \min \left\{ 1, \frac{p(\mathbf{y})q(\mathbf{x}^{(n-1)}|\mathbf{y})}{p(\mathbf{x})q(\mathbf{y}|\mathbf{x}^{(n-1)})} \right\}$
- Faça $\mathbf{X}_n = \mathbf{x}^{(n)} = \begin{cases} \mathbf{y}, & \text{com probabilidade } \alpha(\mathbf{x}^{(n-1)}, \mathbf{y}) > u \\ \mathbf{x}^{(n-1)}, & \text{caso contrário} \end{cases}$

Note que algoritmo MH faz a transição $\mathbf{x} \rightarrow \mathbf{y}$ de acordo com $q(\cdot)$ e aceita este valor com probabilidade $\alpha(\mathbf{x}, \mathbf{y})$, i.e., a probabilidade de transição fica dada por,

$$T(\mathbf{x}, \mathbf{y}) = q(\mathbf{y}|\mathbf{x})\alpha(\mathbf{x}, \mathbf{y}). \quad (1.10)$$

Veja que a expressão acima satisfaz as equações de balanço detalhada,

$$\begin{aligned} T(\mathbf{x}, \mathbf{y})p(\mathbf{x}) &= q(\mathbf{y}|\mathbf{x}) \min \left\{ 1, \frac{p(\mathbf{y})q(\mathbf{x}|\mathbf{y})}{p(\mathbf{x})q(\mathbf{y}|\mathbf{x})} \right\} p(\mathbf{x}) \\ &= \min \{p(\mathbf{y})q(\mathbf{x}|\mathbf{y}), p(\mathbf{x})q(\mathbf{y}|\mathbf{x})\} \\ &= T(\mathbf{y}, \mathbf{x})p(\mathbf{y}) \end{aligned} \quad (1.11)$$

Chib & Greenberg (1995) mostram que, quando as condições de balanço são satisfeitas em relação a uma medida de probabilidade p então a definição (1.4) vale, em relação a p , ou seja, o algoritmo de Metropolis-Hastings gera uma cadeia de Markov com distribuição invariante p . A definição (1.5) é uma condição suficiente. Essa

condição garante, estacionariedade, irredutibilidade e aperiodicidade da cadeia Markoviana.

Uma variante do algoritmo é realizar o procedimento em cada argumento da distribuição alvo, i.e., podemos aplicar o algoritmo para o primeiro argumento, digamos x_1 , e aceitamos ou rejeitamos um novo valor de acordo com a razão de Metropolis fixando todos os outros argumentos e repetindo o procedimento de modo similar para todos os outros argumentos.

Quando a distribuição alvo é de grande dimensão com estrutura de correlação não usual, a performance ideal do algoritmo se torna algo de difícil alcance. O algoritmo dificilmente irá percorrer todo o espaço do vetor aleatório devido a forma complexa da distribuição alvo. Além disso, a cadeia carregará dependência excessiva de valores anteriores gerados demorando para a convergência na distribuição estacionária.

2

Monte Carlo Hamiltoniano

O método de Monte Carlo Hamiltoniano (HMC) é um técnica baseada na dinâmica Hamiltoniana concomitante a regra de transição de Metropolis. Produz um modo eficiente de simulação de valores aleatórios de uma distribuição alvo, percorre rapidamente o suporte desta não se fazendo uso de distribuições auxiliares. É uma construção mais sistemática e direta na configuração de um algoritmo mais próximo do ideal para a amostragem de valores de distribuições com dimensão relativamente alta e forte estrutura de correlação.

O algoritmo HMC foi originalmente proposto por Duane et al. (1987) na simulação dinâmica de sistemas moleculares e primeiramente denominado Monte Carlo híbrido. Posteriormente, Neal (1995) introduziu o método em aplicações estatísticas de redes neurais Bayesianas. Mackay (2003), Neal (2011) e Lan (2013) mostram propriedades importantes deste método e provam sua convergência para alguma distribuição alvo de interesse.

A essência do método é simular o movimento de uma partícula se deslocando sob uma energia potencial igual ao logaritmo negativo da densidade de probabilidade alvo. A cada iteração a velocidade da partícula é aleatorizada simulando seu movimento por algum tempo. Ao final, obtemos sua nova posição, que será o novo valor proposto da distribuição alvo, aceitando-o ou não de acordo com a regra de Metropolis.

2.1 Dinâmica Hamiltoniana

As equações de Hamilton formam um conjunto de equações diferenciais ordinárias vistas como uma reformulação das Leis de Newton na mecânica clássica. Descrevem de modo determinístico como partículas (ou corpos) evoluem no tempo em um sistema fechado e conservativo.

Considere um disco deslizando sem atrito sobre uma superfície de altura variável. Suponha que a posição do disco, digamos θ , é algum valor específico do espaço de estado $S = \mathbb{R}^p$. Considere também uma velocidade inicial relacionada com um momento $\mathbf{p} \in \mathbb{R}^d$ ($\mathbf{p} = m\mathbf{v}$, \mathbf{v} é a velocidade).

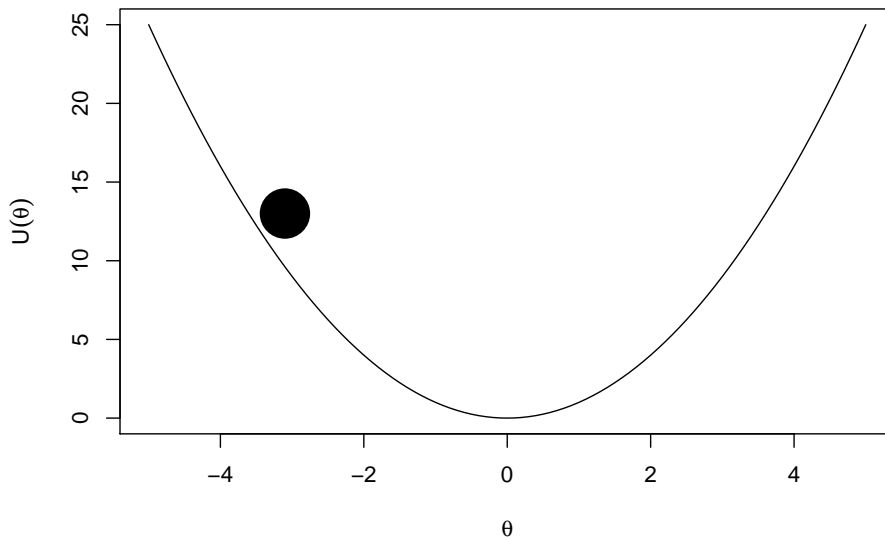


Figura 2.1: Movimento do disco na superfície.

Com o movimento do disco sobre a superfície, a energia potencial do disco $U(\theta)$ e a energia cinética $C(\mathbf{p}) = \mathbf{p}^t \mathbf{p} / 2m$ (m é a massa do disco) variam de acordo com sua posição sobre a superfície. Se o disco tem um movimento ascendente sua energia potencial aumenta e sua energia cinética diminui até um ponto em que o disco atinge energia cinética nula. Neste momento, o disco inicia um movimento descendente passando por uma energia cinética máxima, potencial mínima e tomando o movimento ascendente novamente.

A energia total constante deste sistema é representada por uma função chamada

Hamiltoniana que é definida abaixo.

Definição 2.1 (Hamiltoniano) *A energia total de um sistema fechado e conservativo é dada pela função Hamiltoniana,*

$$\begin{aligned} H(\boldsymbol{\theta}, \mathbf{p}) &= U(\boldsymbol{\theta}) + C(\mathbf{p}) \\ &= U(\boldsymbol{\theta}) + \mathbf{p}^t M^{-1} \mathbf{p} \end{aligned} \quad (2.1)$$

em que $U(\boldsymbol{\theta})$ é a energia potencial e $C(\mathbf{p})$ a energia cinética do sistema.

A evolução determinística de partículas no decorrer do tempo é dada pela solução da dinâmica Hamiltoniana. A definição de dinâmica Hamiltoniana se descreve na forma,

Definição 2.2 (Dinâmica Hamiltoniana) *Dado que a energia total de um sistema é representada por $H(\boldsymbol{\theta}, \mathbf{p})$, a dinâmica Hamiltoniana é definida pelo sistema de equações diferenciais,*

$$\begin{aligned} \frac{d\boldsymbol{\theta}}{dt} &= \frac{\partial H}{\partial \mathbf{p}} = \nabla_{\mathbf{p}} C(\mathbf{p}) \\ \frac{d\mathbf{p}}{dt} &= -\frac{\partial H}{\partial \boldsymbol{\theta}} = \nabla_{\boldsymbol{\theta}} U(\boldsymbol{\theta}) \end{aligned} \quad (2.2)$$

Suponha agora que desejamos determinar a posição e a velocidade de uma partícula após a evolução do sistema para uma época $t_1 = t_0 + \tau$. A solução destas equações fixados valores iniciais $[\boldsymbol{\theta}(t_0) \ \mathbf{p}(t_0)]$ fornece a posição e a velocidade da partícula no tempo t_1 , $[\boldsymbol{\theta}(t_1) \ \mathbf{p}(t_1)]$.

No entanto, soluções analíticas para as equações Hamiltonianas são dificilmente obtidas sendo necessário empregar uma classe especial de método numérico para a solução aproximada do sistema de equações diferenciais. O método de Störmer-Verlet (leapfrog) é um método ideal para sistemas conservativos e tem propriedades fundamentais na garantia da convergência da cadeia de Markov para a distribuição estacionária de interesse, são elas,

- (i) conservação da energia total, i.e., $H(\boldsymbol{\theta}(\tau), \mathbf{p}(\tau)) = H(\boldsymbol{\theta}(0), \mathbf{p}(0))$
- (ii) é reversível no tempo, i.e, o método numérico garante que, se partimos de um ponto inicial $(\boldsymbol{\theta}_0, \mathbf{p}_0)$ e chegamos em $(\boldsymbol{\theta}_1, \mathbf{p}_1)$, tomando o ponto final com o momento negativo, $-\mathbf{p}_1$ voltamos ao ponto $(\boldsymbol{\theta}_0, -\mathbf{p}_0)$.

(iii) preserva o volume infinitesimal.

Uma revisão extensiva sobre este método pode ser consultada em Leimkuhler & Reich 2005.

2.2 O algoritmo de Monte Carlo Hamiltoniano (HMC)

Baseada na Definição 2.1, assumamos que a energia potencial do sistema Hamiltoniano é dada pelo logaritmo negativo da densidade a posteriori que gostaríamos de simular valores, i.e.,

$$U(\boldsymbol{\theta}) = -\ell(\boldsymbol{\theta}) = -\log[L(\boldsymbol{\theta}|D)] - \log[\pi(\boldsymbol{\theta})]. \quad (2.3)$$

Suponha também que o momento \mathbf{p} na função Hamiltoniana é um vetor aleatório no mesmo domínio de $\boldsymbol{\theta}$, i.e., \mathbb{R}^d . Note então que a função densidade conjunta não normalizada de $(\boldsymbol{\theta}, \mathbf{p})$ fica definida pela exponencial negativa da função Hamiltoniana, uma vez que a função $C(\mathbf{p})$ (energia cinética) tem a forma do núcleo de uma normal multivariada $N_d[\mathbf{0}, M]$,

$$\begin{aligned} f(\boldsymbol{\theta}, \mathbf{p}) &\propto \exp[-H(\boldsymbol{\theta}, \mathbf{p})] \\ &\propto L(\boldsymbol{\theta}|D)\pi(\boldsymbol{\theta}) \exp[-\mathbf{p}^t M^{-1} \mathbf{p}]. \end{aligned} \quad (2.4)$$

Ainda, note que (2.4) é uma função de densidade conjunta fatorável. Pela definição (2.2) podemos obter valores de $\pi(\boldsymbol{\theta}|D)$ uma vez que a energia potencial está definida pela distribuição a posteriori. Isto é, para valores iniciais $(\boldsymbol{\theta}^{(\tau)}, \mathbf{p}^{(\tau)})$ com tempo inicial fictício $\tau = 0$, aplique a solução numérica de Störmer-Verlet na dinâmica Hamiltoniana (veja Mackay 2003),

$$\begin{aligned} \mathbf{p}^{(\tau+\varepsilon/2)} &= \mathbf{p}^{(\tau)} + (\varepsilon/2)\nabla_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta}^{(\tau)}) \\ \boldsymbol{\theta}^{(\tau+\varepsilon)} &= \boldsymbol{\theta}^{(\tau)} + \varepsilon\nabla_{\mathbf{p}}C(\mathbf{p}^{(\tau+\varepsilon/2)}) \\ \mathbf{p}^{(\tau+\varepsilon)} &= \mathbf{p}^{(\tau+\varepsilon/2)} + (\varepsilon/2)\nabla_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta}^{(\tau+\varepsilon)}). \end{aligned} \quad (2.5)$$

O símbolo $\nabla_{\mathbf{x}}$, indica o vetor gradiente em relação a uma variável \mathbf{x} , sendo mais conhecido como vetor escore em Estatística. Veja então, que o algoritmo faz uso de derivadas primeiras da função log-densidade não normalizada e portanto aponta para regiões da maior probabilidade, fazendo o algoritmo alcançar mais rapidamente

uma distribuição estacionária da cadeia Markoviana, que é a distribuição alvo da qual buscamos gerar valores.

Ao final da trajetória note que, no ponto $(\boldsymbol{\theta}^{(t_1)}, \mathbf{p}^{(t_1)})$, a diferença entre os Hamiltonianos é próxima de zero devido a discretização do sistema de equações diferenciais. Neste momento a regra de transição de Metropolis é aplicada para corrigir o erro introduzido no sistema de equações. Portanto podemos fazer a transição para o novo valor proposto $(\boldsymbol{\theta}^{(t_1)}, \mathbf{p}^{(t_1)})$ com probabilidade,

$$\begin{aligned} \alpha[(\boldsymbol{\theta}^{(t_1)}, \mathbf{p}^{(t_1)}), (\boldsymbol{\theta}^{(0)}, \mathbf{p}^{(0)})] &= \min \left[\frac{\pi(\boldsymbol{\theta}^{(t_1)}, \mathbf{p}^{(t_1)})}{\pi(\boldsymbol{\theta}^{(0)}, \mathbf{p}^{(0)})}, 1 \right] \\ &= \min \left[\exp[H(\boldsymbol{\theta}^{(t_0)}, \mathbf{p}^{(t_0)}) - H(\boldsymbol{\theta}^{(t_1)}, \mathbf{p}^{(t_1)})], 1 \right], \end{aligned} \quad (2.6)$$

uma vez que o método de Störmer-Verlet garante reversibilidade e simetria (veja Calderhead 2012).

As variáveis ε e M são parâmetros livres. Ambos determinam a rapidez com que a cadeia alcança a distribuição estacionária bem como a mistura (mixing) dos valores propostos. O parâmetro livre ε , representa a discretização do sistema de equações diferenciais e para valores muito grande, a solução do sistema se torna sem sentido. Para valores muito pequenos, não há uma mistura adequada de valores da posteriori. O parâmetro livre M será tomado como a matriz identidade, pois é algo de difícil especificação e mais significativo em problemas complexos de estimação como visto em Calderhead & Girolami (2011).

O algoritmo em sua configuração ideal mantém uma taxa de aceitação em torno de 70% à 95% dos valores gerados, quantidades muito maiores comparado as taxas de aceitação do algoritmo de Metropolis-Hastings. Por fim, o método HMC, somente vale para variáveis aleatórias na reta, sendo então necessário encontrar uma transformação bijetora que faça esta relação.

Suponha que estamos interessados em simular valores de $\pi(\boldsymbol{\theta}|D)$ em que $\boldsymbol{\theta} \in \mathbb{R}^d$. O algoritmo HMC em sua forma simples tomando $M = I$, é dado por,

(i) Forneça uma posição inicial, $\boldsymbol{\theta}^{(0)}$.

(ii) Inicie um contador $i = 1, 2, \dots, N$ (tamanho da cadeia).

- Gere $\mathbf{p}^* \sim N_d(\mathbf{0}, I)$ e $u \sim U(0, 1)$,

- Faça $(\boldsymbol{\theta}^I, \mathbf{p}^I) = (\boldsymbol{\theta}^{(i-1)}, \mathbf{p}^*)$, $H_0 = H(\boldsymbol{\theta}^I, \mathbf{p}^I)$
- Repita em número adequado de loops a solução numérica de Störmer-Verlet,

$$\mathbf{p}^* = \mathbf{p}^* + \frac{\epsilon}{2} \nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}^{(i-1)})$$

$$\boldsymbol{\theta}^{(i-1)} = \boldsymbol{\theta}^{(i-1)} + \epsilon \nabla_{\mathbf{p}} C(\mathbf{p}^*)$$

$$\mathbf{p}^* = \mathbf{p}^* + \frac{\epsilon}{2} \nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}^{(i-1)})$$
- Ao final da trajetória, faça $(\boldsymbol{\theta}^L, \mathbf{p}^L) = (\boldsymbol{\theta}^{(i-1)}, \mathbf{p}^*)$ e $H_1 = H(\boldsymbol{\theta}^L, \mathbf{p}^L)$
- Determine $\alpha[(\boldsymbol{\theta}^L, \mathbf{p}^L), (\boldsymbol{\theta}^I, \mathbf{p}^I)] = \min[\exp(H_0 - H_1), 1]$
- Faça $\boldsymbol{\theta}^{(i)} = \begin{cases} \boldsymbol{\theta}^L, & \text{com probabilidade } \alpha[(\boldsymbol{\theta}^L, \mathbf{p}^L), (\boldsymbol{\theta}^I, \mathbf{p}^I)] > u \\ \boldsymbol{\theta}^I, & \text{caso contrário} \end{cases}$

O número de vezes que repetimos a solução de Störmer-Verlet representa o quão longe avançamos no tempo. Para valores muito pequenos, o algoritmo irá percorrer pequenas regiões do espaço paramétrico. Para valores grandes, o algoritmo explora maiores regiões do espaço paramétrico. No entanto, há uma limitação computacional no sentido de que não há necessidade de avançar em um tempo muito longo. Com algumas iterações o algoritmo já apresenta boa performance.

Exemplo 2.1 Para visualizar a performance do método, vamos retomar o exemplo dado em Coles (2004), página 59, via inferência Bayesiana. O objetivo da análise é ajustar o modelo generalizado de valor extremo (GEV) a um conjunto de dados obtidos a partir da observação anual máxima do nível do mar (em metros) entre os períodos de 1923 à 1987, em Port-Pirie, Austrália. Vale notar que embora as observações tenham sido obtidas no decorrer do tempo, o autor assumiu independência entre observações, uma vez que estas não apresentam dependência temporal.

O histograma dos dados pode ser visto na figura 2.2. Para proceder com o método de Monte Carlo Hamiltoniano, precisamos determinar o vetor gradiente da função log-posteriori. Através da Equação 3.1 e fazendo $\sigma = e^\delta$ para uma parametrização na reta, este vetor é dado por,

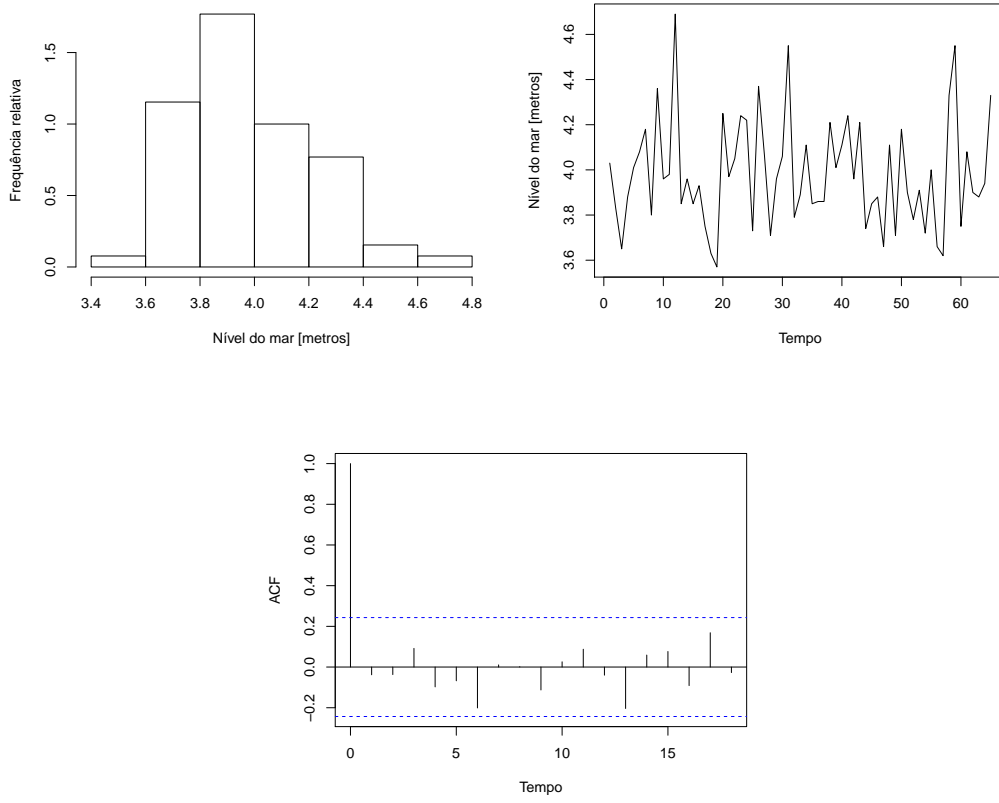


Figura 2.2: Histograma, s rie temporal e autocorrela o para os dados de n vel do mar em Port-Pirie, Australia.

$$\nabla_{\theta} \ell = \begin{bmatrix} \sum_{i=1}^n \frac{1}{\sigma} z_i^{-1} \left((1 + \xi) - z_i^{-1/\xi} \right) + \nabla_{\mu} \log \pi(\mu) \\ \sum_{i=1}^n (1 + \xi) \left(\frac{y_i - \mu}{\sigma} \right) z_i^{-1} - 1 - z_i^{-(1/\xi+1)} \left(\frac{y_i - \mu}{\sigma} \right) + \nabla_{\delta} \log \pi(\delta) \\ \sum_{i=1}^n \frac{\log z_i}{\xi^2} - \left(\frac{1}{\xi} + 1 \right) \left(\frac{y_i - \mu}{\sigma} \right) z_i^{-1} + \frac{1}{\xi} \left(\frac{y_i - \mu}{\sigma} \right) z_i^{-(1/\xi+1)} - \frac{\log z_i}{\xi^2} z_i^{-1/\xi} \\ + \nabla_{\xi} \log \pi(\xi) \end{bmatrix} \quad (2.7)$$

em que, $\theta = [\mu, \delta, \xi]$, $z_i = 1 + \xi(y_i - \mu)/\sigma$ e $\mu, \delta, \xi \stackrel{i.i.d}{\sim} N[0, 25]$

Com o vetor gradiente obtido, o algoritmo foi implementado em linguagem R. Geramos uma amostra de tamanho 6000 descartando 1000 valores com aquecimento. Ap s algumas pr  gera es para adequar a converg ncia do m todo, o par metro livre ε foi tomado igual   0.012 e a solu o de St rmer-Verlet replicada 27 vezes seguidas.

Para efeito de compara o, utilizamos o pacote `evdBayes` para a gera o de valores pelo algoritmo de Metropolis-Hastings com as mesmas quantidades de valores

e aquecimento acima. A configuração de distribuições auxiliares foram obtidas do próprio pacote que já possui os valores ótimos para os parâmetros das distribuições auxiliares. Este pacote também faz uma reparametrização na reta, usando a normal como distribuição auxiliar normal. Além disso faz uso do algoritmo em blocos e aplica a regra de aceitação para cada parâmetro fixado o restante. Desse modo há três taxas de aceitação. Todas as taxas de aceitação ficam em torno de 30 – 50%.

Os resultados são ilustrados pelos gráficos de séries no tempo e autocorrelação nas figuras 2.3 e 2.4, respectivamente para o HMC e MH.

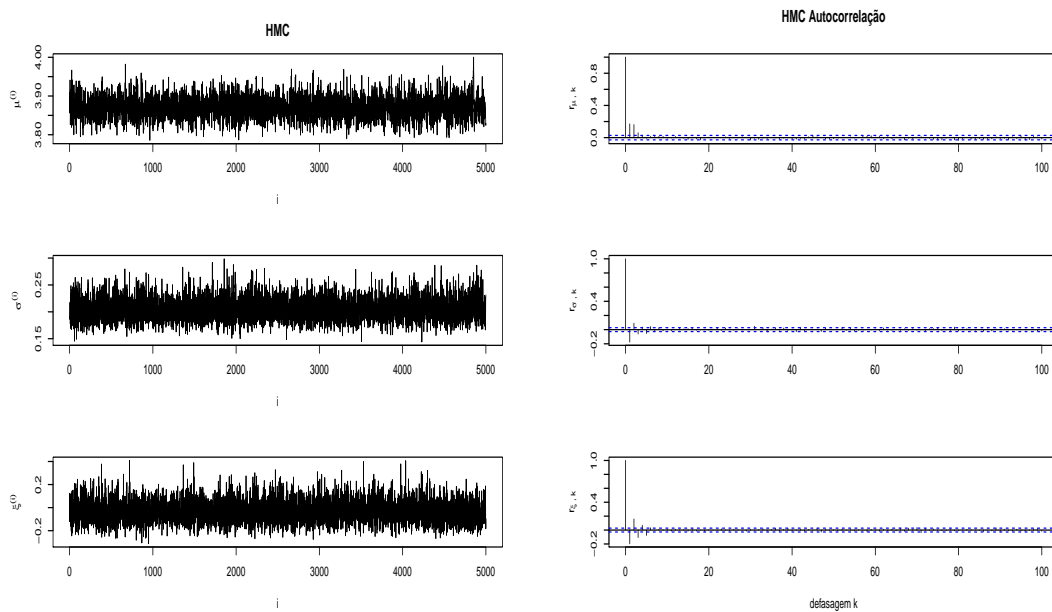


Figura 2.3: Gráficos da série de tempo e autocorrelação para os valores gerados pelo algoritmo HMC.

Note que o algoritmo HMC, com uma taxa de aceitação em torno de 95%, alcança regime estacionário muito mais rápido que o algoritmo MH. Além disso, não há praticamente autocorrelação, não sendo então necessário tomar valores defasados para uma melhor representação da distribuição a posteriori. O algoritmo MH atinge a distribuição estacionário muito lentamente e com alta autocorrelação entre os valores gerados.

2.3 Estudos de simulação

Com o objetivo de avaliar e comparar a performance do algoritmo HMC e do algoritmo de MH, realizamos dois estudos de simulação para a estimação dos parâmetros da

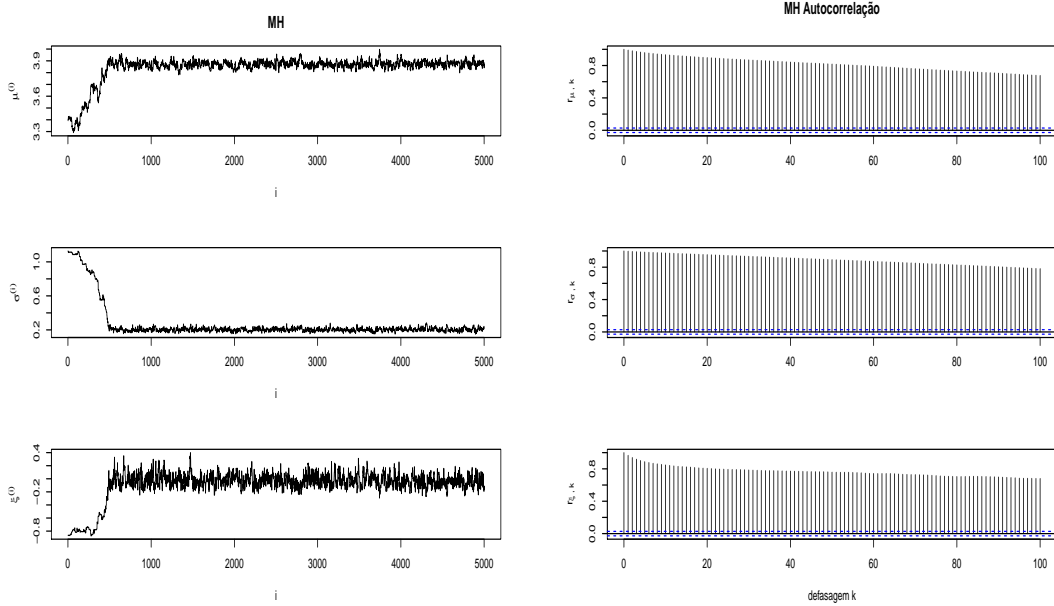


Figura 2.4: Gráficos da série de tempo e autocorrelação para os valores gerados pelo algoritmo MH.

distribuição $GEV(\mu = 2, \sigma = 0.5, \xi = -0.1)$. A escolha dos parâmetros para μ e σ é ao acaso, porém para ξ , escolhemos um valor que segundo Coles (2004) é incomum em aplicações práticas de estimação, pois levam a distribuição com caudas muito pesadas sendo mais difícil fazer inferência.

Em ambos os estudos repetimos o procedimento de replicar 1000 amostras de tamanhos $n = 15, 30, 50, 100$ e calculamos medidas de vício e erro quadrático médio dadas respectivamente por

$$\hat{b} = \frac{\sum_{m=1}^{1000} \hat{\theta}^{(m)} - \theta^{(True)}}{1000}$$

$$\widehat{eqm} = \frac{\sum_{m=1}^{1000} [\hat{\theta}^{(m)} - \theta^{(True)}]^2}{1000}. \quad (2.8)$$

O algoritmo HMC foi implementado na linguagem R assim como o algoritmo MH. Para algoritmo MH utilizamos o pacote `evdBayes` na geração de valores da posteriori. No pacote está implementado o algoritmo MH em blocos com reparametrização na reta. Deste modo têm-se três taxas de aceitação com distribuições auxiliares normais univariadas. Em ambos estudos configuramos as distribuições auxiliares para obter taxas de aceitação em torno de 30% – 50%.

1000 réplicas					
20000 valores / 1100 valores		HMC		MH	
n		viés	EQM	viés	EQM
15	μ	-0.0008	0.0255	-0.0028	0.0250
		0.5169	0.5196	-1.7424	6.0973
	σ	-0.0119	0.0135	-0.0121	0.0130
		0.4572	1.5135	5.0180	51.007
	ξ	-0.0352	0.0737	-0.0364	0.0727
		-0.0681	0.1867	-1.0650	2.5525
30	μ	0.0000	0.0107	-0.0005	0.0108
		-0.2183	0.3943	-2.3592	8.4136
	σ	-0.0098	0.0057	-0.0084	0.0058
		0.3655	1.0837	7.0782	78.279
	ξ	-0.0090	0.0248	-0.0114	0.0256
		-0.0651	0.0965	-1.4178	3.3511
50	μ	-0.0059	0.0079	-0.0045	0.0063
		-0.2202	0.3505	-2.6573	9.8133
	σ	0.0026	0.0336	-0.0028	0.0034
		0.3362	0.8232	8.5333	103.20
	ξ	-0.0124	0.0149	-0.0108	0.0127
		-0.0582	0.0542	-1.5587	3.9191
100	μ	-0.0012	0.0053	-0.0010	0.0033
		-0.4297	0.6297	-3.2037	12.541
	σ	0.0022	0.0017	-0.0023	0.0016
		0.6450	1.7392	10.203	138.04
	ξ	-0.0050	0.0058	-0.0041	0.0053
		-0.0793	0.1241	-1.7940	4.3145

Tabela 2.1: Resultados da simulação do estudo 1 e estudo 2 para cada parâmetro da distribuição GEV com tamanhos amostrais variados. A primeira linha de cada parâmetro corresponde as cadeias de 20000 valores. A linha abaixo corresponde as cadeias de 1100 valores.

Caso 1

Para ambos os algoritmos geramos uma amostra de tamanho 20000, eliminando as 10000 gerações sem tomar valores defasados. A configuração dos parâmetros livres para o HMC foram tomadas após algumas gerações iniciais para verificação de convergência. Esses valores foram mantidos fixos para todas as outras amostras. Em relação ao algoritmo MH, o procedimento foi similar. Tomamos valores dos parâmetros das distribuições auxiliares de modo que forneçam também convergência do método.

Assumimos prioris $\mu, \delta, \xi \stackrel{i.i.d}{\sim} N[0, 25]$ para ambos os algoritmos e em todas as cadeias geradas foi realizado o teste de Geweke para verificar convergência.

Neste primeiro estudo tomamos como estimativa do parâmetro o valor modal da distribuição a posteriori, pois este valor é uma medida de resumo que melhor representa o valor do parâmetro quando as distribuições são assimétricas. Além disso, devido a grande quantidade de valores gerados com a grande quantidade de descarte para aquecimento, não há diferença prática na qualidade da inferência entre os dois métodos, sendo visto na Tabela 2.1.

Caso 2

Neste estudo procedemos de modo similar ao caso anterior, porém gerou-se 1100 valores descartando 100 primeiros valores com aquecimento. O objetivo principal é observar se o algoritmo HMC tende a alcançar a distribuição estacionária de interesse mais rapidamente do que o algoritmo MH e assim fornecer melhores estimativas para os parâmetros.

Como esperamos que muitas das cadeias não atinjam estacionariedade, aplicamos o teste de Geweke rejeitando as cadeias geradas para valores altos da estatística, assim não rejeitamos muitas cadeias geradas para ambos os algoritmos. Também, tomamos como estimativa dos parâmetros a média das distribuições a posteriori, pois é uma medida que é mais influenciada por valores mais distantes.

Pela Tabela 2.1, na segunda linha de cada parâmetro, notamos que os resultados das simulações apresentam-se com melhor qualidade para o algoritmo HMC, evidenciado pelo viés e erro quadrático médio.

2.4 Monte Carlo Hamiltoniano em variedade Riemanniana (RMHMC)

O método Hamiltoniano de Monte Carlo em variedade Riemanniana abrange os mesmos conceitos do algoritmo HMC, no entanto o RMHMC explora as propriedades geométricas da distribuição a posteriori.

Calderhead & Girolami (2011) e Calderhead (2012), definem a dinâmica Hamiltoniana sobre uma hipersuperfície Riemanniana e mostram propriedades importantes do método. Os autores também realizam uma grande quantidade de aplicações sobre modelos de equações diferenciais, evidenciando a performance superior do método em modelos de alta dimensão e estrutura de correlação não usual.

De modo menos rigoroso, o algoritmo RMHMC em cada iteração percorre geodésicas, que são curvas de comprimento mínimo entre dois pontos da hipersuperfície gerada pela distribuição a posteriori. Este comportamento faz com que o algoritmo RMHMC tenha uma taxa de aceitação próxima 99% dos valores gerados e baixa autocorrelação em sua configuração ideal, mesmo em elevadas dimensões.

É interessante notar que na maioria das aplicações estatísticas, a distribuição a posteriori é uma hipersuperfície (uma generalização direta dos objetos no \mathbb{R}^3) e por meio da geometria diferencial, pode-se generalizar conceitos de distância, ângulos e áreas sobre hipersuperfícies curvas.

Definição 2.3 *Uma superfície diferenciável m -dimensional no \mathbb{R}^{m+1} é denominada uma hipersuperfície.*

Note que a distribuição a posteriori é um gráfico de função, i.e., $\pi(\boldsymbol{\theta}|D) : \Theta \subset \mathbb{R}^d \rightarrow \mathbb{R}_+$, então a hipersuperfície gerada é $S(\cdot) : \Theta \subset \mathbb{R}^d \rightarrow \mathbb{R}^{d+1}$, com $S(\boldsymbol{\theta}) = [\boldsymbol{\theta} \ \pi(\boldsymbol{\theta}|D)]^t$. Porém, devido a facilidade analítica, é comum estudar a distribuição a posteriori na escala logarítmica e então $S(\boldsymbol{\theta}) = [\boldsymbol{\theta} \ \log \pi(\boldsymbol{\theta}|D)]^t$. O termo variedade Riemanniana se origina no fato de se associar uma métrica Riemanniana na hipersuperfície em estudo que permite então a determinação de distâncias, ângulos, áreas e a definição da dinâmica Hamiltoniana sobre uma métrica Riemanniana.

Definição 2.4 (Métrica Riemanniana) *Uma métrica Riemanniana numa hipersuperfície S é uma correspondência que associa cada ponto $p \in S$, um produto interno no espaço tangente $T_p S$.*

O espaço tangente $T_p S$ de S em p ($S(\boldsymbol{\theta}) = p$), é um espaço vetorial com base,

$$B(\boldsymbol{\theta}) = \left\{ \frac{\partial S(\boldsymbol{\theta})}{\partial \theta_1}, \dots, \frac{\partial S(\boldsymbol{\theta})}{\partial \theta_d} \right\},$$

e pode visto como uma aproximação linear desta hiperfície em todo ponto $p \in S$. Além disso, todo vetor $u \in T_p S$ é uma combinação linear dos vetores da base B , podendo ser escrito na forma $\mathbf{u} = \sum_{i=1}^d u_i \frac{\partial S(\boldsymbol{\theta})}{\partial \theta_i}$. Deste modo é possível definir um produto interno $I_p(\cdot) : T_p S \rightarrow \mathbb{R}$, i.e,

$$\begin{aligned} I_p(\mathbf{u}) &= \langle \mathbf{u}, \mathbf{u} \rangle \\ &= \left\langle \sum_{i=1}^d u_i \frac{\partial S(\boldsymbol{\theta})}{\partial \theta_i}, \sum_{i=1}^d u_i \frac{\partial S(\boldsymbol{\theta})}{\partial \theta_i} \right\rangle \\ &= [u_1 \dots u_d] \begin{bmatrix} \left\langle \frac{\partial S(\boldsymbol{\theta})}{\partial \theta_1}, \frac{\partial S(\boldsymbol{\theta})}{\partial \theta_1} \right\rangle & \dots & \left\langle \frac{\partial S(\boldsymbol{\theta})}{\partial \theta_1}, \frac{\partial S(\boldsymbol{\theta})}{\partial \theta_d} \right\rangle \\ \vdots & \ddots & \vdots \\ \left\langle \frac{\partial S(\boldsymbol{\theta})}{\partial \theta_d}, \frac{\partial S(\boldsymbol{\theta})}{\partial \theta_1} \right\rangle & \dots & \left\langle \frac{\partial S(\boldsymbol{\theta})}{\partial \theta_d}, \frac{\partial S(\boldsymbol{\theta})}{\partial \theta_d} \right\rangle \end{bmatrix} \begin{bmatrix} u_1 \\ \vdots \\ u_d \end{bmatrix} \\ &= \mathbf{u} G \mathbf{u}^T \end{aligned} \tag{2.9}$$

em que G é a métrica Riemanniana natural induzida pelo \mathbb{R}^d . No entanto, como colocado em Amari & Nagaoka (2000), pode-se definir uma métrica induzida a partir do modelo estatístico assumido, bem conhecida como matriz de informação de Fisher e tomar os elementos de G como,

$$G_{ij} = - \int \frac{\partial^2 \log f_{\mathbf{X}}(\mathbf{x}|\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j} f_{\mathbf{X}}(\mathbf{x}|\boldsymbol{\theta}) d\mathbf{x} \tag{2.10}$$

A matriz de informação de Fisher possui propriedades importantes, como a invariância sobre reparametrizações para o cálculo de geodésicas e sua inversa fornece o limite inferior de Crámer-Rao para a classe dos estimadores não-viciados (veja Schervish 2011). Para uma discussão mais detalhada sobre a definição de uma métrica Riemanniana veja Calderhead (2012). Baseado nesta métrica, Calderhead & Girolami (2011) e Calderhead (2012) definem a função Hamiltoniana na seguinte forma,

$$H(\boldsymbol{\theta}, \mathbf{p}) = - \log \pi(\boldsymbol{\theta}|D) + \frac{1}{2} \log \det G(\boldsymbol{\theta}) + \frac{1}{2} \mathbf{p}^t G(\boldsymbol{\theta})^{-1} \mathbf{p}. \tag{2.11}$$

Desta forma a distribuição conjunta de $(\boldsymbol{\theta}, \mathbf{p})$ se torna,

$$f(\boldsymbol{\theta}, \mathbf{p}) \propto L(\boldsymbol{\theta}|D) \pi(\boldsymbol{\theta}) N_d(\mathbf{0}, G(\boldsymbol{\theta})), \tag{2.12}$$

logo, a dinâmica Hamiltoniana fica dada por,

$$\begin{aligned}\frac{d\boldsymbol{\theta}}{dt} &= \frac{\partial H}{\partial \mathbf{p}} = G(\boldsymbol{\theta})^{-1} \mathbf{p} \\ \frac{dp_i}{dt} &= -\frac{\partial H}{\partial \theta_i} = \frac{\partial \log \pi(\boldsymbol{\theta}|D)}{\partial \theta_i} - \frac{1}{2} \text{tr} \left[G(\boldsymbol{\theta})^{-1} \frac{\partial G(\boldsymbol{\theta})}{\partial \theta_i} \right] + \frac{1}{2} \mathbf{p}^t G(\boldsymbol{\theta})^{-1} \frac{G(\boldsymbol{\theta})}{\partial \theta_i} G(\boldsymbol{\theta})^{-1} \mathbf{p}\end{aligned}\tag{2.13}$$

Neste caso a distribuição conjunta não é fatorável e para simular valores através da dinâmica Hamiltoniana, é necessário aplicar a solução numérica generalizada de Störmer-Verlet da seguinte maneira (veja Leimkuhler & Reich 2005),

(i) Forneça um valor inicial. $\boldsymbol{\theta}^{(0)}$.

(ii) Inicie um contador $i = 1, 2, \dots, SS$ (tamanho da cadeia).

- Gere $\mathbf{p}^{(i)} \sim N_d[\mathbf{0}, G(\boldsymbol{\theta}^{(i-1)})]$ e $u \sim U(0, 1)$,
- Faça $(\boldsymbol{\theta}^I, \mathbf{p}^I) = (\boldsymbol{\theta}^{(i-1)}, \mathbf{p}^{(i)})$, $H_0 = H(\boldsymbol{\theta}^I, \mathbf{p}^I)$
- Para $n = 1$ até L faça (replicando a solução numérica de Störmer-Verlet)

$$\bar{\mathbf{p}}^{(0)} = \mathbf{p}^{(n)}$$

Para $k = 1$ até nfps (atualize o momento com iterações de ponto fixo)

$$\bullet \bar{\mathbf{p}}^{(k)} = \mathbf{p}^{(n)} - \frac{\epsilon}{2} \nabla_{\boldsymbol{\theta}} H(\boldsymbol{\theta}^{(n)}, \bar{\mathbf{p}}^{(k-1)})$$

$$\nabla_{\boldsymbol{\theta}} H = -\frac{\partial \ell}{\partial \theta_i} + \frac{1}{2} \text{tr} \left[G(\boldsymbol{\theta})^{-1} \frac{\partial G(\boldsymbol{\theta})}{\partial \theta_i} \right] - \frac{1}{2} \mathbf{p}' G(\boldsymbol{\theta})^{-1} \frac{G(\boldsymbol{\theta})}{\partial \theta_i} G(\boldsymbol{\theta})^{-1} \mathbf{p}$$

$$\mathbf{p}^{(n+\frac{1}{2})} = \bar{\mathbf{p}}^{(k)}$$

$$\bar{\boldsymbol{\theta}}^{(0)} = \boldsymbol{\theta}^{(n)}$$

Para $k = 1$ até nfps (atualize os parâmetros com iterações de ponto fixo)

$$\bullet \bar{\boldsymbol{\theta}}^{(k)} = \boldsymbol{\theta}^{(n)} + \frac{\epsilon}{2} \nabla_{\mathbf{p}} H(\boldsymbol{\theta}^{(n)}, \mathbf{p}^{(n+1/2)}) + \frac{\epsilon}{2} \nabla_{\mathbf{p}} H(\bar{\boldsymbol{\theta}}^{(k-1)}, \mathbf{p}^{(n+1/2)})$$

$$\nabla_{\mathbf{p}} H = G(\boldsymbol{\theta})^{-1} \mathbf{p}$$

$$\boldsymbol{\theta}^{(n+1)} = \bar{\boldsymbol{\theta}}^{(k)}$$

Atualize o momento

$$\bullet \mathbf{p}^{(n+1)} = \mathbf{p}^{(n+1/2)} - \frac{\epsilon}{2} \nabla_{\boldsymbol{\theta}} H(\boldsymbol{\theta}^{(n+1)}, \mathbf{p}^{(n+1/2)})$$

- Ao final da trajetória faça

$$\boldsymbol{\theta}^L = \boldsymbol{\theta}^{(n+1)}$$

$$\mathbf{p}^L = \mathbf{p}^{(n+1)}$$

$$H_1 = H(\boldsymbol{\theta}^L, \mathbf{p}^L)$$

- Determine $\alpha[(\boldsymbol{\theta}^L, \mathbf{p}^L), (\boldsymbol{\theta}^I, \mathbf{p}^I)] = \min[\exp(H_0 - H_1), 1]$
- Faça $\boldsymbol{\theta}^{(i)} = \begin{cases} \boldsymbol{\theta}^L, & \text{com probabilidade } \alpha[(\boldsymbol{\theta}^L, \mathbf{p}^L), (\boldsymbol{\theta}^I, \mathbf{p}^I)] > u \\ \boldsymbol{\theta}^I, & \text{caso contrário} \end{cases}$

É extremamente importante observar que o algoritmo RMHMC é um método dinâmico na proposição de valores na evolução da cadeia Markoviana. A cada iteração, a estrutura de proposição de um novo valor se adapta a geometria local da distribuição alvo dado o atual estado da cadeia. No método clássico do algoritmo MH, a estrutura de proposição se mantém essencialmente fixa mesmo se considerarmos a informação geométrica do vetor gradiente e também de alguma métrica.

Uma variante do RMHMC é considerar a matriz G como uma métrica fixa, i.e, avaliar G em algum ponto do espaço paramétrico. Por exemplo, tomar $\hat{\boldsymbol{\theta}} = MAP$ (máximo a posteriori) e calcular $G = G(\hat{\boldsymbol{\theta}})$. A principal vantagem é que, a matriz informação de Fisher G avaliada no MAP, em dimensões não muito elevadas, fornece uma boa performance e simplifica o algoritmo, i.e., é o mesmo que utilizar a solução de Störmer-Verlet com $M = G(\hat{\boldsymbol{\theta}})$.

Além disso, se a determinação da matriz de informação de Fisher é analiticamente complexa, podemos usar a métrica obtida em (2.9), que é o produto interno das derivadas parciais em relação a cada parâmetro do modelo. Se a derivação dos elementos dessa matriz em respeito aos parâmetros é ainda tarefa difícil, podemos utilizar esta métrica avaliada no MAP como feito anteriormente.

Note que quando obtemos por uma métrica para a utilização do algoritmo RMHMC, não precisamos que métrica descreva perfeitamente a curvatura da distribuição alvo, é possível utilizar uma aproximação mais grosseira que informa de modo geral a forma desta distribuição.

Aplicações usando o algoritmo RMHMC serão dadas no capítulo 3 e 4, onde introduzimos e aplicamos modelos Bayesianos paramétricos e não-paramétricos em valores extremos. Provas de convergência para uma distribuição estacionária podem ser vistas em Calderhead (2012).

3

Teoria de valores extremos

Apresentamos neste capítulo os fundamentos e teoremas que embasam os modelos estatísticos aplicados ao estudo de dados com característica de extremos. Introduzimos o modelo Bayesiano não-paramétrico baseado no processo Gaussiano e sua aplicação em modelos de valores extremos. Mostramos ao longo deste capítulo aplicações com o método de Monte Carlo Hamiltoniano na modelagem paramétrica sob a abordagem Bayesiana. As referências principais são Kotz & Nadarajah (2000), Coles (2004), Castillo et al. (2004), Rasmussem & Williams (2006), Coles et al. (2003). Barreto-Souza & Vasconcellos (2011), Schervish (2011) e Neal (1998).

3.1 Fundamentos básicos

Atualmente a análise estatística de valores extremos se concentra em dois tipos. Na modelagem do máximo de um conjunto de variáveis aleatórias i.i.d (block maxima models) e sobre os modelos que consideram a excessão de um certo limiar (threshold exceedance models), digamos u .

No primeiro tipo, nos baseamos na distribuição assintótica de uma variável aleatória $M_n = \max(X_1, \dots, X_n)$ para n suficientemente grande. No segundo tipo, tomamos como valores extremos aqueles valores que excedem um limiar u , que pode ser predefinido ou não. Focamos este trabalho no primeiro tipo de estudo de valores extremos.

Definição 3.1 *Uma variável aleatória X é dita seguir uma distribuição de valor ex-*

tremo generalizada (GEV) se sua função de distribuição é dada por,

$$H(x|\xi, \mu, \sigma) = \begin{cases} \exp \left\{ - \left(1 + \xi \frac{x - \mu}{\sigma} \right)^{-1/\xi} \right\} I_{(0, \infty)}(1 + \xi(x - \mu)/\sigma), & \xi \neq 0 \\ \exp \left\{ - \exp \left(-\frac{x - \mu}{\sigma} \right) \right\} I_{\mathbb{R}}(x), & \xi \rightarrow 0. \end{cases} \quad (3.1)$$

sendo $\mu \in \mathbb{R}$, $\sigma \in \mathbb{R}_+$ e $\xi \in \mathbb{R}$ os parâmetros de locação, escala e forma respectivamente.

A distribuição GEV abrange outras distribuições conhecidas na literatura dependendo do valor do parâmetro de forma que define o comportamento da cauda da distribuição. Se $\xi = 0$ a distribuição tem suporte em \mathbb{R} sendo conhecida como distribuição de Gumbel para máximos. Se $\xi > 0$ a distribuição tem suporte em $(\mu - \sigma/\xi, \infty)$, sendo chamada distribuição de Fréchet. Se $\xi < 0$ a distribuição tem suporte $(-\infty, \mu - \sigma/\xi)$ e é conhecida como distribuição Weibull negativa.

Para máximos ou mínimos de uma sequência de variáveis aleatórias i.i.d adequadamente padronizada, a teoria dos valores extremos nos diz que se uma distribuição limite para os mínimos ou máximos existe ela deve ser uma das três descritas acima. Este resultado é formalizado a seguir considerando-se os valores máximos.

Teorema 3.1 Para X_1, \dots, X_n variáveis aleatórias independentes e identicamente distribuídas com função de distribuição F , suponha que existam as sequências $\{a_n\}$ e $\{b_n\}$ tais que para alguma distribuição limite não degenerada $G(x)$,

$$\lim_{n \rightarrow \infty} P \left(\frac{M_n - b_n}{a_n} \leq x \right) = G(x), \quad x \in \mathbb{R}$$

sendo $M_n = \max\{X_1, \dots, X_n\}$. Então existem, μ , $\xi \in \mathbb{R}$ e $\sigma > 0$ tais que $G(x) = H(x|\xi, \mu, \sigma)$.

Analogamente, para valores extremos mínimos, defina $\tilde{M}_n = -M_n$, pelo método de transformação acumulada segue que,

$$H_{\tilde{M}_n}(x|\xi, \tilde{\mu}, \sigma) = 1 - \exp \left\{ - \left(1 - \xi \frac{x - \tilde{\mu}}{\sigma} \right)^{-1/\xi} \right\} I_{(0, \infty)}(1 - \xi(x - \tilde{\mu})/\sigma), \quad (3.2)$$

sendo $\tilde{\mu} = -\mu$.

Clur (2010) mostra que, se $X|\mu, \sigma, \xi \sim \text{GEV}(\mu, \sigma, \xi)$ a esperança e variância da variável aleatória são respectivamente dadas por,

$$E[X] = \begin{cases} \mu - \frac{\sigma}{\xi} + \frac{\sigma}{\xi}\Gamma(1 - \xi), & \xi < 1 \\ \infty, & \text{caso contrário} \end{cases}, \quad (3.3)$$

$$V[X] = \begin{cases} \frac{\sigma^2}{\xi^2} [\Gamma(1 - 2\xi) - \Gamma^2(1 - \xi)], & \xi < \frac{1}{2} \\ \infty, & \text{caso contrário} \end{cases} \quad (3.4)$$

A moda da distribuição GEV é dada por,

$$\operatorname{argmax}_{1+\xi(x-\mu)/\sigma > 0} f(x|\mu, \sigma, \xi) = \mu + \frac{\sigma}{\xi} [(1 + \xi)^{-\xi} - 1], \quad \forall \mu, \sigma, \xi \in \Theta. \quad (3.5)$$

Prescott & Walden (1980) obtém a matriz de informação de Fisher esperada $I(\mu, \sigma, \xi)$ para uma amostra aleatória de tamanho n ,

$$n \begin{bmatrix} \frac{A}{\sigma^2} & -\frac{1}{\sigma^2\xi}[A - \Gamma(2 + \xi)] & -\frac{1}{\sigma\xi} \left(B - \frac{A}{\xi} \right) \\ \cdot & \frac{1}{\sigma^2\xi^2}[1 - 2\Gamma(2 + \xi) + A] & -\frac{1}{\sigma\xi^2} \left[1 - \gamma + \frac{1 - \Gamma(2 + \xi)}{\xi} - B + \frac{A}{\xi} \right] \\ \cdot & \cdot & \frac{1}{\xi^2} \left[\frac{\pi^2}{6} + \left(1 - \gamma + \frac{1}{\xi} \right)^2 - \frac{2B}{\xi} + \frac{A}{\xi^2} \right] \end{bmatrix}, \quad (3.6)$$

sendo $A = (1 + \xi)^2\Gamma(1 + 2\xi)$, $B = \Gamma(2 + \xi)[\psi(1 + \xi) + (1 + \xi)\xi^{-1}]$, $\Gamma(x)$ a função gamma, $\psi(x)$ a função digamma e γ a constante de Euler ($\cong 0.577215$). Notar que a matriz $I(\mu, \sigma, \xi)$ somente existe para valores de $\xi > -0.5$, pois a função gamma possui domínio em \mathbb{R}_+ .

Smith (1985) estuda as propriedades ótimas do estimador de máxima verossimilhança notando que para $\xi > -0.5$, o EMV é regular no sentido de possuir normalidade e eficiência assintótica, caso contrário, o estimador não possui propriedade assintótica ótima. Para $\xi < -1$ o EMV é dificilmente obtido. Segundo Coles (2004), tais situações são raramente encontradas em prática e não limitam o uso do EMV.

3.2 Autoregressão temporal em valores extremos

Modelos paramétricos de valores extremos são aplicados em várias áreas de ciência. Castillo et al. (2004) fazem um extensa revisão com vários artigos nas mais variadas

áreas de estudo. Para citar algumas aplicações, há estudos em, meteorologia, poluição, enchentes, fadiga de materiais, etc. Na área da engenharia, temos estudos em engenharia oceânica, engenharia hidráulica, engenharia de estruturas, tráfego de rodovias, etc. Em suma, a distribuição GEV permite modelar locação, forma e escala supondo qualquer tipo de função sobre os parâmetros respeitando seus respectivos espaços paramétricos, i.e.,

$$Y(x) \sim GEV[\mu(x), \sigma(x), \xi(x)]. \quad (3.7)$$

Em que $\mu(x)$, $\sigma(x)$, $\xi(x)$, são funções de covariáveis ou funções do tempo. A suposição de alguma forma funcional para o parâmetro ξ é tarefa difícil, uma vez que este parâmetro torna a estimação do modelo de difícil alcance devido ao pequeno intervalo em que a distribuição GEV possui momentos de primeira e segunda ordem.

Estudos específicos em séries temporais são propostos por Balakrishnan & Shiji (2013), em que os autores supõem uma distribuição de valores extremos tipo I para a distribuição marginal do processo $\{y_t\}$ e determinam a partir disso a distribuição do ruído. Toulemonde et al. (2010) realizam um estudo teórico e aplicado na investigação de poluição atmosférica.

Zhao et al. (2011) realizam uma abordagem Bayesiana sobre um modelo autoregressivos e autoregressivos condicionalmente heterocedásticos AR-ARCH(1, 1), em que a estrutura de dependência condicional do modelo é dada por

$$\begin{aligned} x_t &\sim GEV(\mu_t, \sigma_t, \xi) \\ x_t &= \beta_0 + \beta_1 x_{t-1} + e_t \\ \sigma_t^2 &= \alpha_0 + \alpha_1 \sigma_{t-1}^2 + \alpha_2 e_{t-1}^2 \\ e_t &= x_{t-1} - \mu_{t-1}. \end{aligned} \quad (3.8)$$

As distribuições a priori são tomadas não-informativas num sentido 'flat', porém, vale notar, que o autor impõe prioris uniformes sobre pequenos intervalos na reta. Uma interessante seção é dedicada ao estudo do parâmetro de forma também variar no tempo, fornecendo assim uma melhor qualidade de ajuste do modelo GEV.

Seguindo neste mesmo sentido, propomos estudar um modelo de série temporal autoregressivo de ordem p , assumindo que o ruído siga uma distribuição generalizada de valor extremo. A justificativa é que, não há na literatura, nenhum estudo

de dependência temporal autoregressiva de qualquer ordem através da distribuição GEV. Além disso, características de assimetria e caudas pesadas aparecem em vários comportamentos de séries temporais como poder ser visto em Balakrishnan & Shiji (2013) e Zhao et al. (2011).

Uma quantidade importante obtida no desenvolvimento do modelo é a respectiva matriz de informação de Fisher, objeto de fundamental importância para a implementação do algoritmo RMHMC. Num contexto clássico, a matriz fornece o limite inferior de Cramér-Rao e é necessária para construção de intervalos de confiança.

3.3 Modelo autoregressivo-GEV

Considere a estrutura geral de um modelo autoregressivo de ordem p e assuma que o ruído segue uma distribuição GEV, escrevemos

$$Y_t = \mu + \sum_{j=1}^p \theta_j Y_{t-j} + e_t, \quad (3.9)$$

em que

$$e_t \stackrel{i.i.d.}{\sim} GEV(0, \sigma, \xi), \quad \forall t. \quad (3.10)$$

Assumindo estacionariedade fraca (ou de segunda ordem) e restringindo que $\xi \in (-0.5, 0.5)$, temos que,

$$\begin{aligned} \mu_{Y_t} = E[Y_t] &= \frac{\mu_{e_t} + \mu}{1 - \sum_{j=1}^p \theta_j}, \quad \forall t \\ \mu_{e_t} = E[e_t] &= -\frac{\sigma}{\xi} + \frac{\sigma}{\xi} \Gamma(1 - \xi), \\ \sigma_{e_t}^2 = V[e_t] &= \frac{\sigma^2}{\xi^2} [\Gamma(1 - 2\xi) - \Gamma^2(1 - \xi)]. \end{aligned} \quad (3.11)$$

É fácil observar que Y_t , dado toda informação anterior, possui distribuição GEV, i.e.,

$$Y_t | D_{t-1}, \mu, \boldsymbol{\theta}, \sigma, \xi \sim GEV\left(\mu + \sum_{j=1}^p \theta_j y_{t-j}, \sigma, \xi\right) \quad (3.12)$$

sendo $D_{t-1} = \{y_{t-1}, \dots, y_{t-p}\}$ e $\boldsymbol{\theta} = [\theta_1 \dots \theta_p]'$.

Deste modo, para uma série suficientemente grande e pelo teorema da probabilidade composta a função log-verossimilhança pode ser aproximadamente reescrita como,

$$\ell(\mu, \boldsymbol{\theta}, \sigma, \xi) = \sum_{t=p+1}^T \log[f_{Y_t}(y_t | D_{t-1}, \mu, \boldsymbol{\theta}, \sigma, \xi) I_{\Omega_t}(y_t)], \quad (3.13)$$

sendo

$$\begin{aligned}
\Omega_t &= \{y_t : z_t > 0\}, \quad z_t = 1 + \xi(y_t - \mu_t)/\sigma, \\
\mu_t &= \mu + \sum_{j=1}^p \theta_j y_{t-j}, \\
D_{t-1} &= \{y_{t-1}, \dots, y_{t-p}\}.
\end{aligned} \tag{3.14}$$

O vetor gradiente da função log-verossimilhança fica dado por,

$$\nabla \ell = \begin{bmatrix} \sum_{t=p+1}^T \frac{1}{\sigma} z_t^{-1} \left((1 + \xi) - z_t^{-1/\xi} \right) \\ \sum_{t=p+1}^T \frac{1}{\sigma} z_t^{-1} \left((1 + \xi) - z_t^{-1/\xi} \right) y_{t-1} \\ \vdots \\ \sum_{t=p+1}^T \frac{1}{\sigma} z_t^{-1} \left((1 + \xi) - z_t^{-1/\xi} \right) y_{t-p} \\ \sum_{t=p+1}^T (1 + \xi) \left(\frac{y_t - \mu_t}{\sigma^2} \right) z_t^{-1} - \frac{1}{\sigma} - z_t^{-(1/\xi+1)} \left(\frac{y_t - \mu_t}{\sigma^2} \right) \\ \sum_{t=p+1}^T \frac{\log z_t}{\xi^2} - \left(\frac{1}{\xi} + 1 \right) \left(\frac{y_t - \mu_t}{\sigma} \right) z_t^{-1} + \frac{1}{\xi} \left(\frac{y_t - \mu_t}{\sigma} \right) z_t^{-(1/\xi+1)} \\ - \frac{\log z_t}{\xi^2} z_t^{-1/\xi} \end{bmatrix}. \tag{3.15}$$

3.4 Matriz informação de Fisher

Numa primeira abordagem, para determinar a matriz informação de Fisher, seria necessário determinar as derivadas de segunda ordem da função log-verossimilhança em relação a todos os parâmetros e aplicar a esperança negativa em todas expressões. No entanto, não precisamos obter tais quantidades diretamente. Podemos utilizar os elementos da matriz informação de Fisher obtidos para o caso de uma amostra aleatória e determinar a matriz informação de Fisher para o modelo autoregressivo-GEV. As propriedades utilizadas e a prova detalhada da obtenção desta matriz pode ser consultada no apêndice.

Uma vez que é necessária a parametrização do modelo na reta para utilizar o algoritmo RMHMC, podemos escolher alguma transformação $\boldsymbol{\delta} = \boldsymbol{\delta}(\boldsymbol{\theta}) \in \mathbb{R}^d$ e determinar a matriz informação de Fisher nesta parametrização. Esta transformação é dada por

$G(\boldsymbol{\delta}) = J(\boldsymbol{\theta} \rightarrow \boldsymbol{\delta})'G(\boldsymbol{\delta})J(\boldsymbol{\theta} \rightarrow \boldsymbol{\delta})$. Em que J é a matriz jacobiana da transformação $\boldsymbol{\delta} = \boldsymbol{\delta}(\boldsymbol{\theta})$.

Os elementos das matriz informação de Fisher para o modelo autoregressivo-GEV são dados por,

$$\begin{aligned}
G_{\mu\mu} &= +(T-p)\frac{A}{\sigma^2} \\
G_{\mu\theta_j} &= +\mu_{Y_t}(T-p)\frac{A}{\sigma^2}, \quad j = 1, \dots, p \\
G_{\mu\sigma} &= -(T-p)\frac{1}{\sigma^2\xi}[A - \Gamma(2 + \xi)] \\
G_{\mu\xi} &= -(T-p)\frac{1}{\sigma\xi}\left(B - \frac{A}{\xi}\right) \\
G_{\theta_i\theta_j} &= +(T-p)\frac{A}{\sigma^2}E[Y_{t-i}Y_{t-j}] \quad i, j = 1, \dots, p \\
G_{\sigma\sigma} &= +(T-p)\frac{1}{\sigma^2\xi^2}[1 - 2\Gamma(2 + \xi) + A] \\
G_{\sigma\theta_j} &= -(T-p)\frac{1}{\sigma^2\xi}[A - \Gamma(2 + \xi)]\mu_{Y_t}, \quad j = 1, \dots, p \\
G_{\sigma\xi} &= -(T-p)\frac{1}{\sigma\xi^2}\left[1 - \gamma + \frac{1 - \Gamma(2 + \xi)}{\xi} - B + \frac{A}{\xi}\right] \\
G_{\xi\theta_j} &= -(T-p)\frac{1}{\sigma\xi}\left(B - \frac{A}{\xi}\right)\mu_{Y_t}, \quad j = 1, \dots, p \\
G_{\xi\xi} &= +(T-p)\frac{1}{\xi^2}\left[\frac{\pi^2}{6} + \left(1 - \gamma + \frac{1}{\xi}\right)^2 - \frac{2B}{\xi} + \frac{A}{\xi^2}\right] \tag{3.16}
\end{aligned}$$

Note que as quantidades $E[Y_{t-i}Y_{t-j}]$ são facilmente obtidas para $p \leq 2$, caso contrário essas quantidades são de difícil determinação. Para o uso do algoritmo RMHCM, podemos determinar tais elementos aproximados para $p \geq 3$, calculando a autocovariância amostral mais a média ao quadrado da série (μ_{Y_t}). Embora isto seja uma aproximação da matriz de informação, ainda é algo ideal para modelos de dimensões relativamente alta. Além disso, se obtamos tais quantidades aproximadas não podemos usar o algoritmo com métrica variável, pois neste caso é necessário derivar a matriz de informação em relação a cada parâmetro.

3.5 Estudo de simulação

Neste estudo, iremos prosseguir de modo similar ao estudo do capítulo 2. O objetivo principal é observar se, com o aumento do número de parâmetros do modelo AR-GEV concomitante ao aumento do tamanho amostral, os algoritmos HMC e RMHMC

tendem a alcançar distribuição estacionária mais rapidamente em relação um ao outro, mesmo com cadeias pequenas.

A configuração do procedimento é como segue. Geramos 1000 amostras de tamanho $n = 60, 150, 300$, para os modelos $AR-GEV(p)$, $p = 1, 2, 3$. Para cada amostra, geramos cadeias de tamanho 600 e descartamos os 100 primeiros valores como aquecimento para ambos algoritmos. Os modelos simulados são todos estacionários e dados por,

$$\begin{aligned}
M_1 : Y_t &= -1 + 0.8Y_{t-1} + e_t \\
M_2 : Y_t &= -1 + 0.9Y_{t-1} - 0.8Y_{t-2} + e_t \\
M_3 : Y_t &= -1 - 1.56Y_{t-1} - 0.55Y_{t-2} + 0.04Y_{t-3} + e_t \\
e_t &\stackrel{i.i.d}{\sim} GEV(0, \sigma = 1, \xi = 0.3) \quad \forall t.
\end{aligned} \tag{3.17}$$

Todos os parâmetros são independentes a priori com distribuições vagas, exceto para ξ , em que limitamos sua região de variação no intervalo $(-0.5, 0.5)$, assegurando assim a existência da média e da variância dos processos autoregressivos estudados, i.e,

$$\begin{aligned}
\mu &\sim N(0, 25), \\
\theta_j &\sim N(0, 25) \quad j = 1, \dots, p, \\
\sigma &\sim IG(0.1, 0.01) \\
\xi &\sim U(-0.5, 0.5)
\end{aligned} \tag{3.18}$$

Para o algoritmo HMC tomamos $\epsilon = 0.006$ e repetimos a solução de Störmer-Verlet 13 vezes. Para o RMHMC, usamos o algoritmo com métrica fixa dada pela matriz de informação do respectivo modelo avaliada na estimativa MAP . Nos modelos $AR-GEV(1)$ e $AR-GEV(2)$ determinamos o elementos $E[Y_t^2]$ e $E[Y_t Y_{t+1}]$ de forma fechada $\forall t$. Para o modelo $AR-GEV(3)$ usamos aproximadamente que $E[Y_t Y_{t+i}] \approx \mu_{Y_t}^2 + \widehat{C}(Y_t, Y_{t+i})$, para $i = 0, 1, 2$, em que \widehat{C} é a covariância amostral. Tomamos $\epsilon = 0.15$ e repetimos a solução de Störmer-Verlet 13 vezes.

Em ambos algoritmos fixamos os valores iniciais dados por, $[0, 0.4, 2.718, 0.01]$, $[0, 0.5, -0.4, 2.718, 0.01]$ e $[0, -1, -0.2, 0, 2.718, 0.01]$ para as respectivas ordens dos modelos.

Os resultados indicam que, para o modelo de ordem 1 e tamanhos amostrais variados, a performance do algoritmo HMC e RMHMC são similares. No modelo de ordem 2 os algoritmos apresentam-se também com performance similar. No modelo de ordem 3, com o aumento do tamanhos amostral, observa-se diferença significativa no viés e erro quadrático médio em relação aos amostradores. Isto decorre do fato dos valores iniciais passarem a ficar distantes da região de maior probabilidade da posteriori com o aumento da amostra, sendo que para o algoritmo RMHMC indica não ter influência.

1000 réplicas AR-GEV(p)		1			2			3					
$T = 600$		HMC		RMHMC		HMC		RMHMC		HMC		RMHMC	
n		viés	eqm	viés	eqm	viés	eqm	viés	eqm	viés	eqm	viés	eqm
60	μ	-0.0236	0.5600	-0.0292	0.7966	-0.0175	0.3382	0.0086	0.0829	-0.0323	1.1502	-0.0339	1.2462
	σ	-0.0238	0.5701	-0.0269	0.6677	-0.0039	0.0173	-0.0124	0.1506	-0.0124	0.1701	-0.0129	0.1817
	ξ	0.0276	0.7667	0.0322	0.9849	0.0111	0.1300	0.0291	0.8020	0.0290	0.9294	0.0283	0.8730
	θ_1	0.0233	0.5459	0.0173	0.2704	0.0058	0.0381	0.0076	0.0570	0.0021	0.0048	-0.0038	0.0159
	θ_2					-0.0050	0.0278	-0.0058	0.0332	0.0121	0.1611	0.0033	0.0119
	θ_3									0.0095	0.0995	0.0085	0.0787
150	μ	0.0018	0.0035	0.0007	0.0005	-0.0100	0.1115	-0.0077	0.0668	-0.0953	10.006	-0.0376	1.5560
	σ	-0.0242	0.5899	-0.0135	0.1843	-0.0007	0.0005	-0.0011	0.0015	-0.0863	8.2037	-0.0323	1.1485
	ξ	0.0053	0.0282	-0.0016	0.0027	0.0022	0.0053	0.0025	0.0071	0.0369	1.5019	0.0170	0.3194
	θ_1	0.0144	0.2095	0.0009	0.0926	0.0005	0.0002	0.0008	0.0006	-0.0082	0.0745	-0.0087	0.0815
	θ_2					-0.0014	0.0023	-0.0018	0.0038	-0.0060	0.0406	-0.0055	0.0334
	θ_3									-0.0004	0.0002	0.0020	0.0047
300	μ	-0.0009	0.0008	-0.0002	0.0054	-0.0051	0.0286	-0.0048	0.0257	-0.3205	106.06	-0.0400	1.6533
	σ	-0.0293	0.8555	-0.0058	0.0344	-0.0073	0.0588	-0.0053	0.0315	-0.3208	106.26	-0.0444	2.0343
	ξ	0.0225	0.5082	-0.0007	0.0005	0.0005	0.0003	0.0000	0.0000	-0.0471	2.2938	-0.0136	0.1923
	θ_1	0.0232	0.5391	0.0053	0.0289	0.0012	0.0017	0.0015	0.0027	-0.0046	2.1750	-0.0221	0.5036
	θ_2					-0.0011	0.0014	-0.0016	0.0028	-0.0471	2.2938	-0.0136	0.1923
	θ_3									-0.0136	0.1924	0.0007	0.0005

Tabela 3.1: Resultados da simulação

3.6 Aplicação em dados reais

Nesta aplicação, as observações representam o máximo anual do nível do lago Michigan, obtidas através do valor máximo das médias de cada mês no ano, entre os anos de 1860 à 1955 ($T = 96$). Os dados podem ser encontrados em,

<https://datamarket.com/data/set/22p3/>

Primeiramente notamos que os gráficos de autocorrelação e autocorrelação parcial indicam um modelo autoregressivo de ordem 1 (figura 3.1), então propomos o modelo,

$$\begin{aligned} Y_t &= \mu + \theta Y_{t-1} + e_t, \\ e_t &\stackrel{i.i.d.}{\sim} \text{GEV}(0, \sigma, \xi) \quad \forall t \end{aligned} \quad (3.19)$$

de modo que,

$$Y_t | y_{t-1}, \mu, \theta, \sigma, \xi \sim \text{GEV}(\mu + \theta y_{t-1}, \sigma, \xi), \quad (3.20)$$

e função de log-verossimilhança dada por,

$$\ell = \ell(\mu, \theta, \sigma, \xi) \cong \sum_{t=2}^T \log[f(y_t | \mu + \theta y_{t-1}, \sigma, \xi) I_{\Omega_t}(y_t)] \quad (3.21)$$

com $\Omega_t = \{y_t : 1 + \xi(y_t - \mu - \theta y_{t-1})/\sigma > 0\}$.

Considerando as priors dadas anteriormente em (3.18), o vetor gradiente da função log-posteriori com parametrização $\delta = \log(\sigma)$ torna-se,

$$\nabla \ell = \begin{bmatrix} \sum_{t=2}^T \frac{1}{\sigma} z_t^{-1} \left((1 + \xi) - z_t^{-1/\xi} \right) + \nabla \log \pi(\mu) \\ \sum_{t=2}^T \frac{1}{\sigma} z_t^{-1} \left((1 + \xi) - z_t^{-1/\xi} \right) y_{t-1} + \nabla \log \pi(\theta) \\ \sum_{t=2}^T (1 + \xi) \left(\frac{y_t - \mu_t}{\sigma} \right) z_t^{-1} - 1 - z_t^{-(1/\xi+1)} \left(\frac{y_t - \mu_t}{\sigma} \right) + \nabla \log \pi(\delta) \\ \sum_{t=2}^T \frac{\log z_t}{\xi^2} - \left(\frac{1}{\xi} + 1 \right) \left(\frac{y_t - \mu_t}{\sigma} \right) z_t^{-1} + \frac{1}{\xi} \left(\frac{y_t - \mu_t}{\sigma} \right) z_t^{-(1/\xi+1)} \\ - \frac{\log z_t}{\xi^2} z_t^{-1/\xi} + \nabla \log \pi(\xi) \end{bmatrix} \quad (3.22)$$

sendo $z_t = 1 + \xi(y_t - \mu - \theta y_{t-1})/\sigma$.

Procedendo com a aplicação, retiramos as três últimas observações para checar predições futuras e aplicamos o algoritmo RMHMC com métrica fixa para geração dos valores da posteriori. Determinamos a métrica fixa na estimativa MAP e aplicamos

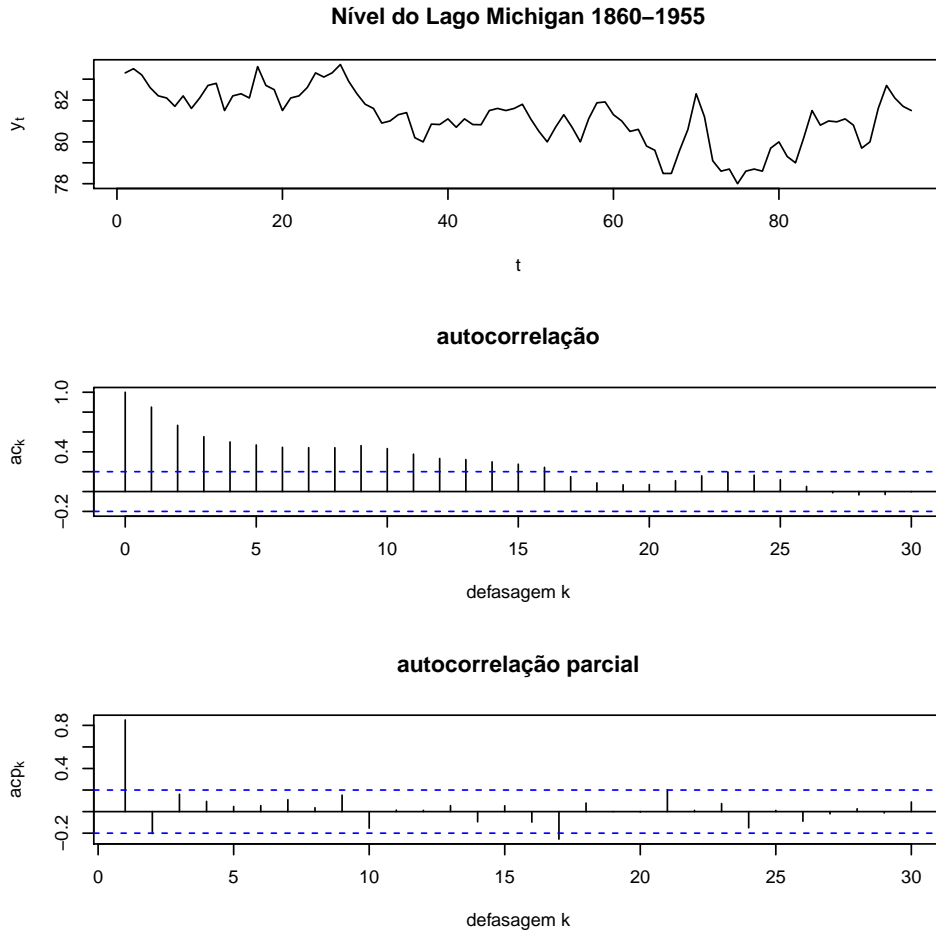


Figura 3.1: Série temporal e gráficos de autocorrelação e autocorrelação parcial.

o algoritmo com stepsize 0.06, repetindo a solução de störmer-verlet 11 vezes. Assim, após a geração de 21000 valores descartando as 1000 primeiras com aquecimento, obtivemos os resultados na tabela 3.2 e nas figuras 3.2 e 3.4.

Pela tabela 3.2, observamos que o modelo estimado é estacionário ao nível de 95% de credibilidade. Além disso, a estimativa de ξ está em torno de -0.25 com pequeno desvio-padrão, caracterizando assim uma distribuição com assimetria moderada. A figura 3.2 indica convergência e baixa correlação entre os valores gerados da distribuição a posteriori. A figura 3.4 ilustra a alta dependência (inversamente proporcional) a posteriori entre os parâmetros μ e θ . Note que o gráfico de dispersão ilustra uma região de variação altamente estreita.

Em relação as previsões futuras, uma vez que seguimos a abordagem Bayesiana, precisamos determinar primeiramente a distribuição preditiva aproximada via método da composição e seguidamente tomar a média da distribuição, ou outra medida de

$N = 20000$	μ	θ	σ	ξ
$\widehat{E}[\cdot D]$	5.929	0.923	0.692	-0.258
$\widehat{DP}[\cdot D]$	3.350	0.041	0.055	0.058
$\widehat{\text{Moda}}$	6.369	0.922	0.687	-0.261
$\widehat{\text{Mediana}}$	5.945	0.923	0.689	-0.259
$IC\ 95\%$	[0.443, 11.437]	[0.856, 0.991]	[0.609, 0.790]	[-0.351, -0.160]

Tabela 3.2: Aproximações para média, desvio-padrão, mediana e intervalo de credibilidade a posteriori

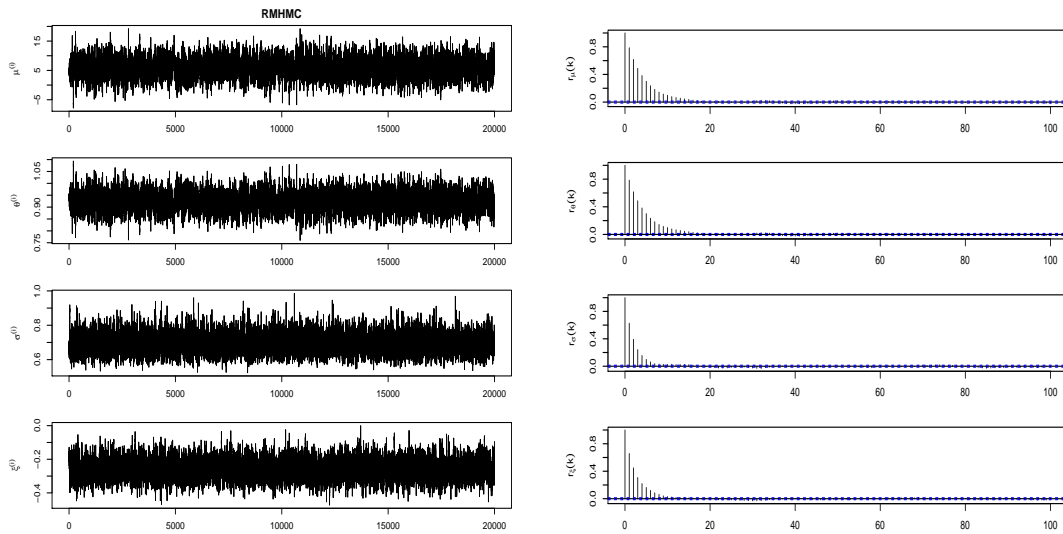


Figura 3.2: Gráficos da série de tempo e autocorrelação para os valores gerados pelo algoritmo RMHMC.

posição como valor preditivo, i.e.,

$$\begin{aligned}
\pi(y_{T+j}|D) &= \int_{\Theta} f(y_{T+j}|\mu + \theta y_{T+j-1}, \sigma, \xi) \pi(\mu, \theta, \sigma, \xi|D) d(\mu, \theta, \sigma, \xi) \\
&= E_{\mu, \theta, \sigma, \xi|D}[f(y_{T+j}|\mu + \theta y_{T+j-1}, \sigma, \xi)]
\end{aligned} \tag{3.23}$$

e então calcular,

$$\begin{aligned}
\hat{y}_{T+j} &= E[y_{T+j}|D] \\
&= E\left[E[y_{T+j}|\mu, \theta, \sigma, \xi, D]\right] \\
&\cong \frac{1}{N} \sum_{i=1}^N y_{T+j}^{(i)} |\mu^{(i)} + \theta^{(i)} y_{T+j-1}^{(i)}, \sigma^{(i)}, \xi^{(i)},
\end{aligned} \tag{3.24}$$

para $j = 1, 2, 3$.

Na figura 3.3 vemos como as previsões se comportam. Todos os valores observados estão dentro do intervalo de credibilidade das distribuições preditivas que tendem a acompanhar a série temporal.

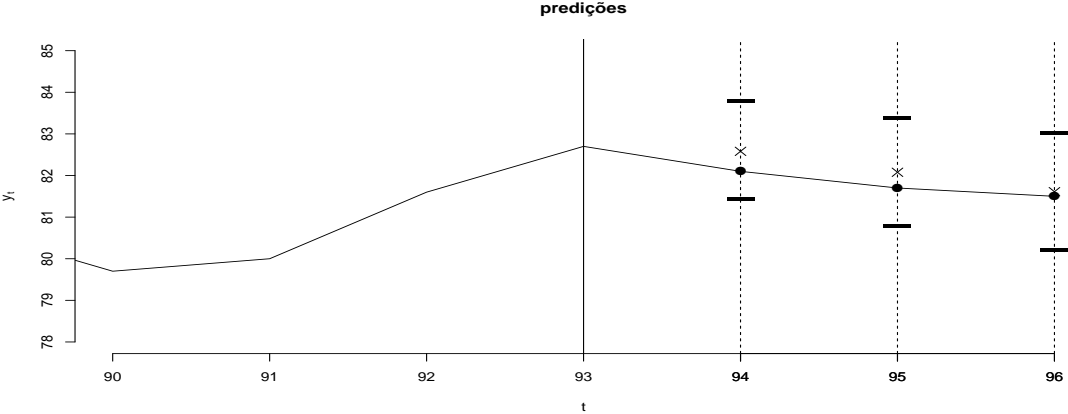


Figura 3.3: Valores preditos marcados com 'x'. Valores observados ponto cheio. Intervalo de credibilidade 95% - barra horizontal.

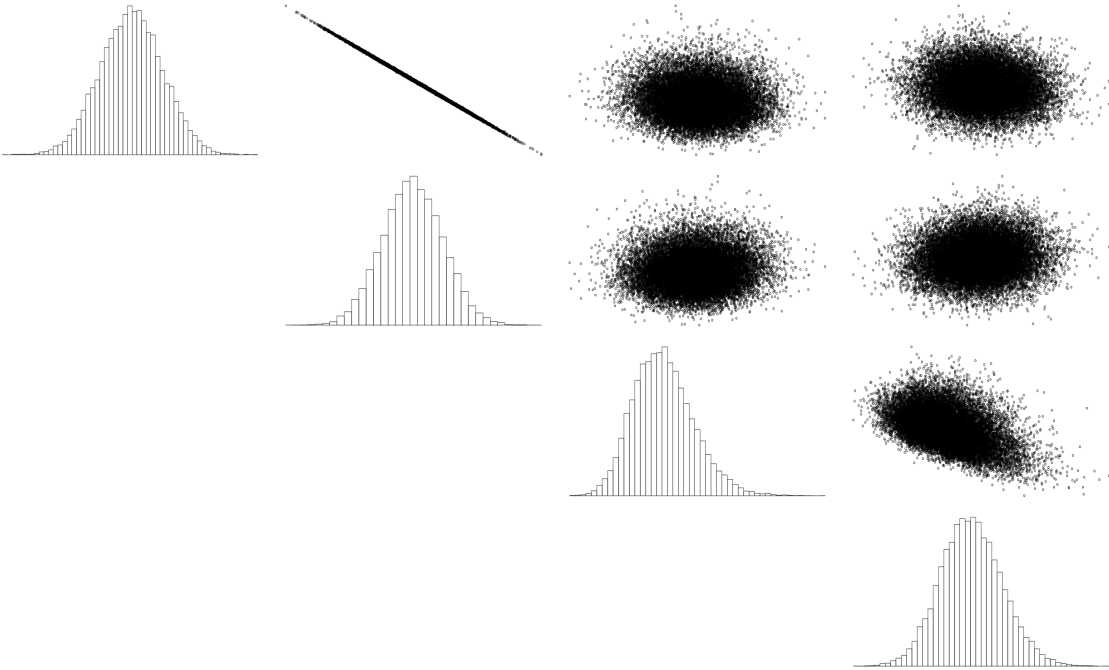


Figura 3.4: Histogramas e gráficos de dispersão respectivamente para μ , θ , σ , ξ .

4

Inferência Bayesiana não-paramétrica via processo Gaussiano

Modelos paramétricos são compostos por uma quantidade finita e fixa de parâmetros. A qualidade do ajuste deste tipo de modelos depende claramente de várias suposições, do tipo de função densidade observacional (comumente pertencente à família exponencial), função de regressão, função de ligação, etc. Até mesmo em modelos lineares generalizados é necessária suposições de funções que melhor se adaptam aos dados.

Além disso, problemas como falta de ajuste ou super ajustamento são inerentes a modelagem paramétrica quando existe uma discrepância significativa entre o verdadeiro mecanismo gerador desconhecido e o modelo escolhido.

No estudo não-paramétrico de modelos Bayesianos não há imposição de qualquer forma funcional paramétrica única subjacente aos dados. As observações permitem estimar a função ao invés de condicioná-la a uma determinada classe paramétrica de funções.

O *trade-off* entre falta de ajuste e a complexidade do modelo é automática. A inferência Bayesiana não-paramétrica irá tender naturalmente a representar uma forma simples para um modelo que melhor se adapta os dados.

Sob o enfoque Bayesiano qualquer tipo de função pode ser tratada como uma função aleatória. No entanto a construção da metodologia Bayesiana não-paramétrica requer a suposição de prioris sobre o espaço de funções e podem assumir diferentes tipos de processos estocásticos. Por exemplo, poderíamos supor um processo de Dirichlet sobre todas as possíveis funções densidade quando a especificação de uma função

de confiabilidade é tarefa difícil na modelagem de dados de sobrevivência (veja Ferguson 1973 e Müller & Quintana 2004). Num problema de regressão não-linear podemos supor um processo estocástico Gaussiano a priori sobre todas as possíveis funções de regressão, que abrangem funções lineares e não-lineares como em Rasmussem & Williams (2006).

Neste trabalho focamos no estudo de processos Gaussianos. Veremos como fazer inferência sobre um modelo Bayesiano baseado num processo Gaussiano com origens nos trabalhos de O'Hagan (1978) e Rasmussem & Williams (2006). Mostramos também como especificar um modelo não-paramétrico no estudo de valores extremos e os problemas computacionais relacionados a inferência do modelo.

4.1 Processo Gaussiano

Os modelos Gaussianos formam uma ampla e flexível classe de modelos estatísticos frequentemente utilizados em aplicações (ver por exemplo Rue & Held 2005 e Rasmussem & Williams 2006 para várias aplicações e outras referências). A utilização de tais modelos é atraente por várias razões. A especificação de sua função de distribuição é relativamente simples e tal modelo descreve razoavelmente vários fenômenos próprios da natureza. Além disso, as propriedades de marginalização e condicionamento são ideais e sua tratabilidade algébrica é relativamente simples, veja Rue & Held (2005).

Definição 4.1 *Uma função aleatória $f(t)$ é dita ser um processo Gaussiano, se qualquer coleção finita de valores da função (conjunto de v.a's) possuem distribuição Normal multivariada, i.e., para qualquer t_1, \dots, t_n e $\forall n \in \mathbb{N}$,*

$$\mathbf{f} \sim N_n[\mathbf{m}, K] \tag{4.1}$$

em que,

$$\mathbf{f} = \begin{bmatrix} f(t_1) \\ \vdots \\ f(t_n) \end{bmatrix}, \quad \mathbf{m} = \begin{bmatrix} m(t_1) \\ \vdots \\ m(t_n) \end{bmatrix} \quad e \quad K = \begin{bmatrix} k(t_1, t_1) & \cdots & k(t_1, t_n) \\ \vdots & \ddots & \vdots \\ k(t_n, t_1) & \cdots & k(t_n, t_n) \end{bmatrix}.$$

É comum usar a notação,

$$f(\cdot) \sim PG[m(\cdot), k(\cdot, \cdot)]. \tag{4.2}$$

As funções $m(\cdot)$ e $k(\cdot, \cdot)$ carregam propriedades importantes bem como especificam unicamente o processo. A função média do processo, $E[f(t)] = m(t)$ é uma função determinística e especifica o valor médio da função em um ponto fixo t . A função $k(t, t^*)$ é como a função anterior porém especifica o grau de dependência entre dois pontos distintos da função, i.e., $k(t, t^*) = Cov[f(t), f(t^*)] = E[(f(t) - m(t))(f(t^*) - m(t^*))]$

Exemplo 4.1 Considere um processo Gaussiano com função média e função de covariância respectivamente dadas por,

$$m(t) = \frac{1}{4}t^2 \quad e \quad k(t, t^*) = e^{-\frac{1}{2}(t-t^*)^2}. \quad (4.3)$$

Fixando um conjunto de pontos no eixo do tempo, $\mathbf{t} = (t_1, \dots, t_n)$, temos que,

$$\mathbf{f} = \begin{bmatrix} f(t_1) \\ \vdots \\ f(t_n) \end{bmatrix} \sim N_n[\mathbf{m}, K]$$

sendo,

$$\mathbf{m} = \begin{bmatrix} \frac{1}{4}t_1^2 \\ \vdots \\ \frac{1}{4}t_n^2 \end{bmatrix} \quad e \quad K = \begin{bmatrix} e^{-\frac{1}{2}(t_1-t_1)^2} & \dots & e^{-\frac{1}{2}(t_1-t_n)^2} \\ \vdots & \ddots & \vdots \\ e^{-\frac{1}{2}(t_n-t_1)^2} & \dots & e^{-\frac{1}{2}(t_n-t_n)^2} \end{bmatrix}.$$

Uma vez definido o processo em (4.3), algumas funções aleatórias foram obtidas a partir da geração de um vetor aleatório normalmente distribuído. A Figura 4.1 ilustra as funções geradas.

4.2 Funções de covariância

A função de covariância especifica o quão correlacionados estão dois valores distintos da função e sua escolha tem fundamental importância. Suas características irão determinar propriedades das funções amostrais que gostaríamos de estudar. Tais aspectos, além das propriedades de estacionariedade e não-estacionariedade, são controlados por alguns parâmetros que vamos denotar por ϕ .

Toda função de covariância que gera uma matriz positiva definida, i.e., $v^t K v \geq 0, \forall v \in \mathbb{R}^n$, pode ser considerada uma função válida para processos Gaussianos.

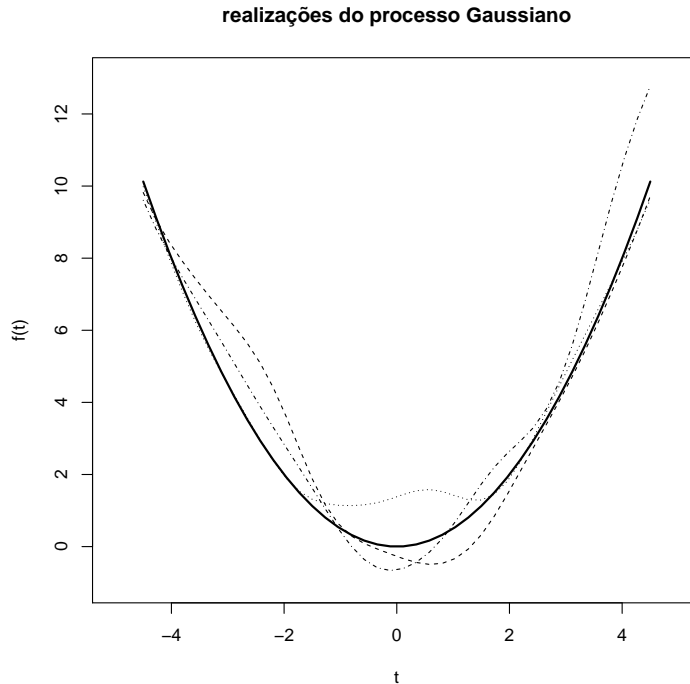


Figura 4.1: Realizações do processo Gaussiano. Linhas tracejadas são funções amostrais do processo. Linha cheia é a média do processo.

Mostraremos algumas funções comumente utilizadas em modelos Bayesianos não-paramétricos e algumas combinações destas que também geram funções de covariância válidas.

Vale notar que os tipos de funções aleatórias estudadas neste trabalho são funções suaves, i.e., funções que possuem derivadas de ordem infinita. Outros tipos de processo Gaussiano como o movimento Browniano não serão estudados aqui, pois as funções de covariância desses processo geram funções que não são deriváveis e envolvem um profundo estudo de cálculo estocástico.

Uma função de covariância muito utilizada é a função exponencial quadrática dada por,

$$k(t, t^*) = \phi_1 \exp \left\{ -\frac{1}{\phi_2} (t - t^*)^2 \right\}. \quad (4.4)$$

O parâmetro $\phi_1 > 0$ controla a variabilidade total do processo, já o parâmetro $\phi_2 > 0$ controla quão dependentes estão dois pontos distintos do processo. Note que, quando $\phi_2 \rightarrow \infty$, $k(.,.) \rightarrow \phi_1$ para $(t - t^*)^2 \neq 0$, tornando os valores da função extremamente dependentes. Caso contrário o processo não possui dependência.

A Figura 4.2 ilustra o comportamento dessa função para valores dos parâmetros

$\phi_1 = 1$ e $\phi_2 = 2$. Um outro tipo de função de covariância é a periódica, que gera

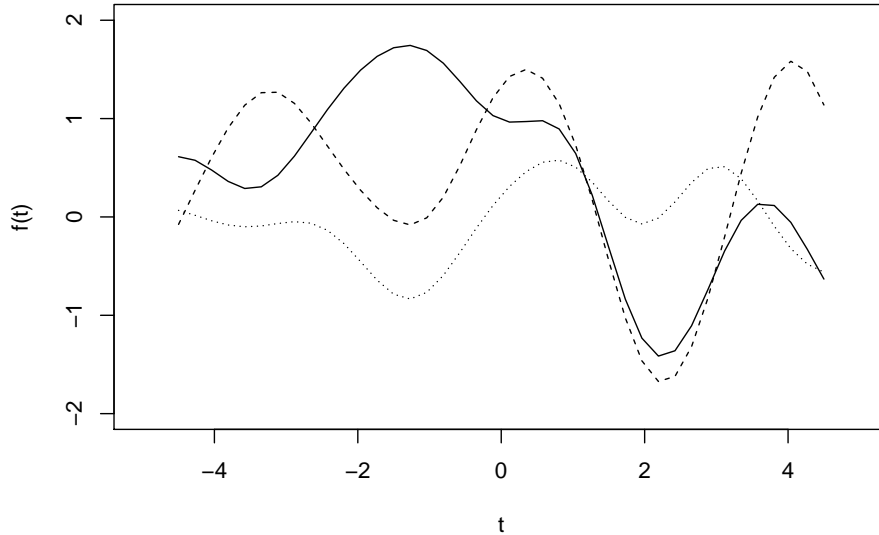


Figura 4.2: Realizações do processo Gaussiano com função exponencial quadrática. $\phi_1 = 1$ e $\phi_2 = 2$.

funções amostrais periódicas,

$$k(t, t^*) = \phi_1 \exp \left\{ -\frac{1}{\phi_2} \sin^2(\gamma|t - t^*|) \right\}. \quad (4.5)$$

Os parâmetros ϕ_1 e ϕ_2 possuem o mesmo papel da função anterior, porém o parâmetro $\gamma = \pi/\omega$ controla a periodicidade das funções amostrais. Numa outra parametrização, ω descreve exatamente o período das funções, veja figura 4.3. Ambas funções de covariância acima geram processos que são estacionários.

Uma função do tipo $k(t, t^*) = \phi_1(t - \phi_2)(t^* - \phi_2)$, gera processos não estacionários, i.e., as funções amostrais tendem a crescer ou decrescer com o aumento de t . Neste caso o parâmetro ϕ_2 indica onde as funções aleatórias tendem a tocar o eixo das abscissas.

Além disso, é possível combinar funções de covariância se somente uma função não é adequada na modelagem dos dados. Por exemplo, a soma de dois processos Gaussianos independentes implica que a função de covariância do processo Gaussiano resultante é dada pela soma das funções de covariância dos respectivos processos. Essas propriedades e outras podem ser revistas em Rasmussem & Williams (2006).

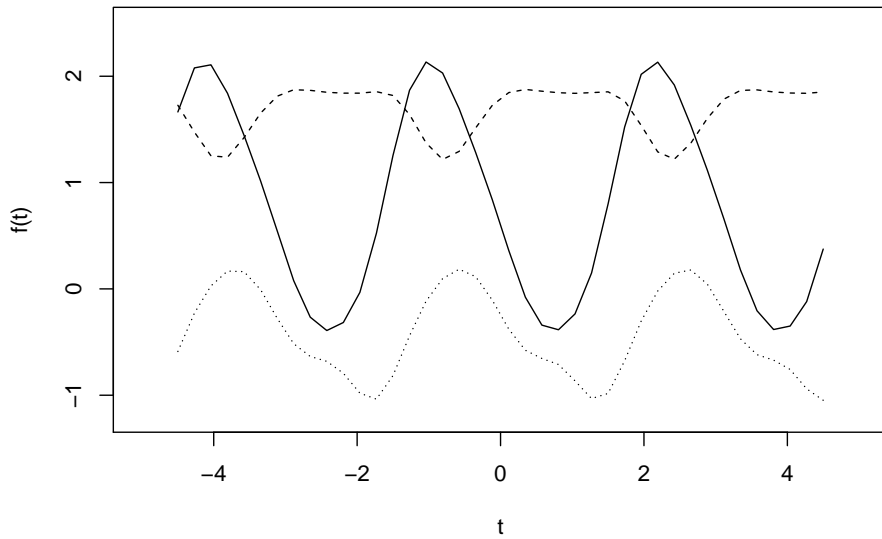


Figura 4.3: Realizações do processo Gaussiano com função periódica e parâmetros $\phi_1 = 1$, $\phi_2 = 2$ e $\gamma = 1$.

A função $m(t)$ também pode depender de parâmetros, digamos $\boldsymbol{\theta}$, porém em problemas de inferência é comum supor um processo com média nula, pois queremos neste tipo de problemas que os dados forneçam a forma da função ao invés impor uma determinada forma a priori.

Em muitos problemas práticos, as combinações lineares de funções de covariância mostram-se ideal para os mais variados comportamentos na modelagem de dados observados, veja Duvenaud (2014). Essa combinações são fundamentais uma vez que não trabalhamos a função média do processo, ou quando modelamos a média de alguma distribuição. Note que estudamos diretamente a dependência do processo.

4.3 Processo Gaussiano como representação a priori para $f(\cdot)$

Suponha que estamos interessados em estudar um modelo de regressão. Assuma também que a especificação de um modelo linear geral do tipo $\mathbf{y} = X\boldsymbol{\theta}$ não é razoável e a escolha de um modelo não linear nos parâmetros do tipo, $\mathbf{y} = f(\mathbf{x}; \boldsymbol{\theta})$ não é tarefa fácil. Neste caso podemos propor o modelo Bayesiano não-paramétrico atribuindo um

processo Gaussiano a priori para a função desconhecida $f(\cdot)$, i.e.,

$$\begin{aligned} y &= f(x) + e \\ f(x) &\sim PG[m(x), k(x, x^*)] \\ e &\sim N[0, \sigma^2]. \end{aligned} \tag{4.6}$$

Uma vez obtido um conjunto de dados de tamanho n e pela Definição 4.1 podemos escrever o modelo como,

$$\begin{aligned} \mathbf{y}|\mathbf{f} &\sim N_n[\mathbf{f}, I\sigma^2] \\ \mathbf{f}|\mathbf{m}, K &\sim N_n[\mathbf{m}, K]. \end{aligned} \tag{4.7}$$

Note que, por construção, não observamos diretamente os valores da função mas seu valor mais um erro. O modelo agora possui um nível maior de hierarquia e logo pode ser visto como um modelo latente com os valores desconhecidos da função sendo parâmetros a serem estimados. É neste sentido que o modelo Bayesiano não-paramétrico é visto como um modelo de dimensões infinitas, uma vez que sua dimensão aumenta de acordo com o tamanho da amostra.

Para completar o modelo Bayesiano precisamos assumir distribuições a priori para os hiperparâmetros, tanto na função de covariância como para os parâmetros da função média. Comumente serão tomados hiperparâmetros independentes a priori com distribuições difusas (variância grande). Ambas suposições é devido ao alto nível de hierarquia. Logo,

$$\begin{aligned} \mathbf{y}|\mathbf{f}, \sigma^2 &\sim N_n[\mathbf{f}, I\sigma^2] \\ \mathbf{f}|\boldsymbol{\theta}, \boldsymbol{\phi} &\sim N_n[\mathbf{m}, K] \\ \sigma^2, \boldsymbol{\theta}, \boldsymbol{\phi} &\sim \pi(\sigma^2, \boldsymbol{\theta}, \boldsymbol{\phi}). \end{aligned} \tag{4.8}$$

Aplicando o teorema de Bayes, temos que a posteriori conjunta do processo latente e dos hiperparâmetros é dada por,

$$\pi(\mathbf{f}, \sigma^2, \boldsymbol{\theta}, \boldsymbol{\phi}|D) \propto N_n[\mathbf{y}|\mathbf{f}, I\sigma^2]N_n[\mathbf{f}|\mathbf{m}, K]\pi(\sigma^2, \boldsymbol{\theta}, \boldsymbol{\phi}). \tag{4.9}$$

Observe que a posteriori (4.9) possui dimensão maior que o tamanho da amostra e não possui forma fechada. Para obter resumos (média, variância, ...) da distribuição a posteriori é necessário a implementação do algoritmo MCMC para gerar valores

de (4.9). Isto é algo de difícil configuração pois a posteriori tem dimensão elevada e possui forte estrutura de correlação (veja Calderhead & Girolami 2011). No entanto podemos tomar outra direção. A posteriori (4.9) é o produto de duas distribuições normais multivariadas e pela fórmula de Sherman-Morrison-Woodbury-Schur, pode ser reescrita na seguinte forma,

$$\pi(\mathbf{f}, \sigma^2, \boldsymbol{\theta}, \boldsymbol{\phi} | D) \propto N_n[\mathbf{y} | \mathbf{m}, K + I\sigma^2] N_n[\mathbf{f} | \mathbf{c}, C] \pi(\sigma^2, \boldsymbol{\theta}, \boldsymbol{\phi}), \quad (4.10)$$

em que,

$$\begin{aligned} \mathbf{c} &= \mathbf{m} + K[K + I\sigma^2]^{-1}(\mathbf{y} - \mathbf{m}) \\ \mathbf{C} &= K - K[K + I\sigma^2]^{-1}K. \end{aligned} \quad (4.11)$$

Agora podemos integrar (4.10) em relação ao processo latente e obter distribuição marginal dos hiperparâmetros em forma fechada, pois a integral em relação ao processo latente é igual a 1. Logo

$$\pi(\sigma^2, \boldsymbol{\theta}, \boldsymbol{\phi} | D) \propto N_n[\mathbf{y} | \mathbf{m}, K + I\sigma^2] \pi(\sigma^2, \boldsymbol{\theta}, \boldsymbol{\phi}). \quad (4.12)$$

Deste modo, podemos obter de modo menos trabalhoso valores da posteriori marginal (4.12) via algum método de simulação estocástica e obter os valores do processo latente a posteriori. Através do método da composição (veja Tanner (1996)) escrevemos na seguinte forma,

$$\begin{aligned} \pi(\mathbf{f} | D) &= \int \pi(\mathbf{f}, \sigma^2, \boldsymbol{\theta}, \boldsymbol{\phi} | D) d(\sigma^2, \boldsymbol{\theta}, \boldsymbol{\phi}) \\ &= \int \pi(\mathbf{f} | \sigma^2, \boldsymbol{\theta}, \boldsymbol{\phi}, D) \pi(\sigma^2, \boldsymbol{\theta}, \boldsymbol{\phi} | D) d(\sigma^2, \boldsymbol{\theta}, \boldsymbol{\phi}) \\ &= E_{\sigma^2, \boldsymbol{\theta}, \boldsymbol{\phi} | D} \left[\pi(\mathbf{f} | \sigma^2, \boldsymbol{\theta}, \boldsymbol{\phi}, D) \right] \\ &= E_{\sigma^2, \boldsymbol{\theta}, \boldsymbol{\phi} | D} \left[N_n(\mathbf{f} | \mathbf{c}, C) \right]. \end{aligned} \quad (4.13)$$

A última passagem de (4.13) é através dos mesmos motivos de (4.9), porém com distribuição a posteriori do processo latente condicionada nos hiperparâmetros e nos dados observados.

Uma parte importante da inferência Bayesiana é avaliar se o modelo proposto fornece boas predições sobre valores futuros. Para determinar a distribuição preditiva de um vetor de valores futuros, digamos $\mathbf{y}_* | D$, é necessário condicionar o vetor \mathbf{y}_*

sobre valores futuros do processo Gaussiano \mathbf{f}_* e marginalizar para obter distribuição preditiva desejada, i.e.,

$$\begin{aligned}\pi(\mathbf{y}_*|D) &= \int \pi(\mathbf{y}_*|\mathbf{f}_*, \sigma^2)\pi(\mathbf{f}_*, \sigma^2|D)d(\mathbf{f}_*, \sigma^2) \\ &= E_{\mathbf{f}_*, \sigma^2|D} \left[N(\mathbf{y}_*|\mathbf{f}_*, I\sigma^2) \right].\end{aligned}\quad (4.14)$$

Note ainda que é necessário obter valores de $\mathbf{f}_*|D$. Pela propriedade de distribuição condicional normal da distribuição normal multivariada, segue que,

$$\begin{aligned}\pi(\mathbf{f}_*|D) &= \int N(\mathbf{f}_*|\mathbf{f}, \boldsymbol{\theta}, \boldsymbol{\phi})\pi(\mathbf{f}, \boldsymbol{\theta}, \boldsymbol{\phi}|D)d(\mathbf{f}, \boldsymbol{\theta}, \boldsymbol{\phi}) \\ &= E_{\mathbf{f}, \boldsymbol{\theta}, \boldsymbol{\phi}|D} \left[N(\mathbf{f}_*|\mathbf{c}_*, C_*) \right]\end{aligned}\quad (4.15)$$

com,

$$\begin{aligned}\mathbf{c}_* &= \mathbf{m}_* + K_*K^{-1}(\mathbf{f} - \mathbf{m}) \\ C_* &= K_{**} - K_*K^{-1}K_*^t.\end{aligned}\quad (4.16)$$

Exemplo 4.2 Considere um conjunto de $n = 50$ pontos (figura 4.4) simulados a partir da função não-linear (4.17) com erro aditivo normal e valores na abscissa espaçados aleatoriamente no intervalo $(0, 10)$.

$$\begin{aligned}y_i &= x_i + 2 \sin(0.5\pi x_i) + 4 + e_i \\ e_i &\sim N[0, 0.25].\end{aligned}\quad (4.17)$$

Em alguns problemas práticos especificar previamente a forma funcional da função regressora pode ser algo extremamente difícil. Suponha a observação do conjunto de dados na figura 4.4 sem conhecer (4.17). Notamos que os dados da figura 4.4 indicam tendência com sazonalidade. Desde modo, supondo um vago estado de conhecimento sobre a forma funcional da regressão, assumo o modelo Bayesiano não-paramétrico com função média nula e função de covariância dada por,

$$K = \phi_1(x - \phi_2)(x^* - \phi_2) + \phi_3 \exp \left[-\frac{1}{\phi_5} \sin^2(\phi_4|x - x^*|) \right]. \quad (4.18)$$

Deste modo as equações (4.12) e (4.13) se tornam,

$$\begin{aligned}\pi(\sigma^2, \boldsymbol{\phi}|D) &\propto N_n[\mathbf{y}|\mathbf{0}, K + I\sigma^2]\pi(\sigma^2, \boldsymbol{\phi}) \\ \pi(\mathbf{f}|D) &= E_{\sigma^2, \boldsymbol{\phi}|D} \left[N_n(\mathbf{f}|\mathbf{c}, C) \right],\end{aligned}\quad (4.19)$$

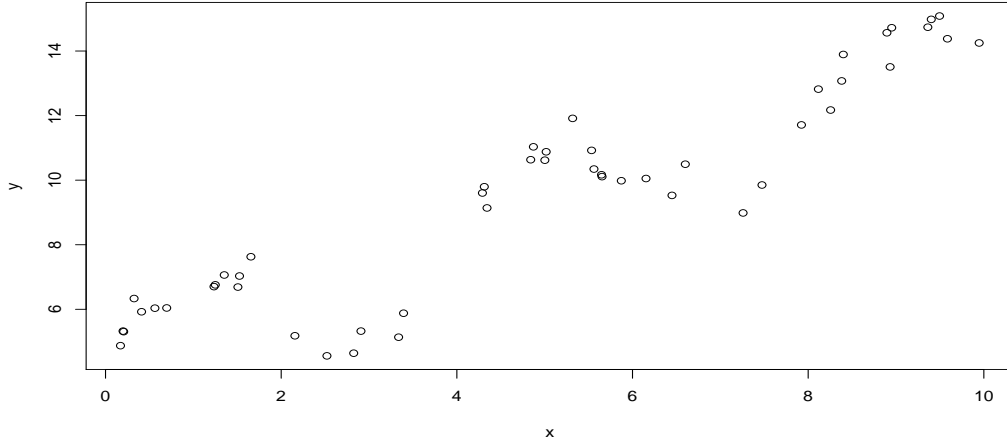


Figura 4.4: Pontos amostrados.

em que, $\boldsymbol{\phi} = [\phi_1 \ \phi_2 \ \phi_3 \ \phi_4 \ \phi_5]'$,

$$\begin{aligned} \mathbf{c} &= K[K + I\sigma^2]^{-1}\mathbf{y} \\ \mathbf{C} &= K - K[K + I\sigma^2]^{-1}K, \end{aligned} \quad (4.20)$$

com parâmetros independentes a priori e distribuições,

$$\begin{aligned} \sigma^2 &\sim IG(0.1, 0.1) \\ \phi_{-2} &\sim IG(0.1, 0.1) \\ \phi_2 &\sim N(0, 50). \end{aligned} \quad (4.21)$$

A distribuição preditiva se simplifica para,

$$\begin{aligned} \pi(\mathbf{y}_*|D) &= E_{\mathbf{f}_*, \sigma^2|D} \left[N(\mathbf{y}_*|\mathbf{f}_*, I\sigma^2) \right] \\ \pi(\mathbf{f}_*|D) &= E_{\mathbf{f}, \phi|D} \left[N(\mathbf{f}_*|\mathbf{c}_*, C_*) \right]. \end{aligned} \quad (4.22)$$

com,

$$\begin{aligned} \mathbf{c}_* &= K_*K^{-1}\mathbf{f} \\ C_* &= K_{**} - K_*K^{-1}K_*^t. \end{aligned} \quad (4.23)$$

Note que a expressão (4.19) tem forma analítica de difícil manipulação devido ao vetor de hiperparâmetros $\boldsymbol{\phi}$ serem parte da função de covariância. Para obter valores de (4.19) fazemos uso do algoritmo RMHMC com métrica fixa, sendo necessário determinar vetor gradiente da função log-posteriori com reparametrização na

reta e obter a métrica avaliada no MAP. Deste modo, tomamos a reparametrização, $[\sigma^2 \phi_1 \phi_2 \phi_3 \phi_4 \phi_5]' = [\exp(\delta_1) \exp(\delta_2) \delta_3 \exp(\delta_4) \exp(\delta_5) \exp(\delta_6)]'$ e obtemos as derivadas parciais em relação aos hiperparâmetros da função de covariância dadas por,

$$\nabla_{\boldsymbol{\delta}} \ell = \begin{bmatrix} \frac{e^{\delta_1}}{2} \text{tr} \left((K + Ie^{\delta_1})^{-1} [\mathbf{y}\mathbf{y}^t (K + Ie^{\delta_1})^{-1} - I] \right) + \frac{\partial \log \pi(\delta_1)}{\partial \delta_1} \\ \frac{1}{2} \text{tr} \left((K + Ie^{\delta_1})^{-1} [\mathbf{y}\mathbf{y}^t (K + Ie^{\delta_1})^{-1} - I] \frac{\partial K}{\partial \delta_2} \right) + \frac{\partial \log \pi(\delta_2)}{\partial \delta_2} \\ \vdots \\ \frac{1}{2} \text{tr} \left((K + Ie^{\delta_1})^{-1} [\mathbf{y}\mathbf{y}^t (K + Ie^{\delta_1})^{-1} - I] \frac{\partial K}{\partial \delta_5} \right) + \frac{\partial \log \pi(\delta_5)}{\partial \delta_5} \end{bmatrix}, \quad (4.24)$$

com as derivadas da função de covariância com respeito a cada hiperparâmetro dadas por

$$\begin{aligned} \frac{\partial K}{\partial \delta_2} &= \left\{ \frac{\partial K}{\partial \delta_2} \right\}_{ij} = \phi_1 [x_i x_j - (x_i + x_j) \phi_2 + \phi_2] \\ \frac{\partial K}{\partial \delta_3} &= \left\{ \frac{\partial K}{\partial \delta_3} \right\}_{ij} = -\phi_1 [(x_i + x_j) - 2\phi_2] \\ \frac{\partial K}{\partial \delta_4} &= \left\{ \frac{\partial K}{\partial \delta_4} \right\}_{ij} = \phi_3 \exp \left[-\frac{1}{\phi_5} \sin^2(\phi_4 |x_i - x_j|) \right] \\ \frac{\partial K}{\partial \delta_5} &= \left\{ \frac{\partial K}{\partial \delta_5} \right\}_{ij} = -\frac{2\phi_4}{\phi_5} \sin(\phi_4 |x_i - x_j|) \cos(\phi_4 |x_i - x_j|) |x_i - x_j| \\ &\quad \times \phi_3 \exp \left[-\frac{1}{\phi_5} \sin^2(\phi_4 |x_i - x_j|) \right] \\ \frac{\partial K}{\partial \delta_6} &= \left\{ \frac{\partial K}{\partial \delta_6} \right\}_{ij} = \frac{\phi_3 \sin^2(\phi_4 |x_i - x_j|)}{\phi_5} \exp \left[-\frac{1}{\phi_5} \sin^2(\phi_4 |x_i - x_j|) \right], \end{aligned} \quad (4.25)$$

e os elementos da matriz de informação de Fisher dados por

$$G_{kl} = \frac{1}{2} \text{tr} \left[(K + Ie^{\delta_1})^{-1} \frac{\partial (K + Ie^{\delta_1})}{\partial \delta_k} (K + Ie^{\delta_1})^{-1} \frac{\partial (K + Ie^{\delta_1})}{\partial \delta_l} \right], \quad (4.26)$$

para $i, j = 1, \dots, n$ e $k, l = 1, \dots, 6$. Mostramos estes resultados e sua provas no apêndice, pois são resultados que não aparecem facilmente na literatura de modelos estatísticos.

Primeiramente determinamos a estimativa MAP e tomamos a métrica fixa como $M = G(\boldsymbol{\delta}_{\text{MAP}})$. Após algumas gerações para adequar os parâmetros livres e obter boa

	σ^2	ϕ_1	ϕ_2	ϕ_3	ϕ_4	ϕ_5
$\widehat{E}[\cdot D]$	0.29	14.59	-2.25	406.79	0.78	155.37
$\widehat{sd}[\cdot D]$	0.06	46.48	4.95	1320.95	0.01	478.41
$\widehat{\text{moda}}$	0.27	0.42	-3.88	12.59	0.78	3.45
$\widehat{\text{med}}$	0.28	1.87	-2.84	39.66	0.78	20.05
<i>IC</i> 95%	[0.20, 0.40]	[0.27, 68.55]	[-9.68, 6.59]	[2.20, 2017.90]	[0.76, 0.80]	[1.40, 739.58]

Tabela 4.1: Resumos de interesse dos hiperparâmetros a posteriori (4.19)

convergência do algoritmo, tomamos $\varepsilon = 0.1$ e replicamos a solução de Störmer-Verlet quatorze vezes. Com estes valores geramos uma cadeia de tamanho 22000 com burn-in de 1000 sem tomar valores defasados.

Notamos que as distribuições marginais a posteriori dos hiperparâmetros são extremamente assimétrica, deste modo, tomamos a moda a posteriori como estimativa dos hiperparâmetros, veja figura 4.5.

A variância das funções aleatórias representada pela soma dos hiperparâmetros, $\hat{\phi}_1$ e $\hat{\phi}_3$ fica em torno de 13 e o ruído bem próximo do valor verdadeiro $\hat{\sigma}^2 = 0.26$. O parâmetro $\hat{\phi}_2$, indica que as funções aleatórias tendem a cortar o eixo das abscissas em -3.88 . A periodicidade das funções fica em torno de $\hat{\omega} = \pi/\hat{\phi}_5 = 4.02$. Os resultados podem ser vistos na tabela 4.1.

O valor esperado do processo latente para cada valor na abscissa gerado é calculado através de (4.19). A figura 4.6 esboça o valor esperado da função nos valores observados da abscissa junto com os intervalos de credibilidade. As cruzez ilustram os pontos simulados.

Na determinação da função preditiva esperada (aproximadamente a média de todas as funções geradas), varremos um intervalo $(0, 20)$ com 300 pontos e usamos as equações preditivas (4.22) na forma aproximada via Monte Carlo. Uma vez obtido os valores da distribuição marginal $\mathbf{f}_*|D$ através do método da composição, fazemos,

$$\begin{aligned}
\hat{\mathbf{y}}_* &= E[\mathbf{y}_*|D] \\
&= E\left[E[\mathbf{y}_*|\mathbf{f}_*, I\sigma^2]\right] \\
&\cong \frac{1}{N} \sum_{i=1}^N \mathbf{y}_*^{(i)}|\mathbf{f}_*^{(i)}, \sigma^{2(i)}.
\end{aligned} \tag{4.27}$$

A estimativa da função fica também bem representada tanto na região de variação

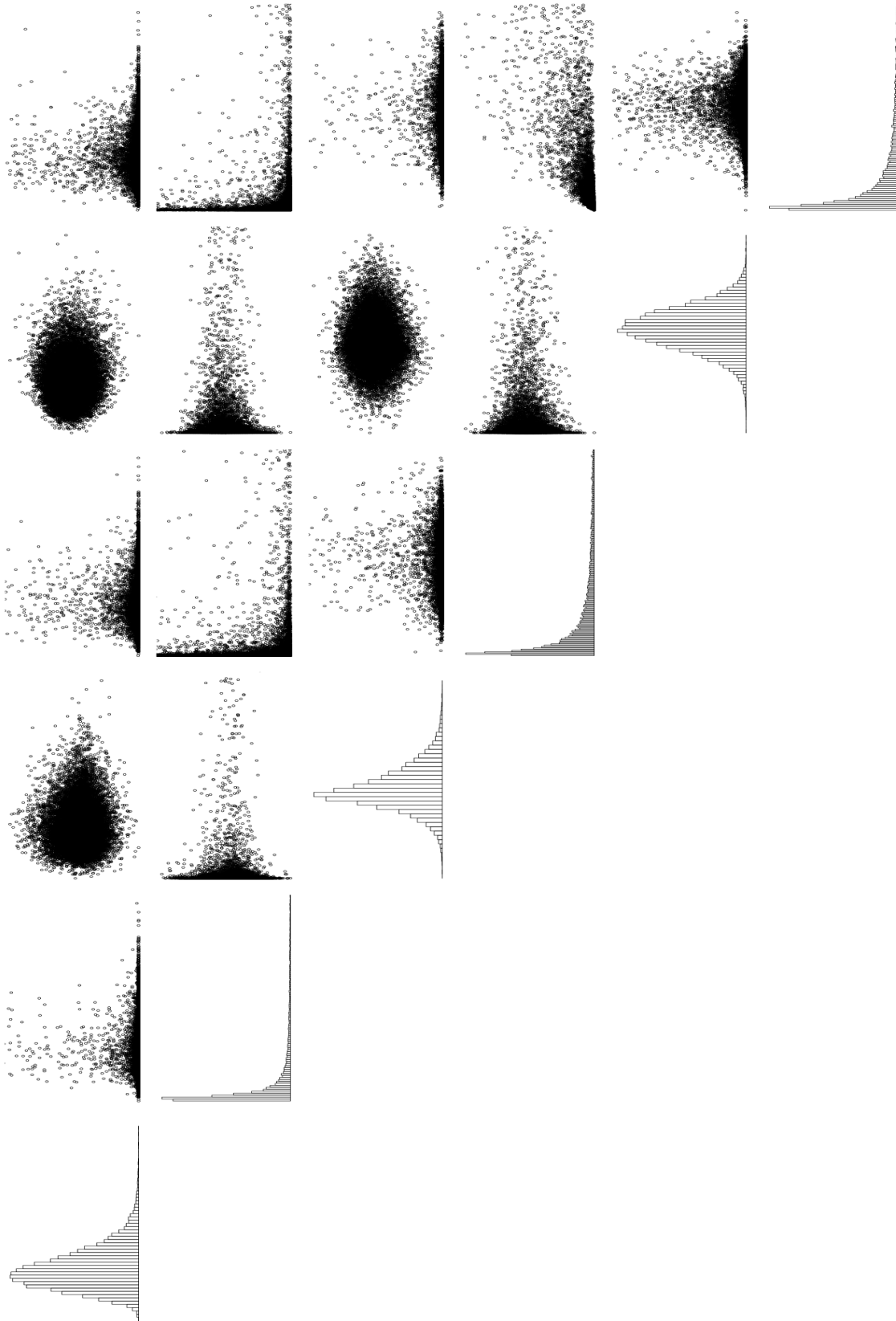


Figura 4.5: Histogramas (distribuições marginais) e gráficos de dispersão (distribuições marginais biviariadas) para os hiperparâmetros σ^2 , ϕ_1 , ϕ_2 , ϕ_3 , ϕ_4 , ϕ_5 respectivamente.

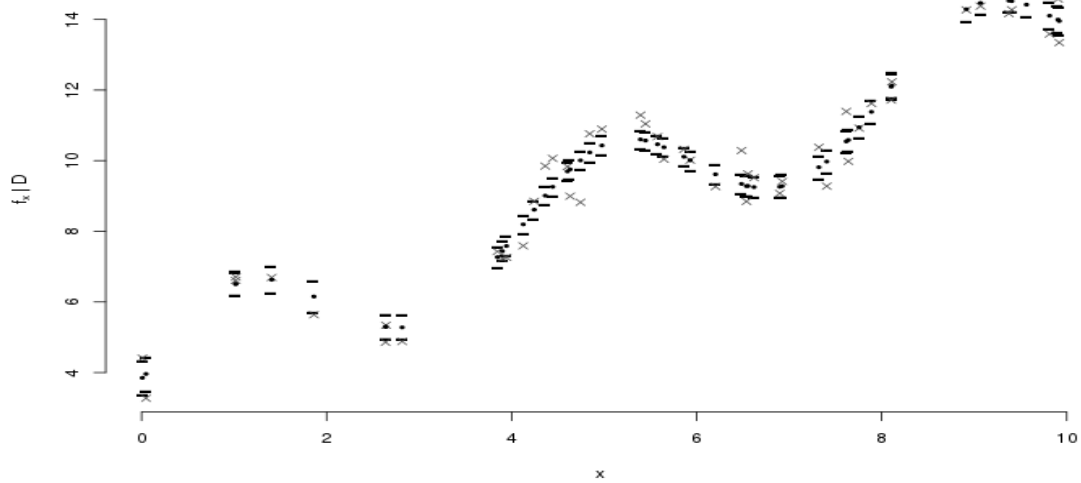


Figura 4.6: Estimativa (média) da valor da função nas abscissas simuladas

dos dados como fora desta. O processo Gaussiano latente captura de forma bastante consistente a verdadeira forma da função somente com a observação dos dados sem a imposição de qualquer função média no processo. Em regiões distantes da região de variação dos dados, a qualidade das previsões é adequada mesmo sem qualquer informação sobre a forma média do processo.

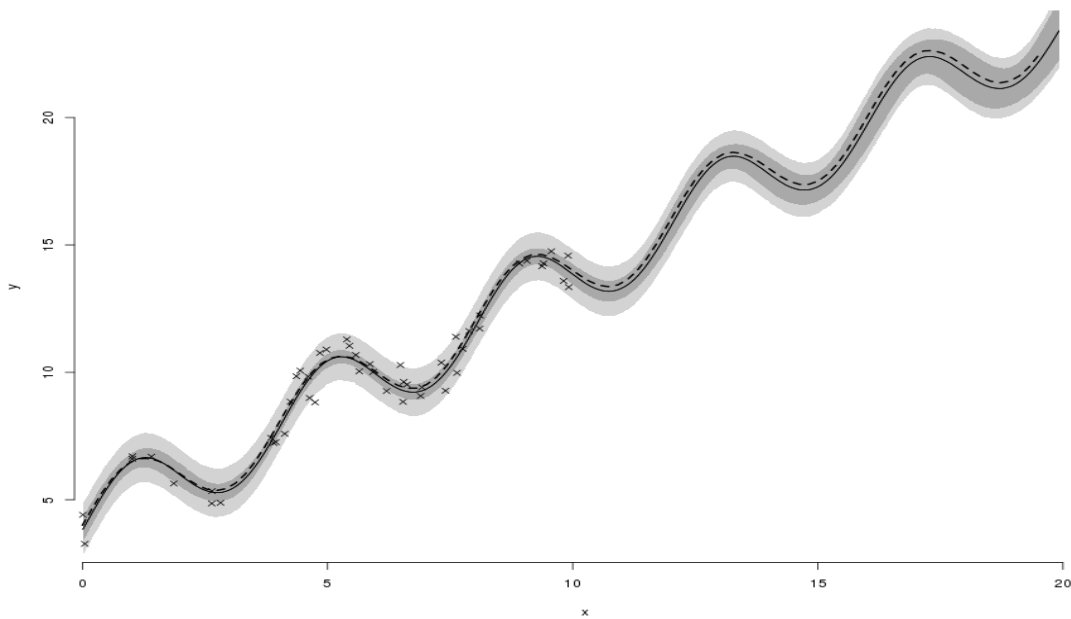


Figura 4.7: previsão (média) da função desconhecida.

O gráfico 4.7 ilustra a predição (média) da função desconhecida. A linha sólida representa a função preditiva média de $\mathbf{y}_*|D$ e de $\mathbf{f}_*|D$ que coincidem. A linha tracejada é a função verdadeira dada por (4.17). A região cinza claro ilustra a região de credibilidade 95% de $\mathbf{y}_*|D$. A região cinza escura ilustra a região de credibilidade 95% de $\mathbf{f}_*|D$. As cruces representam os pontos observados. Notar que a região de credibilidade aumenta vagarosamente. Isto é devido a seleção correta da combinação linear das funções de covariância para a representação dos dados observados.

4.4 Inferência Bayesiana não-paramétrica em valores extremos

Como visto na seção anterior, a tratabilidade analítica do modelo Bayesiano não-paramétrico é ideal sob a suposição de normalidade dos dados. Porém, em muitas situações práticas, há forte evidência que a normalidade não se adequa em determinados comportamentos, por exemplo, em dados de valores extremos. Nestas situações, as distribuições passam a apresentar forte assimetria concomitante a caudas pesadas.

Para introduzir o modelo Bayesiano não-paramétrico, na seção de modelagem paramétrica de valores extremos, o parâmetro de locação μ era tomado como algum tipo de função regressora ou um modelo de série temporal, neste mesmo sentido, tome $\mu = f(x)$ com uma função desconhecida, deste modo segue que, obtido um conjunto de dados de tamanho n e pela definição 4.1 escrevemos,

$$\begin{aligned} \mathbf{y}|\mathbf{f}, \boldsymbol{\theta} &\sim GEV_n(\mathbf{f}, \boldsymbol{\theta}) \\ \mathbf{f}|\boldsymbol{\phi} &\sim N_n(\mathbf{m}, K), \\ \boldsymbol{\theta}, \boldsymbol{\phi} &\sim \pi(\boldsymbol{\theta}, \boldsymbol{\phi}) \end{aligned} \tag{4.28}$$

em que $\boldsymbol{\theta} = [\sigma \ \xi]'$. Pelo teorema de Bayes segue que,

$$\pi(\mathbf{f}, \boldsymbol{\theta}, \boldsymbol{\phi}|D) \propto L(\mathbf{y}|\mathbf{f}, \boldsymbol{\theta})N_n(\mathbf{f}|\boldsymbol{\phi})\pi(\boldsymbol{\theta}, \boldsymbol{\phi}), \tag{4.29}$$

em que $L(\mathbf{y}|\mathbf{f}, \boldsymbol{\theta}) = \prod_{i=1}^n GEV(y_i|f_i, \sigma, \xi)$. Note que (4.29) é analiticamente intratável, não sendo possível proceder como no caso de normalidade observacional. Deste modo, usaremos o algoritmo RMHMC para simular valores de (4.29) e obter resumos a posteriori de interesse. Para proceder com o algoritmo é necessário reparametrizar

todos os parâmetros na reta, assim, tomamos $\sigma = \exp(\delta_1)$, $\boldsymbol{\phi} = \exp(\boldsymbol{\delta})$ para os parâmetros da função de covariância e determinamos o vetor gradiente da função log-posteriori de (4.29),

$$\nabla \ell = \begin{bmatrix} \nabla_{\mathbf{f}} \log L(\mathbf{y}|\mathbf{f}, \boldsymbol{\theta}) + K^{-1}(\mathbf{m} - \mathbf{f}) \\ \nabla_{\delta_1} \log L(\mathbf{y}|\mathbf{f}, \boldsymbol{\theta}) + \nabla_{\delta_1} \log \pi(\delta_1) \\ \nabla_{\xi} \log L(\mathbf{y}|\mathbf{f}, \boldsymbol{\theta}) + \nabla_{\xi} \log \pi(\xi) \\ \nabla_{\boldsymbol{\delta}} \log N[\mathbf{f}|\mathbf{m}, K] + \nabla_{\boldsymbol{\delta}} \log \pi(\boldsymbol{\delta}) \end{bmatrix} \quad (4.30)$$

em que

$$\nabla_{f_i} L(\mathbf{y}|\mathbf{f}, \boldsymbol{\theta}) = \frac{1}{\sigma} z_i^{-1} \left((1 + \xi) - z_i^{-1/\xi} \right) \quad (4.31)$$

para $i = 1, \dots, n$.

$$\nabla_{\delta_1} \log L(\mathbf{y}|\mathbf{f}, \boldsymbol{\theta}) = \sum_{i=1}^n (1 + \xi) \left(\frac{y_i - f_i}{\sigma} \right) z_i^{-1} - 1 - z_i^{-(1/\xi+1)} \left(\frac{y_i - f_i}{\sigma} \right) \quad (4.32)$$

$$\begin{aligned} \nabla_{\xi} \log L(\mathbf{y}|\mathbf{f}, \boldsymbol{\theta}) &= \sum_{i=1}^n \frac{\log z_i}{\xi^2} - \left(\frac{1}{\xi} + 1 \right) \left(\frac{y_i - f_i}{\sigma} \right) z_i^{-1} + \frac{1}{\xi} \left(\frac{y_i - f_i}{\sigma} \right) z_i^{-(1/\xi+1)} \\ &\quad - \frac{\log z_i}{\xi^2} z_i^{-1/\xi} \end{aligned} \quad (4.33)$$

$$\nabla_{\delta_k} \log N[\mathbf{f}|\mathbf{m}, K] = \frac{1}{2} \text{tr} \left[K^{-1}[(\mathbf{f} - \mathbf{m})(\mathbf{f} - \mathbf{m})' K^{-1} - I] \frac{\partial K}{\partial \delta_k} \right] \quad (4.34)$$

com, $z_i = 1 + \xi \left(\frac{y_i - f_i}{\sigma} \right)$, para $i = 1, \dots, n$ e $k = 2, \dots, p$.

Note ainda que é necessário obter uma métrica G considerando a estrutura hierárquica do modelo. Na seção de discussão de Calderhead & Girolami (2011), página 191, Ian Murray e Ryan Prescott Adams propõem uma métrica para modelos hierárquicos dada por,

$$\begin{aligned} G(\mathbf{f}, \boldsymbol{\theta}, \boldsymbol{\delta}) &= - E_{\mathbf{y}|\mathbf{f}, \boldsymbol{\theta}}(\nabla \nabla \ell) \\ &\quad - E_{\mathbf{f}|\boldsymbol{\delta}}[\nabla \nabla \log N_n(\mathbf{m}, K)] \\ &\quad - \nabla \nabla \log \pi(\boldsymbol{\theta}, \boldsymbol{\delta}) \end{aligned} \quad (4.35)$$

que gera uma matriz G positiva definida. O primeiro elemento de (4.35) é dado pela matriz informação de Fisher para o modelo observacional. O segundo elemento é também a matriz de informação para o processo latente. O último elemento é a matriz de derivadas segundas cruzadas das funções log-priori.

A desvantagem dessa métrica está no fato de não levar em consideração a possível covariância entre os parâmetros do modelo observacional e os parâmetros do processo

Gaussiano latente, não sendo então uma métrica ideal para a configuração ótima do algoritmo RMHMC. Esta métrica G será usada no decorrer do trabalho uma vez que não há outras propostas na literatura para este modelo específico e é um problema em aberto para qualquer modelo com variáveis latentes.

4.5 Exemplo simulado com dados faltantes

Para uma aplicação simulada ao modelo proposto, vamos considerar inicialmente a distribuição valor extremo tipo I, que é um caso particular da distribuição GEV quando $\xi \rightarrow 0$. Dizemos que uma variável aleatória Y segue uma distribuição de valor extremos tipo I quando sua densidade é dada por,

$$\pi(y|\mu, \sigma) = \frac{1}{\sigma} \exp \left[-\frac{y - \mu}{\sigma} - \exp \left(-\frac{y - \mu}{\sigma} \right) \right]. \quad (4.36)$$

Considere um cenário similar ao exemplo 4.2, i.e., assuma agora que o ruído segue uma distribuição valor extremo tipo, escrevemos,

$$\begin{aligned} y_i &= t_i + 1.5 \sin(0.5\pi t_i) + 2 + e_i \\ e_i &\sim EV(0, 0.5) \\ y_i &\sim EV(\mu_{t_i}, 0.5), \quad \mu(t_i) = t_i + 1.5 \sin(0.5\pi t_i) + 2. \end{aligned} \quad (4.37)$$

Neste caso, geramos $n = 100$ valores simulados e alocamos 50 valores aleatoriamente no intervalo $(-7, 3)$ e o restante no intervalo $(7, 13)$. A figura 4.8 ilustra os pontos observados. Note que há uma intervalo na abcissa, entre $(3, 13)$, que não exibe observações.

Assuma o modelo da seção anterior com a distribuição valor extremos tipo I e considere o processo Gaussiano latente para $\mu_t = f(t)$, com função média nula e função de covariância dada por (4.18). Assumindo também que os hiperparâmetros são independentes a priori com distribuições vagas, pelo teorema de Bayes segue que,

$$\pi(\mathbf{f}, \sigma, \boldsymbol{\phi}|D) \propto L(\mathbf{f}, \sigma|\mathbf{y}) N_T(\mathbf{f}|\boldsymbol{\phi}) \pi(\sigma) \pi(\boldsymbol{\phi}), \quad (4.38)$$

com,

$$\begin{aligned} \sigma &\sim IG(0.1, 0.1) \\ \boldsymbol{\phi}_{-2} &\sim IG(0.1, 0.1) \\ \phi_2 &\sim \log IG(0.1, 0.1). \end{aligned} \quad (4.39)$$

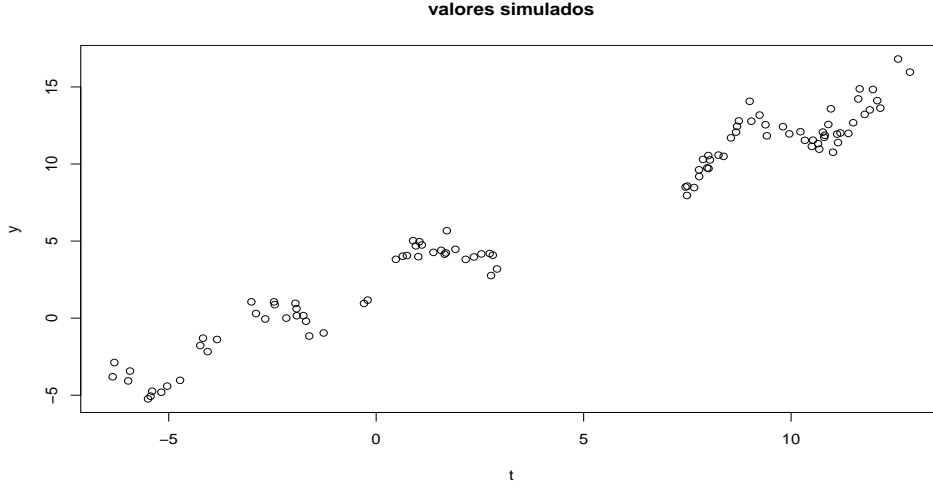


Figura 4.8: Pontos amostrados.

Note que a posteriori é 106-dimensional. Para proceder com o algoritmo RMHMC, fazemos $[\sigma \phi_1 \phi_2 \phi_3 \phi_4 \phi_5]' = [\exp(\delta_1) \exp(\delta_2) \delta_3 \exp(\delta_4) \exp(\delta_5) \exp(\delta_6)]'$ e determinamos o vetor gradiente da função log-posteriori ($\ell = \log \pi(\mathbf{f}, \boldsymbol{\delta} | D)$) nesta parametrização, i.e.,

$$\nabla \ell = \begin{bmatrix} \nabla_{\mathbf{f}} \log L - K^{-1} \mathbf{f} \\ -n + \sum_{i=1}^n \frac{y_i - \mathbf{f}_i}{\sigma} - \frac{y_i - \mathbf{f}_i}{\sigma} \exp \left[-\frac{y_i - \mathbf{f}_i}{\sigma} \right] + \nabla_{\delta_1} \log \pi(\delta_1) \\ \frac{1}{2} \text{tr} \left[K^{-1} (\mathbf{f} \mathbf{f}' K^{-1} - I) \frac{\partial K}{\partial \delta_2} \right] + \nabla_{\delta_2} \log \pi(\delta_2) \\ \vdots \\ \frac{1}{2} \text{tr} \left[K^{-1} (\mathbf{f} \mathbf{f}' K^{-1} - I) \frac{\partial K}{\partial \delta_6} \right] + \nabla_{\delta_6} \log \pi(\delta_6) \end{bmatrix} \quad (4.40)$$

com,

$$\frac{\partial \log L}{\partial \mathbf{f}_i} = \frac{1}{\sigma} \left[1 - \exp \left(-\frac{y_i - \mathbf{f}_i}{\sigma} \right) \right] \quad (4.41)$$

para $i = 1, \dots, n$. As derivadas da função de covariância com respeito a cada hiperparâmetro são dadas pelas equações (4.25). A métrica G dada por (4.35) se resume para uma métrica que depende somente dos hiperparâmetros, sendo uma métrica 'flat'

para os valores do processo latente, i.e.,

$$\begin{aligned}
G(\mathbf{f}, \boldsymbol{\delta}) &= -E_{\mathbf{y}|\mathbf{f}, \boldsymbol{\delta}_1}(\nabla\nabla\ell) \\
&\quad - E_{\mathbf{f}|\boldsymbol{\delta}_{-1}}[\nabla\nabla \log N_n(\mathbf{0}, K)] \\
&\quad - \nabla\nabla \log \pi(\boldsymbol{\delta}) \\
&= \begin{bmatrix} G_1 & \mathbf{0} \\ \mathbf{0} & G_2 \end{bmatrix}
\end{aligned} \tag{4.42}$$

com

$$G_1 = \begin{bmatrix} \text{diag} \left[\frac{1}{\sigma^2} \right]_{T \times T} + K^{-1} & \frac{(\gamma - 1)}{\sigma} \mathbf{1}_{T \times 1} \\ \frac{(\gamma - 1)}{\sigma} \mathbf{1}_{1 \times T} & T \left[\frac{\pi^2}{6} + (1 - \gamma)^2 \right] - \nabla\nabla_{\boldsymbol{\delta}_1} \log \pi(\boldsymbol{\delta}_1) \end{bmatrix} \tag{4.43}$$

e

$$G_2 = \{G_2\}_{kl} = \frac{1}{2} \text{tr} \left[K^{-1} \frac{\partial K}{\partial \delta_k} K^{-1} \frac{\partial K}{\partial \delta_l} \right] - \nabla_{\delta_k} \nabla_{\delta_l} \log \pi(\boldsymbol{\delta}) \tag{4.44}$$

para $k, l = 2, \dots, 6$.

Neste caso, optei ainda em usar o algoritmo RMHMC com métrica fixa para gerar valores da posteriori de alta dimensão, pois a métrica depende somente dos hiperparâmetros. A desvantagem é que para determinar o valor MAP precisamos maximizar uma função de dimensão 106. No entanto como a métrica é um parâmetro livre no algoritmo RMHMC, realizei o seguinte procedimento. Assumi uma distribuição observacional normal para os dados e obtemos o MAP da função log-posteriori. Substitui estes valores em (4.42).

Ainda, notamos que a matriz de covariância K é extremamente mal-condicionada, sendo assim, adicionamos uma matriz diagonal de valores pequenos (0.0085) para melhorar o condicionamento da matriz de covariância. Segundos os autores Rasmussem & Williams (2006), este procedimento não traz problemas e melhora a estabilidade numérica na inversão computacional da matriz K .

Geramos uma amostra de 500000 vetores de dimensão 106 e tomamos vetores com defasagem de tamanho 20 (gerou-se no total 53 milhões de valores em aproximadamente 22 horas). Os resultados podem ser vistos na tabela 4.2. A figura 4.9 ilustra a dependência entre alguns valores da função desconhecida, $f(x_1)$, $f(x_2)$, $f(x_{99})$ e $f(x_{100})$. A figura 4.10 ilustra a estrutura de dependência entre os hiperparâmetros (na reta) a posteriori.

	σ	ϕ_1	ϕ_2	ϕ_3	ϕ_4	ϕ_5
$\widehat{E}[\cdot D]$	0.46	25.08	-1.59	270.09	0.78	117.78
$\widehat{sd}[\cdot D]$	0.04	122.28	5.13	871.26	0.01	365.60
$\widehat{\text{moda}}$	0.46	0.42	-1.43	1.74	0.78	0.83
$\widehat{\text{med}}$	0.46	1.87	-0.15	17.74	0.78	9.61
$IC\ 95\%$	[0.40, 0.54]	[0.27, 82.47]	[-2.51, 11.96]	[1.01, 1412.61]	[0.77, 0.79]	[0.46, 597.41]

Tabela 4.2: Resumos de interesse dos hiperparâmetros a posteriori (4.38)

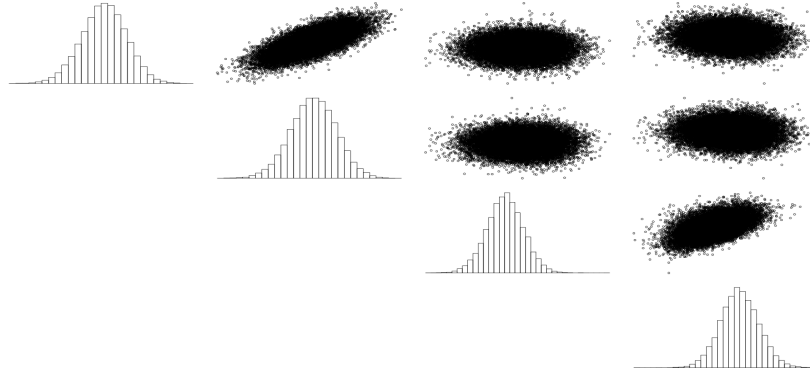


Figura 4.9: Histogramas (distribuições marginais) e gráficos de dispersão para os valores das função, $f(x_1)|D$, $f(x_2)|D$, $f(x_{99})|D$ e $f(x_{100})|D$.

Para determinar valores preditivos de $\mathbf{y}_*|D$, procede-se de modo similar a seção anterior, porém notando que,

$$\begin{aligned}
\pi(\mathbf{y}_*|D) &= \int \pi(\mathbf{y}_*, \mathbf{f}_*, \sigma|D) d\mathbf{f}_* d\sigma \\
&= \int \pi(\mathbf{y}_*|\mathbf{f}_*, \sigma) \pi(\mathbf{f}_*, \sigma|D) d\mathbf{f}_* d\sigma \\
&= E_{\mathbf{f}_*, \sigma|D}[\pi(\mathbf{y}_*|\mathbf{f}_*, \sigma)].
\end{aligned} \tag{4.45}$$

Varremos um intervalo $I_* = (-6.5, 23)$ com 300 pontos. Geramos para cada t no intervalo I_* , o valor preditivo do processo latente e seguidamente usamos (4.45). Uma vez que a distribuição valor extremo é assimétrica, tomamos tanto a média como a moda para cada t fixo no intervalo T_* . Os resultados são dados na figura 4.11 e na figura 4.12.

É fácil notar que a moda a posteriori da distribuição preditiva $\mathbf{y}_*|D$ fica mais próxima da função verdadeira em relação a esperança a posteriori ($E[\mathbf{y}_*|D]$). Fato que corrobora com o uso da distribuição GEV como distribuição observacional dos dados com característica de máximos (ou mínimos).

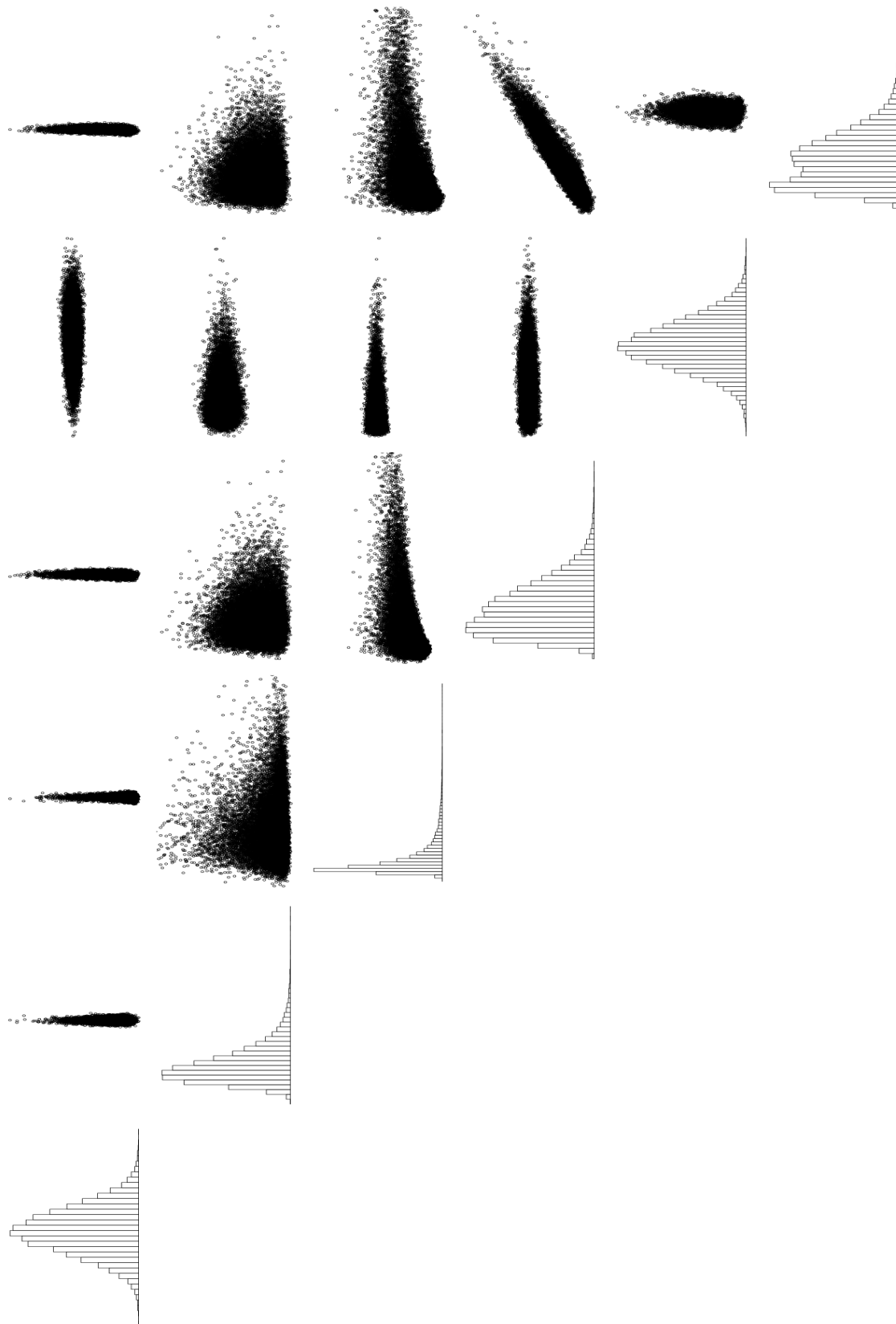


Figura 4.10: Histogramas (distribuições marginais) e gráficos de dispersão dos hiperparâmetros na reta

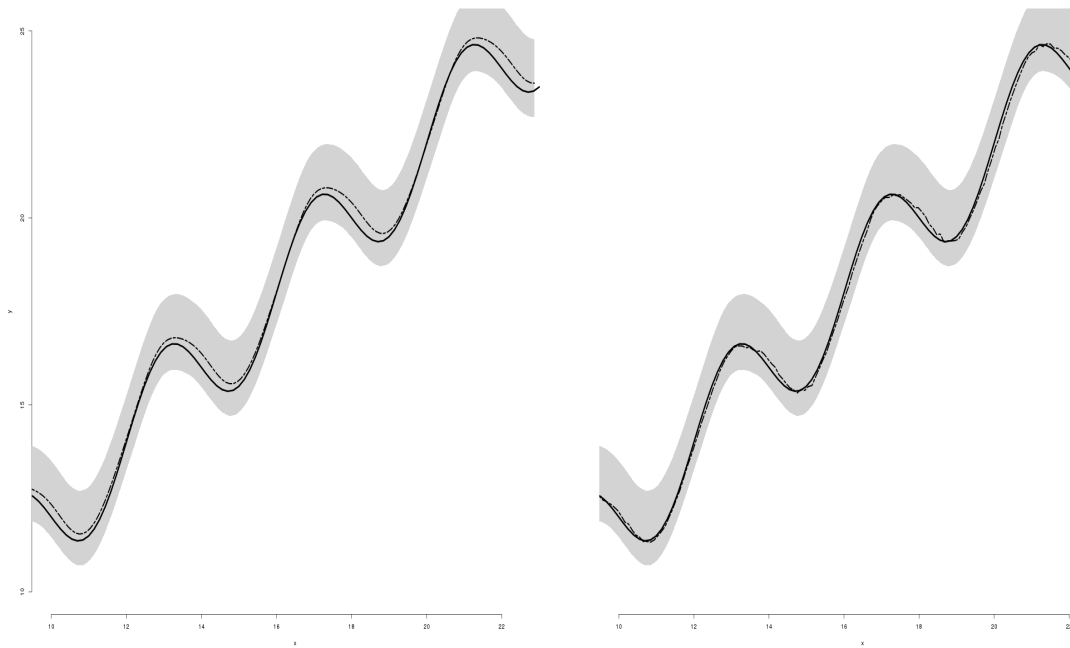


Figura 4.11: Intervalo de previsão $I_* = (10, 22)$. Média da distribuição preditiva $\mathbf{y}_*|D$ à esquerda com linha tracejada. Moda da distribuição preditiva $\mathbf{y}_*|D$ à direita com linha tracejada. Função verdadeira com linhas sólidas.

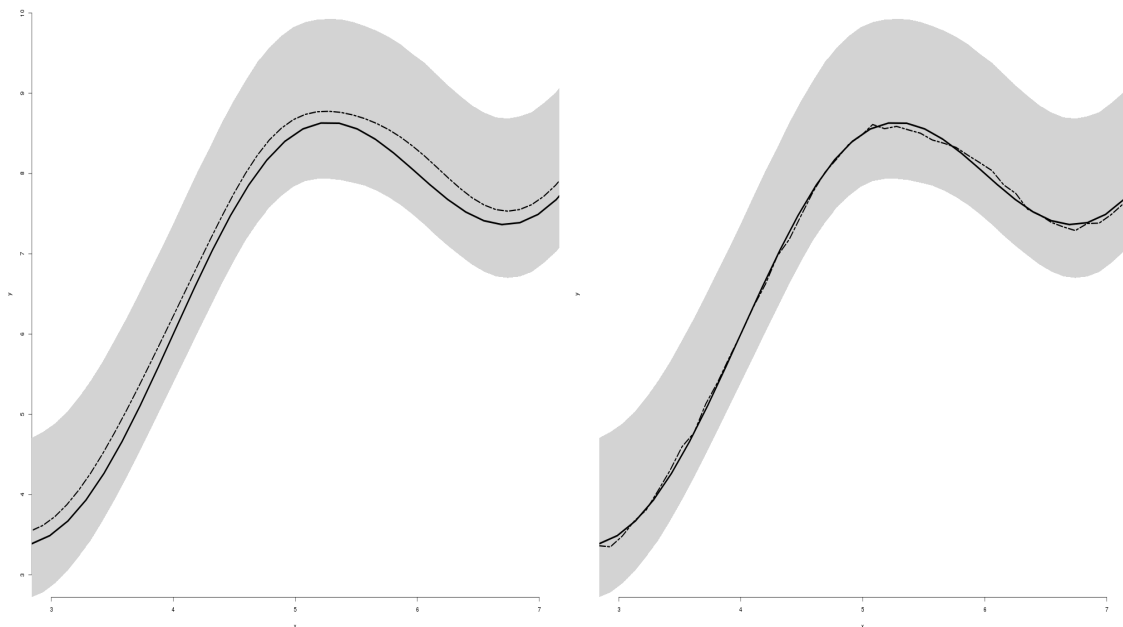


Figura 4.12: Intervalo previsão para os dados faltantes $I_* = (3, 7)$. Média da distribuição preditiva $\mathbf{y}_*|D$ à esquerda com linha tracejada. Moda da distribuição preditiva $\mathbf{y}_*|D$ à direita com linha tracejada. Função verdadeira com linhas sólidas.

4.6 Aplicação em dados reais

O conjunto de dados nesta aplicação representa a temperatura máxima mensal entre as épocas de 9/1993 e 8/1998, na estação Hothera, localizada na Antártica. Este conjunto de dados é encontrado através do site,

<http://www.antarctica.ac.uk/met/data.html>

O total de observações é de 60, porém deixamos as 10 últimas para realizar previsões. Assim $T = 50$. A série observada é vista no gráfico 4.13. Procedemos de modo similar a seção anterior, porém neste caso, consideramos a distribuição GEV e a mesma combinação linear de funções de covariância da seção anterior dada em (4.18), devido a forma periódica das observações. Deste modo segue que a posteriori de dimensão 57 é

$$\pi(\mathbf{f}, \sigma, \xi, \boldsymbol{\phi} | D) \propto L(\mathbf{f}, \sigma, \xi | \mathbf{y}) N_T(\mathbf{f} | \boldsymbol{\phi}) \pi(\sigma) \pi(\xi) \pi(\boldsymbol{\phi}), \quad (4.46)$$

assumimos parâmetros independentes a priori e distribuições vagas dadas por,

$$\begin{aligned} \sigma &\sim IG(0.1, 0.1) \\ \xi &\sim U(-0.5, 0.5) \\ \boldsymbol{\phi}_{-2} &\sim IG(0.1, 0.1) \\ \phi_2 &\sim N(0, 25). \end{aligned} \quad (4.47)$$

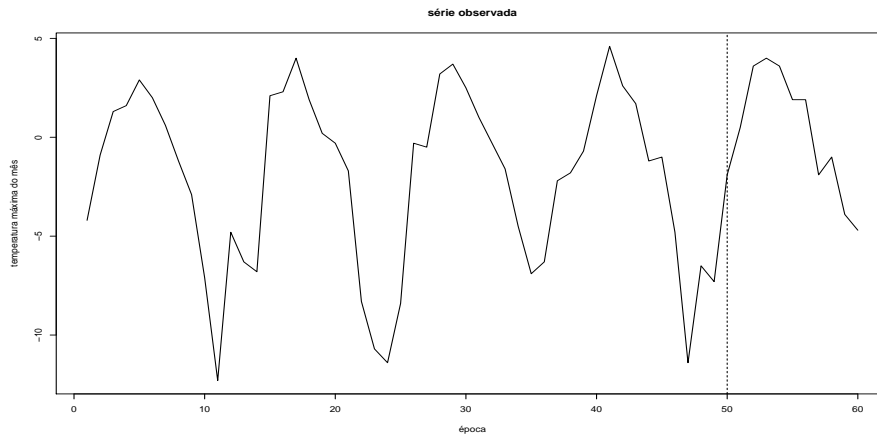


Figura 4.13: Temperatura máxima na estação Rothera, Antártica.

O vetor gradiente da função log-posteriori com parametrização na reta é dado por,

$$\nabla \ell = \begin{bmatrix} \nabla_{\mathbf{f}} \log L - K^{-1} \mathbf{f} \\ \sum_{i=1}^n (1 + \xi) \left(\frac{y_i - f_i}{\sigma} \right) z_i^{-1} - 1 - z_i^{-(1/\xi+1)} \left(\frac{y_i - f_i}{\sigma} \right) + \nabla_{\delta_1} \log \pi(\delta_1) \\ \sum_{i=1}^n \frac{\log z_i}{\xi^2} - \left(\frac{1}{\xi} + 1 \right) \left(\frac{y_i - f_i}{\sigma} \right) z_i^{-1} + \frac{1}{\xi} \left(\frac{y_i - f_i}{\sigma} \right) z_i^{-(1/\xi+1)} - \frac{\log z_i}{\xi^2} z_i^{-1/\xi} \\ \frac{1}{2} \text{tr} \left[K^{-1} (\mathbf{f} \mathbf{f}' K^{-1} - I) \frac{\partial K}{\partial \delta_2} \right] + \nabla_{\delta_2} \log \pi(\delta_2) \\ \vdots \\ \frac{1}{2} \text{tr} \left[K^{-1} (\mathbf{f} \mathbf{f}' K^{-1} - I) \frac{\partial K}{\partial \delta_6} \right] + \nabla_{\delta_6} \log \pi(\delta_6) \end{bmatrix} \quad (4.48)$$

com,

$$\frac{\partial \log L}{\partial f_i} = \frac{1}{\sigma} z_i^{-1} \left((1 + \xi) - z_i^{-1/\xi} \right) \quad (4.49)$$

em que $z_i = 1 + \xi(y_i - f_i)/\sigma$ para $i = 1, \dots, n$.

A métrica utilizada também com parametrização na reta é dada na fórmula abaixo. Notar que a métrica também é 'flat' para os valores desconhecidos da função.

$$\begin{aligned} G(\mathbf{f}, \boldsymbol{\delta}) &= -E_{\mathbf{y}|\mathbf{f}, \delta_1, \xi}(\nabla \nabla \ell) \\ &\quad - E_{\mathbf{f}|\delta_{-1}}[\nabla \nabla \log N_n(\mathbf{0}, K)] \\ &\quad - \nabla \nabla \log \pi(\boldsymbol{\delta}) \\ &= \begin{bmatrix} G_1 & \mathbf{0} \\ \mathbf{0} & G_2 \end{bmatrix} \end{aligned} \quad (4.50)$$

com

$$G_1 = \begin{bmatrix} \text{diag} \left[\frac{A}{\sigma^2} \right]_{T \times T} + K^{-1} & -\frac{1}{\sigma^2 \xi} [A - \Gamma(2 + \xi)] \mathbf{1}_{T \times 1} & -\frac{1}{\sigma \xi} \left(B - \frac{A}{\xi} \right) \mathbf{1}_{T \times 1} \\ \frac{T}{\xi^2} [1 - 2\Gamma(2 + \xi) + A] & -\frac{T}{\sigma \xi^2} \left[1 - \gamma + \frac{1 - \Gamma(2 + \xi)}{\xi} - B + \frac{A}{\xi} \right] & \\ & \frac{T}{\xi^2} \left[\frac{\pi^2}{6} + \left(1 - \gamma + \frac{1}{\xi} \right)^2 - \frac{2B}{\xi} + \frac{A}{\xi^2} \right] & \end{bmatrix} \quad (4.51)$$

e

$$G_2 = \{G_2\}_{kl} = \frac{1}{2} \text{tr} \left[K^{-1} \frac{\partial K}{\partial \delta_k} K^{-1} \frac{\partial K}{\partial \delta_l} \right] - \nabla_{\delta_k} \nabla_{\delta_l} \log \pi(\boldsymbol{\delta}) \quad (4.52)$$

para $k, l = 2, \dots, 6$.

Nesta caso, também obtemos por gerar valores da posteriori via RMHMC com métrica fixa. Realizamos o mesmo procedimento da seção anterior, porém não há informação alguma sobre o valor de ξ . Uma vez que a priori para ξ possui suporte em

um intervalo pequeno da reta, geramos algumas cadeias (para a posteriori) usando diferentes valores de ξ na métrica fixa até obter um valor que forneça uma convergência razoável da cadeia.

Usamos o algoritmo com stepsize 0.06 e repetimos a solução de störmer-verlet 12 vezes. No total, geramos aproximadamente 43 milhões de valores, com período de aquecimento de 57 mil e defasagem de tamanho 40 entre os vetores gerados. O tempo total foi de aproximadamente 6 horas. Ao final, obtemos 18000 valores para cada parâmetros da distribuição a posteriori.

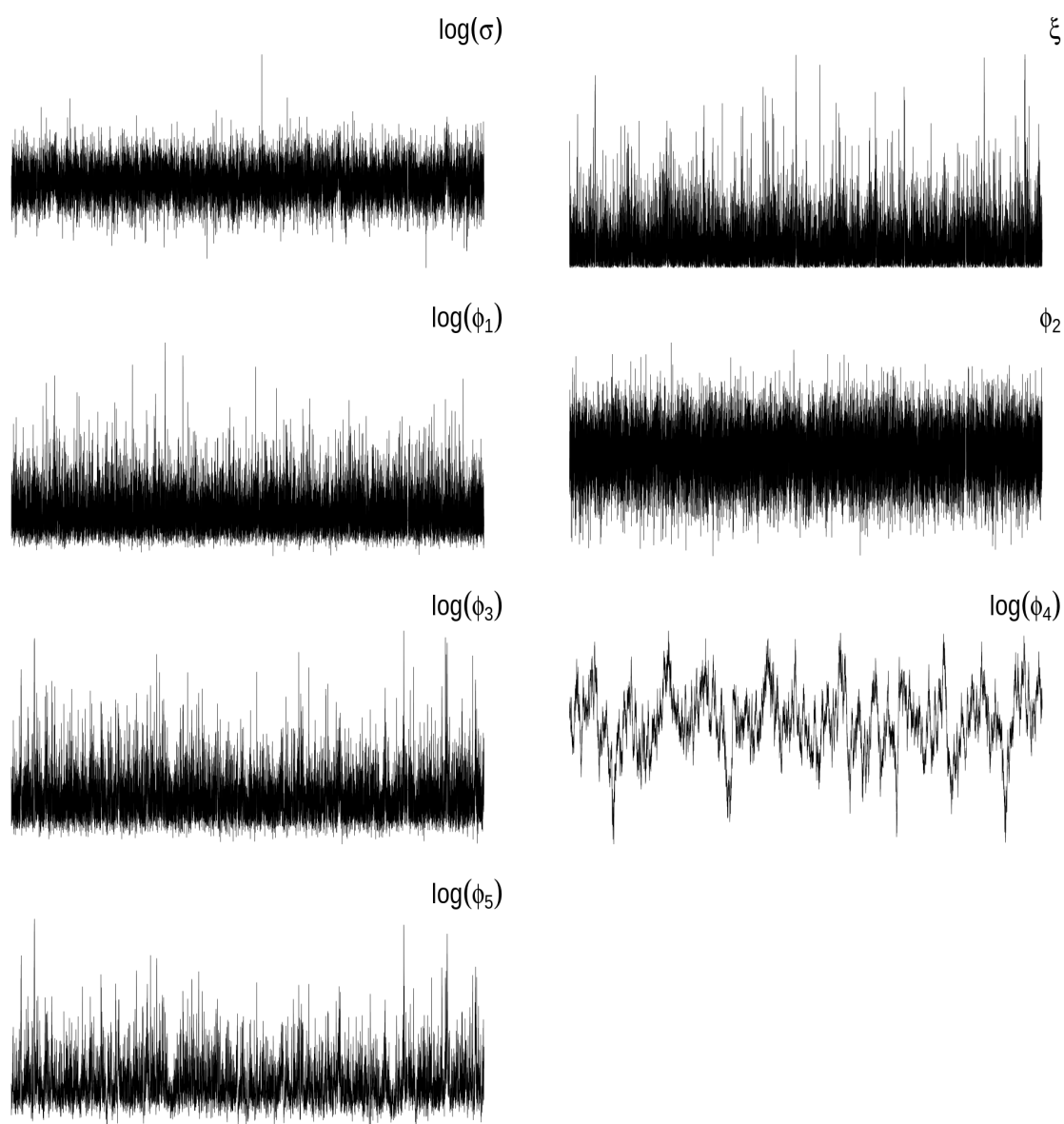


Figura 4.14: Valores gerados da posteriori para os hiperparâmetros na reta.

	σ	ξ	ϕ_1	ϕ_2	ϕ_3	ϕ_4	ϕ_5
$\widehat{E[. D]}$	1.84	-0.40	2.44	0.12	111.11	0.26	2.77
$\widehat{sd[. D]}$	0.21	0.09	8.34	4.91	233.12	0.002	5.29
$\widehat{\text{moda}}$	1.79	-0.48	0.07	0.09	22.79	0.25	0.57
$\widehat{\text{med}}$	1.83	-0.43	0.30	0.12	40.05	0.26	1.06
$IC\ 95\%$	[1.53, 2.22]	[-0.49, -0.21]	[0.04, 10.91]	[-8.17, 8.30]	[10.91, 451.18]	[0.25, 0.26]	[0.33, 11.43]

Tabela 4.3: Resumos de interesse dos hiperparâmetros a posteriori (4.46)

Note que as cadeias para os hiperparâmetros apresentam boa convergência, exceto para a cadeia de ϕ_4 , que apresenta alta correlação, mas que não foi rejeitada no teste de Geweke ao nível de 95%.

Os resultados na tabela 4.3, indicam periodicidade de aproximadamente 12 meses, i.e. $\hat{\omega} = \pi/\hat{\phi}_5 = 12.08$. A variabilidade das funções representada pela soma de $\hat{\phi}_1 + \hat{\phi}_3$ fica em torno de 23, um valor próximo à variabilidade total dos dados. A função tende a cortar o eixo das abcissas próximo da origem, $\hat{\phi}_2 = 0.09$.

As previsões futuras são realizadas de modo similar a seção anterior, notando que a distribuição agora é a GEV. O intervalo de previsões é $I_* = (1, 60)$ com 300 pontos. Os resultados das previsões são dados pela figura 4.15.

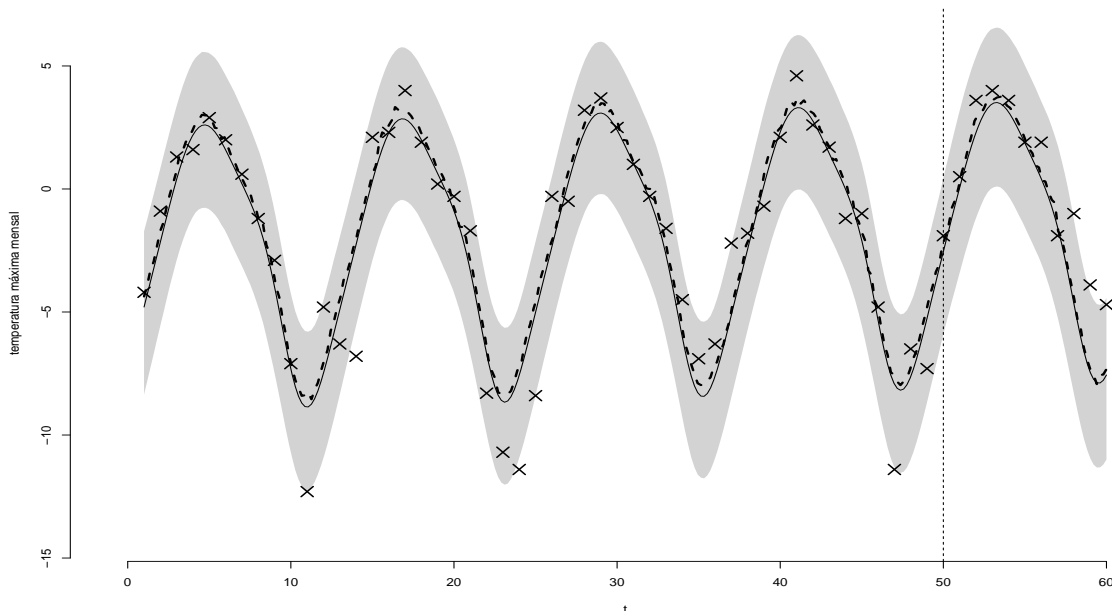


Figura 4.15: Previsões com média e moda a posteriori da distribuição de $\mathbf{y}_*|D$. A linha cheia representa a média a posteriori, linha tracejada representa a moda a posteriori. Os pontos observados são representados pelas cruces. O intervalo de credibilidade 95% é representado pela região cinza.

Conclusão

O desenvolvimento de métodos eficientes de simulação estocástica concomitante ao crescente poder computacional permite a inferência em modelos complexos de alta dimensão. O principal objetivo deste trabalho foi aplicar métodos HMC e suas variantes em modelos de valores extremos que requerem uma manipulação mais refinada da posteriori para obter boas inferências. Realizamos algumas aplicações desse método na área de valores extremos, tanto em modelos paramétricos como não-paramétricos sobre a abordagem Bayesiana.

No Capítulo 2, fizemos uma revisão dos métodos HMC com exemplos em valores extremos e mostramos sua performance superior em relação ao algoritmo de Metropolis-Hastings. Notamos que embora o método possua um esforço computacional maior a sua qualidade é melhor em relação ao algoritmo de Metropolis. Notar também que não levamos em consideração o tempo computacional envolvido nas simulações. Afirmamos isto no sentido de que não foi necessário uma configuração trabalhosa do algoritmo HMC em relação ao algoritmo MH.

No Capítulo 3, revimos a fundamentação teórica de valores extremos e propomos um modelo de série temporal autoregressivo com distribuição generalizada de valor extremo para o ruído. Além disso, determinamos a respectiva matriz de Informação de Fisher para o modelo. Resultado não conhecido na literatura. Conduzimos também um estudo de simulação para a estimação do parâmetros destes modelos, comparando quão rápido os algoritmos HMC e RMHMC tendem a alcançar distribuição estacionária, fornecendo assim melhores estimativas para os parâmetros.

No capítulo 4, introduzimos a modelagem Bayesiana não-paramétrica e vimos

como fazer inferência sobre uma função desconhecida via processo Gaussiano. Ressaltamos também que a inferência realizada sobre o modelo é completamente Bayesiana via RMHMC, pois é praxe fazer a inferência destes modelos sobre a estimativa MAP (máximo a posteriori, ver Rasmussem & Williams 2006) devido a dificuldade de implementação do algoritmo MH.

Seguindo nesta direção, introduzimos a modelo Bayesiano não-paramétrico em valores extremos e fizemos uso do algoritmo RMHMC. Aplicamos o modelo tanto em dados simulados como em dados reais. Embora o modelo Bayesiano não-paramétrico seja analiticamente complexo, notamos que a tratabilidade do parâmetros μ como um função aleatória é uma boa alternativa para o estudo de dados de valores extremos. É importante colocar que a moda, neste caso, fornece uma estimativa melhor que a média para a função desconhecida.

Através do site <http://olddunwich.wordpress.com> é possível acessar os principais códigos-R da dissertação para reproduzir resultados similares.

5.1 Perspectivas Futuras

O método de Monte Carlo Hamiltoniano com suas variantes abre um novo horizonte de aplicações e estudos teóricos.

A utilização de conceitos da geometria diferencial para o desenvolvimento de algoritmos de Monte Carlo fornecem uma nova fonte de pesquisas tanto na estatística aplicada como no estudo teórico de cadeias de Markov. Diferentes métricas podem ser aplicadas juntamente com outras funções de probabilidades multivariadas para a energia cinética do sistema, e outros métodos de solução numérica para EDO's são possíveis de estudos.

A aplicação prática nos algoritmos de salto reversíveis também é alvo de importante investigação. Modelos complexos que anteriormente eram de difícil manipulação podem ser colocados em prática via uma abordagem Bayesiana. Na abordagem clássica, o algoritmo também é fonte de estudo, uma vez que pode-se aplicar o método HMC concomitante a técnica simulated-annealing para obter valores modais de funções de verossimilhança complexas.

Neste sentido, a abordagem Bayesiana para a modelagem estatística é colocada em

outro patamar. Sua relação com conceitos de física, equações diferenciais, geometria diferencial e probabilidade passar ser íntima e fundamental para qualquer tipo de estudo teórico e aplicado.

Bibliografia

- Abrahamsen, P. (1977). A review of gaussian random fields and correlation functions. Norwegian Computer Center, Box 114 Blindern, N-0314 Oslo, Norway.
- Amari, S. I., & Nagaoka, H. (2000). *Methods of Information Geometry*. Oxford University Press.
- Balakrishnan, N., & Shiji, K. (2013). Extreme value autoregressive model and its applications. *Journal of Statistical Theory and Practice*.
- Barreto-Souza, W., & Vasconcellos, K. L. (2011). Bias and skewness in a general extreme-value regression model. *Computational Statistics and Data Analysis*.
- Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society*, (53), 370–418.
- Berger, J. O. (1980). *Statistical Decision Theory and Bayesian Analysis*. Springer Series in Statistics.
- Bernardo, J. M., & Smith, A. (1994). *Bayesian Theory*. John Wiley and Sons.
- Bonassi, F. V. (2009). *Permutabilidade de Quantidades Aleatórias Binárias e a Falácia do Apostador*. Master's thesis, IME-USP.
- Box, G., Jenkins, G. M., & Reinsel, G. (1994). *Time Series Analysis: Forecasting and Control*. Prentice-Hall International, 3 ed.
- Brahim-Belhouari, S., & Bermak, A. (2004). Gaussian process for nonstationary time series prediction. *Computational Statistics and Data Analysis*, (47), 705–712.
- Bronson, R. (1977). *Moderna introdução às equações diferenciais, Coleção Schaum*. McGraw-Hill.

- Burda, M., & Maheu, J. M. (2013). Bayesian adaptively updated hamiltonian monte carlo with an application to high-dimensional bekk garch models. *Studies in Non-linear Dynamics and Econometrics*, 17, 345–372.
- Calderhead, B. (2007). *A Study of Population MCMC for estimating Bayes Factors over Nonlinear ODE Models*. Master's thesis, University of Glasgow.
- Calderhead, B. (2012). *Differential geometric MCMC methods and applications*. Ph.D. thesis, University of Glasgow.
- Calderhead, B., & Girolami, M. (2011). Riemann manifold langevin and hamiltonian monte carlo methods. *Journal of the Statistical Royal Society B*, 73(2), 123–214.
- Carmo, M. P. D. (1976). *Differential Geometry of Curves and Surfaces*. Prentice-Hall Inc, 1 ed.
- Castillo, E., Hadi, A. S., & Balakrishnan, N. (2004). *Extreme Values and Related Models with Applications in Engineering and Science*. A John Wiley and Sons Publication.
- Chib, S., & Greenberg, E. (1995). Understanding the metropolis-hastings algorithm. *The American Statistician*, 49(4), 327–335.
- Chkrebtii, O. A. (2013). *Probabilistic Solution of Differential Equations for Bayesian Uncertainty Quantification and Inference*. Ph.D. thesis, Simon Fraser University.
- Choo, K. (2000). *Learning Hyperparameters for Neural Networks Models using Hamiltonian Dynamics*. Master's thesis, University of Toronto.
- Clur, J.-C. (2010). *Nonparametric Smoothing in Extreme Value Theory*. Master's thesis, University of Cape Town.
- Coles, S. (2004). *An Introduction to Statistical Modelling of Extreme Values*. Springer Series in Statistics.
- Coles, S., Pericchi, L. R., & Sisson, S. (2003). A fully probabilistic approach to extreme rainfall modeling. *Journal of Hidrology*, 273, 35–50.

- Duane, S., Kennedy, A. D., Pendleton, B. J., & Roweth, D. (1987). Hybrid monte carlo. *Physics Letter B*, 195(2), 216–222.
- Duvenaud, D. K. (2014). *Automatic Model Construction with Gaussian Processes*. Ph.D. thesis, University of Cambridge.
- Fergusson, T. S. (1973). A bayesian analysis of some nonparametric problems. *The Annals of Statistic*, 1(2), 209–230.
- Filippone, M., Zhong, M., & Girolami, M. (2013). A comparative evaluation of stochastic-based inference methods for gaussian process models. *Journal of Machine Learning*, 93, 93–114.
- Fox, E. B. (2009). *Bayesian Nonparametric Learning of Complex Dynamical Phenomena*. Ph.D. thesis, Massachusetts Institute of Technology.
- Fuest, M. (2009). *Modelling Temporal Dependencies of Extreme Events via Point Processes*. Master’s thesis, Institut für Statistik Ludwig–Maximilians Universität München.
- Garthwaith, P. H., Kadane, J. B., & O’Hagan, A. (2005). Statistical methods for eliciting probability distributions. *Journal of the American Statistical Association*, 100(470), 680–700.
- Gelfand, A. E., & Kottas, A. (2002). A computational approach for full nonparametric bayesian inference under dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 11(2), 289–305.
- Geman, S., & Geman, D. (1984). Stochastic relaxation, gibbs distribution and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Learning*, 6(6), 721–741.
- Ghil, M., Yiou, P., Hallegatte, S., Malamud, B. D., Naveau, P., Soloviev, A., Friedrichs, P., Keilis-Borok, V., Kondrashov, D., Kossobokov, V., Mestre, O., Nicolis, C., Rust, H. W., Shebalin, P., Vrac, M., Witt, A., & Zaliapin, I. (2011). Extreme events: dynamics, statistics and prediction. *Nonlinear Processes in Geophysics*, (18), 295–350.

- Ghosh, J. K., & Ramamoorthi, R. V. (2003). *Bayesian Nonparametrics*. Springer Series in Statistics.
- Girard, A., Candela, J. Q., Murray-smith, R., & Rasmussen, C. E. (2003). Gaussian process priors with uncertain inputs - application to multiple-step ahead time series forecasting.
- Gregorcic, G., & Lightbody, G. (2002). Gaussian processes for modelling of dynamic non-linear systems. In *Irish Signals and Systems Conference*, (pp. 141–147).
- Grimmett, G., & Stirzaker, D. (2001). *Probability and Random Processes*. Oxford University Press.
- Hastings, W. K. (1970). Monte carlo sampling methods using markov chains and their applications. *Biometrika*, *57*(1), 97–109.
- Jylanki, P., Vanhatalo, J., & Vehtari, A. (2011). Gaussian process regression with a student-t likelihood. *Journal of Machine Learning Research*, (12), 3227–3257.
- Kass, R. E., & Wasserman, L. (1996). The selection of prior distributions by formal rules. *Journal of the American Statistical Association*, *91*(435), 1343–1370.
- Kotz, S., & Nadarajah, S. (2000). *Extreme Value Distributions. Theory and Applications*. Imperial College Press.
- Körding, K. P., & Wolpert, D. M. (2004). Bayesian integration in sensorimotor learning. *Nature*, *427*, 244–247.
- Lan, S. (2013). *Advanced Bayesian Computational Methods through Geometric Techniques*. Ph.D. thesis, University of California - Irvine.
- Leimkuhler, B., & Reich, S. (2005). *Simulating Hamiltonian Dynamics*. Cambridge University Press.
- Mackay, D. J. (1992). *Bayesian Methods for Adaptive Models*. Ph.D. thesis, California Institute of Technology - Pasadena.
- Mackay, D. J. (2003). *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press.

- Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. A., & Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, *21*(6), 1087–1092.
- Metropolis, N., & Ulam, S. (1949). Monte carlo method. *Journal of the American Statistical Association*, *44*(247), 335–341.
- Mikosch, T. (1998). *Elementary Stochastic Calculus with Finance View*. World Scientific.
- Minka, T. (2000). Old and new matrix useful for statistics.
- Minka, T. (2001). *A family of Algorithms for Approximate Bayesian Inference*. Ph.D. thesis, Massachusetts Institute of Technology.
- Müller, P., & Quintana, F. A. (2004). Nonparametric bayesian data analysis. *Statistical Science*, *19*(1), 95–110.
- Neal, R. M. (1995). *Bayesian Learning of Neural Networks*. Ph.D. thesis, University of Toronto.
- Neal, R. M. (1998). Regression and classification using gaussian process priors. *Bayesian Statistics*, *6*.
- Neal, R. M. (2011). *Handbook of Markov Chain Monte Carlo*, chap. 5. Chapman and Hall CRC Press.
- O’Hagan, A. (1978). Curve fitting and optimal design for prediction. *Journal of Royal Statistical Society B*, *40*(1), 1–42.
- O’Hagan, A. (1994). *Kendall’s Advanced Theory of Statistics: Bayesian Inference*. Oxford University Press.
- Osborne, M. (2010). *Bayesian Gaussian Processes for Sequential Prediction, Optimisation and Quadrature*. Ph.D. thesis, University of Oxford.
- Prescott, P., & Walden, A. T. (1980). Maximum likelihood estimation of the parameters of the generalized extreme-value distribution. *Biometrika*, *67*, 723–724.

- Rasmussen, C. E., & Williams, C. K. (2006). *Gaussian Process and Machine Learning*. Massachusetts Institute of Technology Press.
- Rasmussen, C. E. (1996). *Evaluation of Gaussian Processes and other Methods of Non-linear Regression*. Ph.D. thesis, University of Toronto.
- Robert, C. P., & Casella, G. (2004). *Monte Carlo Statistical Methods*. Springer Text in Statistics.
- Rue, H., & Held, L. (2005). *Gaussian Markov Random Fields Theory and Applications*. Chapman and Hall CRC.
- Rue, H., & Martino, S. (2009). Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations. *Journal of the Royal Statistical Society B*, 72(2), 319–392.
- Schervish, M. J. (2011). *Theory of Statistics*. Springer Series in Statistics.
- Schinazi, R. B. (1999). *Classical and Spatial Stochastic Processes*. Birkhäuser Boston.
- Skilling, J. (1992). Bayesian solution of ordinary differential equations. *Maximum Entropy and Bayesian Methods - Fundamental Theories of Physics*, 50, 23–37.
- Smith, E. (2005). *Bayesian Modelling of Extreme Rainfall Data*. Ph.D. thesis, University of Newcastle - Tyne.
- Smith, R. L. (1985). Maximum likelihood estimation in a class of nonregular cases. *Biometrika*, 72(1), 67–90.
- Särkkä, S. (2013). *Bayesian Filtering and Smoothing*. Cambridge University Press.
- Tanner, M. A. (1996). *Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions*. Springer.
- Teh, Y. W., & Orbanz, P. (2010). *Encyclopedia of Machine Learning*, chap. Bayesian Nonparametric Models. Springer-Verlag.
- Tenenblat, K. (2008). *Introdução à Geometria Diferencial*. Blucher, 2 ed.

- Toulemonde, G., Guillaou, A., Naveau, P., Vrac, M., & Chevallier, F. (2010). Autoregressive models for maxima and their applications to methane and nitrous oxide. *Environmetrics*, *21*, 189–2007.
- Wang, Z. (2012). *A Bayesian Nonparametric Modelling Framework for Extreme Value Analysis*. Ph.D. thesis, University of California - Santa Cruz.
- West, M., & Harrison, J. (1997). *Bayesian Forecasting and Dynamic Models*. Springer Series in Statistics.
- Zhang, H. (2004). Inconsistent estimation and asymptotically equal interpolations in model-based geostatistics. *Journal of the American Statistical Association*, *99*(465), 250–261.
- Zhao, X., Scarrott, C. J., Oxley, L., & Reale, M. (2011). Garch dependence in extreme value models with bayesian inference. *Mathematics and Computers in Simulation*, *81*, 1430–1440.
- Zin, W. Z. W., Jemain, A. A., & Ibrahim, K. (2012). Bayesian changepoint analysis of the extreme rainfall events. *Journal of Mathematics and Statistics*, *8*(1), 85–91.

Apêndice A

Fórmulas Matriciais

A.1 Derivadas matriciais

$$\begin{aligned}\frac{\partial}{\partial\phi}\text{tr}(K) &= \text{tr}\left(\frac{\partial}{\partial\phi}K\right) \\ \frac{\partial}{\partial\phi}K^{-1} &= -K^{-1}\frac{\partial K}{\partial\phi}K^{-1} \\ \frac{\partial}{\partial\phi}\log|K| &= \text{tr}\left[K^{-1}\frac{\partial K}{\partial\phi}\right]\end{aligned}\tag{A.1}$$

Toma-se a derivada usual em relação a cada elemento da matriz quadrada K .

A.2 Fórmula de Sherman-Morrison-Woodbury-Schur

$$\begin{aligned}(A + UBV^t)^{-1} &= A^{-1} - A^{-1}U(B^{-1} + V^tA^{-1}U)^{-1}V^tA^{-1} \\ A(A + B)^{-1} &= AB^{-1} - A(A + B)AB^{-1}\end{aligned}\tag{A.2}$$

A e B são matrizes inversíveis. U e V são matrizes retangulares de dimensão compatível.

A.3 Matriz informação de Fisher para o modelo autoregressivo-GEV

Primeiramente, é necessário encontrar as derivadas de segunda ordem da função log-verossimilhança com respeito à todos os parâmetros, em seguida, aplicar as propriedades de esperança condicional sobre todas as expressões com o sinal negativo. Note

que não precisamos reencontrar as derivadas da função de log-verossimilhança, usamos os resultados já conhecidos para o caso de uma amostra aleatória. As propriedades usadas são: regra da cadeia para derivadas, a derivada da soma é a soma das derivadas, a esperança da soma é a soma das esperanças e a esperança condicional, e.g., $E[g(Y_t)] = E[E[g(Y_t)|D_{t-1}]]$, $\forall t$.

$$\begin{aligned}
\frac{\partial^2 \ell}{\partial \mu^2} &= \frac{\partial}{\partial \mu} \left(\frac{\partial \sum_{t=p+1}^T \ell_t}{\partial \mu} \right) = \sum_{t=p+1}^T \frac{\partial}{\partial \mu} \left(\frac{\partial \ell_t}{\partial \mu_t} \frac{\partial \mu_t}{\partial \mu} \right) \\
&= \sum_{t=p+1}^T \frac{\partial}{\partial \mu} \frac{\partial \ell_t}{\partial \mu_t} \\
&= \sum_{t=p+1}^T \frac{\partial^2 \ell_t}{\partial \mu_t^2} \\
-E \left(\frac{\partial^2 \ell_t}{\partial \mu^2} \right) &= - \sum_{t=p+1}^T E \left[\frac{\partial^2 \ell_t}{\partial \mu_t^2} \right] \\
&= - \sum_{t=p+1}^T E \left[E \left(\frac{\partial^2 \ell_t}{\partial \mu_t^2} \middle| D_{t-1} \right) \right] \\
&= (T-p) \frac{A}{\sigma^2} \tag{A.3}
\end{aligned}$$

$$\begin{aligned}
\frac{\partial^2 \ell}{\partial \mu \partial \theta_j} &= \frac{\partial}{\partial \mu} \left(\frac{\partial \sum_{t=p+1}^T \ell_t}{\partial \theta_j} \right) = \sum_{t=p+1}^T \frac{\partial}{\partial \mu} \left(\frac{\partial \ell_t}{\partial \mu_t} \frac{\partial \mu_t}{\partial \theta_j} \right) \\
&= \sum_{t=p+1}^T \frac{\partial}{\partial \mu} \left(\frac{\partial \ell_t}{\partial \mu_t} Y_{t-j} \right) \\
&= \sum_{t=p+1}^T \frac{\partial^2 \ell_t}{\partial \mu_t^2} Y_{t-j} \\
-E \left(\frac{\partial^2 \ell}{\partial \mu \partial \theta_j} \right) &= - \sum_{t=p+1}^T E \left[\frac{\partial^2 \ell_t}{\partial \mu_t^2} Y_{t-j} \right] \\
&= - \sum_{t=p+1}^T E \left[E \left(\frac{\partial^2 \ell_t}{\partial \mu_t^2} Y_{t-j} \middle| D_{t-1} \right) \right] \\
&= - \sum_{t=p+1}^T E \left[Y_{t-j} E \left(\frac{\partial^2 \ell_t}{\partial \mu_t^2} \middle| D_{t-1} \right) \right] \\
&= (T-p) \frac{A}{\sigma^2} E[Y_{t-j}] = \mu_{Y_t} (T-p) \frac{A}{\sigma^2} \tag{A.4}
\end{aligned}$$

$$\begin{aligned}
\frac{\partial^2 \ell}{\partial \mu \partial \sigma} &= \frac{\partial}{\partial \sigma} \left(\frac{\partial \sum_{t=p+1}^T \ell_t}{\partial \mu} \right) \\
&= \sum_{t=p+1}^T \frac{\partial}{\partial \sigma} \left(\frac{\partial \ell_t}{\partial \mu_t} \frac{\partial \mu_t}{\partial \mu} \right) \\
&= \sum_{t=p+1}^T \frac{\partial^2 \ell_t}{\partial \sigma \partial \mu_t} \\
-E \left(\frac{\partial^2 \ell}{\partial \mu \partial \sigma} \right) &= - \sum_{t=p+1}^T E \left[\frac{\partial^2 \ell_t}{\partial \sigma \partial \mu_t} \right] \\
&= - \sum_{t=p+1}^T E \left[E \left(\frac{\partial^2 \ell}{\partial \sigma \partial \mu_t} \middle| D_{t-1} \right) \right] \\
&= -(T-p) \frac{1}{\sigma^2 \xi} [A - \Gamma(2 + \xi)] \tag{A.5}
\end{aligned}$$

$$\begin{aligned}
\frac{\partial^2 \ell}{\partial \mu \partial \xi} &= \frac{\partial}{\partial \xi} \left(\frac{\partial \sum_{t=p+1}^T \ell_t}{\partial \mu} \right) \\
&= \sum_{t=p+1}^T \frac{\partial}{\partial \xi} \left(\frac{\partial \ell_t}{\partial \mu_t} \frac{\partial \mu_t}{\partial \mu} \right) \\
&= \sum_{t=p+1}^T \frac{\partial^2 \ell_t}{\partial \xi \partial \mu_t} \\
-E \left(\frac{\partial^2 \ell}{\partial \mu \partial \xi} \right) &= - \sum_{t=p+1}^T E \left[\frac{\partial^2 \ell_t}{\partial \xi \partial \mu_t} \right] \\
&= - \sum_{t=p+1}^T E \left[E \left(\frac{\partial^2 \ell_t}{\partial \xi \partial \mu_t} \middle| D_{t-1} \right) \right] \\
&= -(T-p) \frac{1}{\sigma \xi} \left(B - \frac{A}{\xi} \right) \tag{A.6}
\end{aligned}$$

$$\begin{aligned}
\frac{\partial^2 \ell}{\partial \theta_i \partial \theta_j} &= \frac{\partial}{\partial \theta_i} \left(\frac{\partial \sum_{t=p+1}^T \ell_t}{\partial \theta_j} \right) \\
&= \sum_{t=p+1}^T \frac{\partial}{\partial \theta_i} \left(\frac{\partial \ell_t}{\partial \mu_t} \frac{\partial \mu_t}{\partial \theta_j} \right) \\
&= \sum_{t=p+1}^T \frac{\partial}{\partial \theta_i} \left(\frac{\partial \ell_t}{\partial \mu_t} Y_{t-j} \right) \\
&= \sum_{t=p+1}^T \frac{\partial^2 \ell_t}{\partial \mu_t^2} \frac{\partial \mu_t}{\partial \theta_i} Y_{t-j} \\
&= \sum_{t=p+1}^T \frac{\partial^2 \ell_t}{\partial \mu_t^2} Y_{t-i} Y_{t-j} \\
-E \left(\frac{\partial^2 \ell}{\partial \theta_i \partial \theta_j} \right) &= - \sum_{t=p+1}^T E \left[\frac{\partial^2 \ell}{\partial \mu_t^2} Y_{t-i} Y_{t-j} \right] \\
&= - \sum_{t=p+1}^T E \left[E \left(\frac{\partial^2 \ell}{\partial \mu_t^2} Y_{t-i} Y_{t-j} \middle| D_{t-1} \right) \right] \\
&= - \sum_{t=p+1}^T E \left[Y_{t-i} Y_{t-j} E \left(\frac{\partial^2 \ell_t}{\partial \mu_t^2} \middle| D_{t-1} \right) \right] \\
&= (T-p) \frac{A}{\sigma^2} E[Y_{t-i} Y_{t-j}] \quad \forall i, j
\end{aligned} \tag{A.7}$$

$$\begin{aligned}
\frac{\partial^2 \ell}{\partial \sigma \partial \theta_j} &= \frac{\partial}{\partial \sigma} \left(\frac{\partial \sum_{t=p+1}^T \ell_t}{\partial \theta_j} \right) \\
&= \sum_{t=p+1}^T \frac{\partial}{\partial \sigma} \left(\frac{\partial \ell_t}{\partial \mu_t} \frac{\partial \mu_t}{\partial \theta_j} \right) \\
&= \sum_{t=p+1}^T \frac{\partial}{\partial \sigma} \left(\frac{\partial \ell_t}{\partial \mu_t} Y_{t-j} \right) \\
&= \sum_{t=p+1}^T \frac{\partial^2 \ell_t}{\partial \sigma \partial \mu_t} Y_{t-j} \\
-E \left(\frac{\partial^2 \ell}{\partial \sigma \partial \theta_j} \right) &= - \sum_{t=p+1}^T E \left[\frac{\partial^2 \ell_t}{\partial \sigma \partial \mu_t} Y_{t-j} \right] \\
&= - \sum_{t=p+1}^T E \left[E \left(\frac{\partial^2 \ell_t}{\partial \sigma \partial \mu_t} Y_{t-j} \middle| D_{t-1} \right) \right] \\
&= - \sum_{t=p+1}^T E \left[Y_{t-j} E \left(\frac{\partial^2 \ell}{\partial \sigma \partial \mu_t} \middle| D_{t-1} \right) \right] \\
&= -(T-p) \frac{1}{\sigma^2 \xi} [A - \Gamma(2 + \xi)] \mu_{Y_t}
\end{aligned} \tag{A.8}$$

$$\begin{aligned}
\frac{\partial^2 \ell}{\partial \xi \partial \theta_j} &= \frac{\partial}{\partial \xi} \left(\frac{\partial \sum_{t=p+1}^T \ell_t}{\partial \theta_j} \right) \\
&= \sum_{t=p+1}^T \frac{\partial}{\partial \xi} \left(\frac{\partial \ell_t}{\partial \mu_t} \frac{\partial \mu_t}{\partial \theta_j} \right) \\
&= \sum_{t=p+1}^T \frac{\partial}{\partial \xi} \left(\frac{\partial \ell_t}{\partial \mu_t} Y_{t-j} \right) \\
&= \sum_{t=p+1}^T \frac{\partial^2 \ell_t}{\partial \xi \partial \mu_t} Y_{t-j} \\
-E \left(\frac{\partial^2 \ell}{\partial \xi \partial \theta_j} \right) &= - \sum_{t=p+1}^T E \left[\frac{\partial^2 \ell_t}{\partial \xi \partial \mu_t} Y_{t-j} \right] \\
&= - \sum_{t=p+1}^T E \left[E \left(\frac{\partial^2 \ell_t}{\partial \xi \partial \mu_t} Y_{t-j} \middle| D_{t-1} \right) \right] \\
&= - \sum_{t=p+1}^T E \left[Y_{t-j} E \left(\frac{\partial^2 \ell_t}{\partial \xi \partial \mu_t} \middle| D_{t-1} \right) \right] \\
&= -(T-p) \frac{1}{\sigma \xi} \left(B - \frac{A}{\xi} \right) \mu_{Y_t} \tag{A.9}
\end{aligned}$$

e de modo análogo,

$$-E \left(\frac{\partial^2 \ell}{\partial \xi \partial \sigma} \right) = -(T-p) \frac{1}{\sigma \xi^2} \left[1 - \gamma + \frac{1 - \Gamma(2 + \xi)}{\xi} - B + \frac{A}{\xi} \right]$$

A.4 Matriz de Informação de Fisher para o processo Gaussiano latente

Seja $\mathbf{f} \sim N_n(\mathbf{m}, K)$ com K dependendo de um vetor de hiperparâmetros ϕ . Considere $a = \mathbf{f} - \mathbf{m}$.

$$\log N_n(\mathbf{f}|\mathbf{m}, K) = -\frac{n}{2} \log(2\pi) - \frac{1}{2} \log |K| - \frac{1}{2} a^t K^{-1} a \quad (\text{A.10})$$

$$\begin{aligned} \frac{\partial}{\partial \phi_j} \log N(\mathbf{f}|\mathbf{m}, K) &= -\frac{1}{2} \text{tr} \left[K^{-1} \frac{\partial K}{\partial \phi_j} \right] - \frac{1}{2} \frac{\partial}{\partial \phi_j} \text{tr}(a^t K^{-1} a) \\ &= -\frac{1}{2} \text{tr} \left[K^{-1} \frac{\partial K}{\partial \phi_j} + \frac{\partial a a^t K^{-1}}{\partial \phi_j} \right] \\ &= -\frac{1}{2} \text{tr} \left[K^{-1} \frac{\partial K}{\partial \phi_j} - a a^t K^{-1} \frac{\partial K}{\partial \phi_j} K^{-1} \right] \\ &= -\frac{1}{2} \text{tr} \left[K^{-1} \frac{\partial K}{\partial \phi_j} - K^{-1} a a^t K^{-1} \frac{\partial K}{\partial \phi_j} \right] \\ &= \frac{1}{2} \text{tr} \left[K^{-1} (a a^t K^{-1} - I) \frac{\partial K}{\partial \phi_j} \right] \end{aligned} \quad (\text{A.11})$$

$$\begin{aligned} \frac{\partial^2}{\partial \phi_i \partial \phi_j} \log N(\mathbf{f}|\mathbf{m}, K) &= \frac{1}{2} \text{tr} \left[\frac{\partial}{\partial \phi_i} \left\{ K^{-1} (a a^t K^{-1} - I) \frac{\partial K}{\partial \phi_j} \right\} \right] \\ &= \frac{1}{2} \text{tr} \left[\left(-K^{-1} \frac{\partial K}{\partial \phi_i} K^{-1} a a^t K^{-1} - K^{-1} a a^t K^{-1} \frac{\partial K}{\partial \phi_i} K^{-1} \right. \right. \\ &\quad \left. \left. + K^{-1} \frac{\partial K}{\partial \phi_i} K^{-1} \right) \frac{\partial K}{\partial \phi_j} + K^{-1} a a^t K^{-1} \frac{\partial^2 K}{\partial \phi_i \partial \phi_j} - K^{-1} \frac{\partial^2 K}{\partial \phi_i \partial \phi_j} \right] \\ &= \frac{1}{2} \text{tr} \left[(I - 2K^{-1} a a^t) K^{-1} \frac{\partial K}{\partial \phi_i} K^{-1} \frac{\partial K}{\partial \phi_j} \right. \\ &\quad \left. + (K^{-1} a a^t K^{-1} - K^{-1}) \frac{\partial^2 K}{\partial \phi_i \partial \phi_j} \right]. \end{aligned} \quad (\text{A.12})$$

Tomando a esperança em relação ao último conjunto de equações e notando que $E(a a^t) = K$, os elementos da matriz informação de Fisher são dados por,

$$\begin{aligned} G_{ij} &= -E \left(\frac{\partial^2}{\partial \phi_i \partial \phi_j} \log N(\mathbf{f}|\mathbf{m}, K) \right) \\ &= \frac{1}{2} \text{tr} \left[K^{-1} \frac{\partial K}{\partial \phi_i} K^{-1} \frac{\partial K}{\partial \phi_j} \right] \end{aligned} \quad (\text{A.13})$$

A derivada de G_{ij} em relação ϕ_k é dada por,

$$\begin{aligned}
\frac{\partial}{\partial \phi_k} G_{ij} = & -\frac{1}{2} \text{tr} \left[K^{-1} \frac{\partial K}{\partial \phi_k} K^{-1} \frac{\partial K}{\partial \phi_i} K^{-1} \frac{\partial K}{\partial \phi_j} \right] \\
& - \frac{1}{2} \text{tr} \left[K^{-1} \frac{\partial K}{\partial \phi_i} K^{-1} \frac{\partial K}{\partial \phi_k} K^{-1} \frac{\partial K}{\partial \phi_j} \right] \\
& + \frac{1}{2} \text{tr} \left[K^{-1} \frac{\partial^2 K}{\partial \phi_i \partial \phi_k} K^{-1} \frac{\partial K}{\partial \phi_j} \right] \\
& + \frac{1}{2} \text{tr} \left[K^{-1} \frac{\partial K}{\partial \phi_i} K^{-1} \frac{\partial^2 K}{\partial \phi_j \partial \phi_k} \right]
\end{aligned} \tag{A.14}$$