

Universidade Federal do Rio de Janeiro

INFERÊNCIA BAYESIANA APROXIMADA EM
MODELOS DE ESPAÇO DE ESTADOS

Camila Maria Casquilho Resende



2011

INFERÊNCIA BAYESIANA APROXIMADA EM MODELOS DE ESPAÇO DE ESTADOS

Camila Maria Casquilho Resende

Dissertação de Mestrado submetida ao Programa de Pós-Graduação em Estatística do Instituto de Matemática da Universidade Federal do Rio de Janeiro - UFRJ, como parte dos requisitos necessários à obtenção do título de Mestre em Estatística.

Orientador: Dani Gamerman

Rio de Janeiro

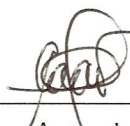
Junho 2011

INFERÊNCIA BAYESIANA APROXIMADA EM MODELOS DE ESPAÇO DE ESTADOS

Camila Maria Casquilho Resende

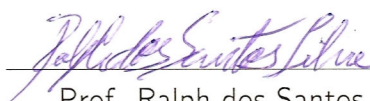
Orientador: Dani Gamerman

Dissertação de Mestrado submetida ao Programa de Pós-Graduação em Estatística do



Prof. Carlos Antonio Abanto Valle, D.Sc.

Instituto de Matemática - UFRJ



Prof. Ralph dos Santos Silva, D.Sc.

Instituto de Ciências Exatas - UFMG

Rio de Janeiro

Junho 2011

R433i Resende, Camila Maria Casquilho

Inferência bayesiana aproximada em modelos de espaço de estados / Camila Maria Casquilho Resende – Rio de Janeiro: UFRJ/IM, 2011.

xiii, 74 f.: il.; 31cm.

Orientador: Dani Gamerman.

Dissertação (mestrado) - UFRJ/Instituto de Matemática, Programa de Pós-graduação em Estatística, 2011.

Referências: f. 75-79

1. Análise numérica - algoritmos. 2. Inferência Bayesiana. 3. Matemática - estatística. 4. Probabilidades. 5. Análise estatística – Tese. I. Gamerman, Dani. II. Universidade Federal do Rio de Janeiro, Instituto de Matemática, Programa de Pós-graduação em Estatística. III. Título

CDD 519.5

*À minha mãe pelo amor e dedicação incondicionais
e à memória de minha avó Jorgina.*

“Nunca deixe que lhe digam
Que não vale a pena acreditar no sonho que se tem
Ou que seus planos nunca vão dar certo
Ou que você nunca vai ser alguém

(...)

Confie em si mesmo
Quem acredita sempre alcança!”

(Trecho da música “Mais uma vez”, composição de Flavio Venturini e Renato Russo)

Agradecimentos

Primeiramente, agradeço a Deus por me permitir chegar onde eu cheguei. À minha mãe, agradeço por seu amor e apoio incondicional, por ter me ensinado a ter força de vontade para vencer e superar os obstáculos que sempre surgem e pelo esforço para que eu tivesse uma boa educação. Ao meu pai, agradeço por todo apoio e torcida.

Agradeço àqueles que perceberam em mim um talento que a minha autocrítica jamais permitiria que se desenvolvesse e que me incentivaram a continuar seguindo o caminho que agora, finalmente, decidi seguir.

Ao meu orientador Professor Dani Gamerman, agradeço pela orientação, exemplo, ensinamentos, dedicação e paciência.

Pelas contribuições dadas a este trabalho, agradeço à Mariana Albi de Oliveira Souza; ao João Batista de Moraes Pereira; à Thainá Alves Malhão por gentilmente me ceder os dados da mortalidade de dengue e ao Professor Roberto de Andrade Medronho pelos ensinamentos passados sobre Epidemiologia. Ao Thiago Guerrera Martins, agradeço por me fornecer a programação utilizada em sua dissertação de Mestrado e pela ajuda dada, mesmo à distância.

A todos os meus amigos, obrigada por fazerem parte da minha vida. Sou grata pela amizade e confiança depositados em mim.

Agradeço à Kelly Cristina Mota Gonçalves, Larissa de Carvalho Alves e ao João Batista de Moraes Pereira, que junto comigo compõem a PANELA, pela amizade e companheirismo. PANELA unida, jamais será vencida. Em especial, agradeço ao João Batista de Moraes Pereira pelo amor, carinho, respeito, paciência e incentivo.

Agradeço às meninas da minha turma de Mestrado: Ana Carolina Carioca da Costa, Carolina Marques Bastos (Jones), Renata Souza Bueno e Priscilla Ferreira da Silva pelos momentos inesquecíveis que passamos. Tivemos que superar muitos desafios e ter bastante paciência para entender que o futuro a gente é quem faz. Sheila Klem Rodrigues das Neves

também deve ser citada, pois apesar de ter entrado da Pós-Graduação antes, também faz parte de nossa turma e compartilhou muitos momentos felizes conosco.

A todos os alunos que frequentam o Laboratório de Pós-Graduação em Estatística, agradeço pelo ambiente de convivência cordial e por tornarem os momentos mais alegres e menos tensos.

Agradeço à UFRJ por ter me proporcionado os melhores anos de minha vida até agora.

Aos professores da UFRJ, em especial, àqueles do Departamento de Métodos Estatísticos e do Programa de Pós-Graduação em Estatística, agradeço por contribuírem de forma exemplar para minha formação Estatística e amadurecimento profissional.

Em particular, à Professora Alexandra Mello Schmidt pelos conselhos, exemplo e paciência; à Professora Beatriz Vaz de Melo Mendes pela oportunidade de ter me apresentado à atividade de pesquisa durante a iniciação científica e à Natalie Haanwinckel Hurtado por todo apoio e incentivo desde a época de caloura do curso de Graduação em Ciências Atuariais na UFRJ.

Aos professores Carlos Antonio Abanto-Valle e Ralph dos Santos Silva, agradeço por aceitarem o convite de fazer parte da minha banca e à Professora Marina Silva Paez pela posição de suplente na banca.

Por fim, agradeço à CAPES e à FAPERJ pelo apoio financeiro, durante o primeiro e o segundo anos de mestrado, respectivamente. Agradeço também ao Instituto de Matemática e ao Departamento de Estatística da UFRJ pelo apoio financeiro para participação de congressos científicos na área de Estatística, o que possibilitou o intercâmbio de ideias entre mim e demais pesquisadores e alunos, certamente contribuindo para a minha formação acadêmica e para o desenvolvimento deste trabalho.

Resumo

A modelagem via espaço de estados é uma metodologia flexível para tratar problemas em análises de séries temporais. Os modelos latentes gaussianos e os não-gaussianos são subclasses dos modelos de espaço de estados. São apresentados procedimentos de inferência bayesiana nesta classe de modelos.

Em problemas de inferência bayesiana, por vezes, as distribuições de interesse não são possíveis de serem obtidas analiticamente, sendo necessários métodos de aproximação. Nesta dissertação, diferentes métodos de aproximação são apresentados, entre eles, métodos determinísticos e de simulação. Em particular, o método de aproximação INLA (do inglês, *integrated nested Laplace approximation*) é descrito e exemplificado na classe modelos latentes gaussianos como alternativa aos métodos de Monte Carlo via cadeias de Markov, amplamente utilizados na literatura estatística.

Por fim, uma proposta de extensão do método de aproximação INLA para viabilizar sua utilização em modelos latentes não-gaussianos é apresentada.

Palavras-Chaves: Modelos de espaço de estados, inferência bayesiana, *integrated nested Laplace approximation*, métodos de Monte Carlo via cadeias de Markov.

Abstract

State-space modeling is a flexible framework in time series analysis. Gaussian and non-Gaussian latent models are subclasses of state-space models. Bayesian inference procedures in this class of models are presented.

In Bayesian inference problems, sometimes, some distributions of interest are not available analytically, so approximated methods are needed. In this work, different approximation methods are presented, among them, deterministic methods and methods of simulation. More specifically, the integrated nested Laplace approximation (INLA) method is described and illustrated in the latent Gaussian class as an alternative to Markov chain Monte Carlo methods, widely used in the statistical literature.

Finally, a proposal to extend the approximation method INLA to allow its use in non-Gaussian latent models is presented.

Keywords: State-space modeling, Bayesian inference, integrated nested Laplace approximation, Markov chain Monte Carlo.

Sumário

1	Introdução	1
2	Análise Numérica	4
2.1	Aproximação de Taylor	4
2.1.1	Aproximação de Taylor de Segunda Ordem	5
2.2	Otimização Sem Restrições	5
2.3	Interpolação	8
2.3.1	Forma de Lagrange	9
2.3.2	<i>Spline</i> Cúbica Interpolante	10
2.4	Integração Numérica	11
2.4.1	Integração por Quadratura	12
3	Métodos Aproximados de Inferência	13
3.1	Métodos Determinísticos	14
3.1.1	Método de Laplace	15
3.1.2	<i>Variational Bayes</i>	17
3.1.3	<i>Expectation-Propagation</i>	18
3.2	Métodos de Simulação Estocástica	19
3.2.1	Algoritmo de Metropolis Hastings	20
3.2.2	Amostrador de Gibbs	21
4	Modelos de Espaço de Estados	22
4.1	Modelos Latentes Gaussianos	23

4.1.1	Modelos Lineares Dinâmicos Normais	24
4.2	Modelos Não-Gaussianos	25
4.3	Família Gama de Modelos Dinâmicos	26
4.4	Dengue na Região Metropolitana do Estado do Rio de Janeiro	28
4.4.1	Introdução	28
4.4.2	Descrição do Modelo	28
4.4.3	Procedimento de Inferência	31
4.4.4	Descrição dos Dados	33
4.4.5	Resultados	35
5	O Método de Aproximação INLA	41
5.1	Definição	41
5.2	Distribuições de Interesse	43
5.3	Parametrização Adequada do Vetor Paramétrico	43
5.4	Esquema de Aproximação	44
5.5	Aproximação para $p(\boldsymbol{\theta} \mathbf{y})$	44
5.5.1	Exploração da Grade de Pontos	45
5.6	Aproximação para $p(x_i \boldsymbol{\theta}, \mathbf{y})$	46
5.6.1	Aproximação Gaussiana	46
5.6.2	Método de Laplace	46
5.6.3	Método de Laplace Simplificado	48
5.7	Algoritmo INLA	50
5.8	Exemplo: Modelo Poisson Dinâmico	50
5.8.1	Resultados	53
6	Extensões	56
6.1	INLA para Modelos Latentes Não-Gaussianos	56
6.2	Esquema de Aproximação	57
6.3	Caso com Independência Condicional	58
6.3.1	Aproximação para $p(\boldsymbol{\theta} \mathbf{y})$	58
6.3.2	Aproximação para $p(x_i \boldsymbol{\theta}, \mathbf{y})$	59

6.4	Caso Geral	60
6.4.1	Aproximação para $p(\boldsymbol{\theta} \mathbf{y})$	60
6.4.2	Aproximação para $p(x_i \boldsymbol{\theta}, \mathbf{y})$	62
6.5	Exemplo: Modelo Poisson não-Gaussiano	62
6.5.1	Descrição do Modelo	62
6.5.2	Procedimento de Inferência	65
7	Considerações Finais	73

Capítulo 1

Introdução

Muitas vezes, não é possível obter a solução analítica de um determinado problema. Em particular, quando se utiliza a abordagem bayesiana em problemas de inferência estatística, a análise se baseia na distribuição *a posteriori*, que contém toda a informação probabilística a respeito do parâmetro de interesse. Por vezes, esta distribuição não pode ser obtida analiticamente, principalmente por depender de integrais que não tem solução ou cuja solução é custosa computacionalmente.

Nestes casos, há a necessidade da utilização de métodos aproximados de inferência, podendo estes ser determinísticos ou de simulação, ou ainda, de simulação estocástica.

A década de 1980 concentrou a maior parte dos trabalhos referentes a aproximações determinísticas. Somente a partir década de 1990, através do avanço tecnológico, os métodos de simulação começaram a ser mais comumente explorados.

Cabe ressaltar, entretanto, que tanto os métodos determinísticos como os métodos de simulação estocástica possuem desvantagens. O primeiro se torna cada vez mais complexo ao se aumentar a dimensão paramétrica e o segundo pode trazer um custo computacional alto em determinados problemas.

Os métodos de Monte Carlo via cadeias de Markov são, ainda, a abordagem mais comumente utilizada na literatura estatística.

Todavia, devido ao fato de muitas vezes a análise se tornar custosa do ponto de vista computacional, há pesquisa estatística a fim de se buscar alternativas para esta abordagem. [Rue et al. \(2009\)](#) propuseram o método determinístico de aproximação INLA (do inglês, *integrated nested Laplace approximation*) como forma de realizar inferência bayesiana de forma aproximada em modelos latentes gaussianos de forma rápida e eficiente.

Os modelos latentes gaussianos fazem parte de uma classe de modelos chamada modelos de espaço de estados, ou ainda, modelos dinâmicos, que nada mais são do que uma metodologia flexível para tratar de séries temporais.

Assim sendo, esta dissertação trata de métodos aproximados de inferência e de modelos de espaços de estados com um foco no método de aproximação INLA para modelos nesta classe. Este método é descrito e exemplificado através de um modelo Poisson dinâmico.

Mais ainda, este trabalho trata de uma proposta para a utilização do método de aproximação INLA também em modelos latentes não-gaussianos.

Estrutura da Dissertação

O Capítulo 2 contém conceitos preliminares, onde estão descritos alguns métodos numéricos, úteis quando da aplicação do método INLA. São citados métodos para solução de problemas de otimização, interpolação polinomial e integração numérica.

O Capítulo 3, por sua vez, fala sobre a necessidade de se utilizar métodos de aproximação, uma vez que pode não ser possível obter de forma analítica algumas distribuições de interesse. Em particular, no caso de inferência bayesiana, por vezes, são necessários métodos de aproximação para a distribuição *a posteriori*. Este capítulo trata de métodos determinísticos como o método de Laplace ([Tierney e Kadane, 1986](#)), o *Variational Bayes* ([Beal, 2003](#)) e o método *Expectation-Propagation* ([Minka, 2001](#)), além dos métodos de Monte Carlo via cadeias de Markov com algoritmos de métodos de simulação estocástica.

Já o Capítulo 4 trata da modelagem via espaço de estados, em que são descritos modelos latentes gaussianos e não-gaussianos. É apresentada a família gama de modelos dinâmicos e é feita uma aplicação dessa modelagem em um problema real.

Uma série temporal de óbitos de dengue e por doenças infecciosas que fazem parte de seu diagnóstico diferencial em indivíduos menores que quinze anos de idade na região metropolitana do estado do Rio de Janeiro é analisada, através de métodos de simulação estocástica.

Em seguida, no Capítulo 5 o método INLA proposto por [Rue et al. \(2009\)](#) é descrito de forma detalhada. É apresentado um exemplo desta abordagem para o modelo Poisson dinâmico.

Por fim, o Capítulo 6 trata de uma proposta para a extensão do método INLA do capítulo anterior, a fim de possibilitar sua aplicação em modelos latentes não-gaussianos.

Capítulo 2

Análise Numérica

Os métodos de análise numérica são usados quando não é possível obter uma solução geral de um determinado problema ou, ainda, quando a forma desta solução é complicada, fazendo com que seu uso não seja prático. Este é um capítulo que contém alguns conceitos preliminares. A importância dele nesta dissertação é que os métodos numéricos aqui descritos devem ser utilizados quando da aplicação do método INLA. Para maiores detalhes sobre esses métodos, sugere-se a leitura dos seguintes livros: [Ruggiero e Lopes \(1996\)](#), [Burden e Faires \(2001\)](#), [Boyd e Vandenberghe \(2004\)](#) e [Nocedal e Wright \(2006\)](#).

2.1 Aproximação de Taylor

Para se aproximar uma função, um dos métodos mais usados é a aproximação de Taylor. Tal aproximação será bastante utilizada quando da aplicação do método INLA.

Seja uma função f tal que $f : [x_0, x] \rightarrow \mathfrak{R}$, em que \mathfrak{R} denota o conjunto dos números reais, e suas n primeiras derivadas são contínuas no intervalo fechado $[x_0, x]$. Além disso, $f^{(n+1)}(\cdot)$ existe em (x_0, x) , sendo $f^{(j)}(\cdot)$ a j -ésima derivada de f . Então, existe um número ξ no intervalo aberto (x_0, x) tal que

$$\begin{aligned} f(x) &= f(x_0) + \frac{(x - x_0)}{1!} f'(x_0) + \frac{(x - x_0)^2}{2!} f''(x_0) + \dots + \\ &+ \frac{(x - x_0)^n}{n!} f^{(n)}(x_0) + R_n(x), \end{aligned} \tag{2.1}$$

em que $R_n(x)$ é conhecido como resto de Lagrange e é dado por $R_n(x) = \frac{(x-x_0)^{n+1}}{(n+1)!} f^{(n+1)}(\xi)$.

2.1.1 Aproximação de Taylor de Segunda Ordem

Um caso particular de interesse é obter uma aproximação polinomial de segundo grau.

Seja uma função $f : \mathfrak{R} \rightarrow \mathfrak{R}$ de classe C^1 e duas vezes diferenciável em x_0 . A aproximação de Taylor de segunda ordem para uma função f de uma variável em torno de x_0 é dada por

$$f(x) \approx f(x_0) + \frac{(x-x_0)}{1!} f'(x_0) + \frac{(x-x_0)^2}{2!} f''(x_0) \quad (2.2)$$

Considere, agora, uma função $f : \mathfrak{R}^n \rightarrow \mathfrak{R}$ de classe C^1 e duas vezes diferenciável em $\mathbf{x}_0 \in \mathfrak{R}^n$. Denotando $\nabla_f(\mathbf{x})$ o vetor gradiente de f em \mathbf{x} e $H_f(\mathbf{x})$ a matriz hessiana de f em \mathbf{x} , a aproximação de Taylor de segunda ordem para uma função de várias variáveis em torno de \mathbf{x}_0 é dada por

$$f(\mathbf{x}) \approx f(\mathbf{x}_0) + \nabla_f(\mathbf{x}_0)^T (\mathbf{x} - \mathbf{x}_0) + \frac{1}{2!} (\mathbf{x} - \mathbf{x}_0)^T H_f(\mathbf{x}_0) (\mathbf{x} - \mathbf{x}_0) \quad (2.3)$$

2.2 Otimização Sem Restrições

Nos problemas de otimização sem restrições, o objetivo é minimizar uma função objetivo que depende de variáveis reais, com nenhuma restrição nessas variáveis. Matematicamente, deseja-se

$$\min_{\mathbf{x}} f(\mathbf{x}), \quad (2.4)$$

em que $\mathbf{x} \in \mathfrak{R}^n$ é um vetor real com $n \geq 1$ componentes e $f : \mathfrak{R}^n \rightarrow \mathfrak{R}$.

Todos os algoritmos de otimização sem restrições exigem o estabelecimento de um valor inicial $\mathbf{x}^{(0)}$. Começando em $\mathbf{x}^{(0)}$, tais algoritmos geram uma sequência $\{\mathbf{x}^{(k)}\}_{k=0}^{\infty}$ que termina conforme uma dada regra de parada. Para mover de uma iteração $\mathbf{x}^{(k)}$ para $\mathbf{x}^{(k+1)}$, os algoritmos usam informação sobre a função f em $\mathbf{x}^{(k)}$ e também informação sobre as iterações anteriores $\mathbf{x}^{(0)}, \mathbf{x}^{(1)}, \dots, \mathbf{x}^{(k-1)}$.

Uma das estratégias para se avançar de um ponto corrente $\mathbf{x}^{(k)}$ para o próximo é a busca em linha. Cada iteração do método de busca em linha obtém uma direção $\mathbf{p}^{(k)}$ e decide quão longe se moverá nesta direção. Cada iteração é dada por

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + t^{(k)} \mathbf{p}^{(k)}, \quad (2.5)$$

em que o escalar positivo $t^{(k)}$ é chamado de tamanho do passo.

Com isso, pode-se resumir tal abordagem através do seguinte algoritmo:

Algoritmo 1 *Algoritmo de Busca em Linha Genérico*

1. Assuma um valor inicial $\mathbf{x}^{(0)}$ que pertence ao domínio da função objetivo.
2. Para $k = 0, 1, 2, \dots$, repita os itens abaixo até que um dado critério de parada seja satisfeito.
 - (i) Determine uma direção $\mathbf{p}^{(k)}$.
 - (ii) Busca em linha: Determine o tamanho do passo $t^{(k)} > 0$.
 - (iii) Atualize $\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + t^{(k)} \mathbf{p}^{(k)}$.

O item (ii) é chamado de busca em linha já que a escolha do tamanho do passo na k -ésima iteração $t^{(k)}$ determina o local ao longo da linha $\{\mathbf{x}^{(k)} + t^{(k)} \mathbf{p}^{(k)} \mid t^{(k)} \in \mathfrak{R}^+\}$ que a próxima iteração estará.

Uma condição de parada possível é dada por $\|\nabla_f(\mathbf{x}^{(k)})\|_2 \leq \epsilon$, fixando-se uma tolerância pequena e positiva ϵ , sendo $\|\cdot\|_2$ a norma Euclidiana.

A maior parte dos algoritmos de busca em linha requerem que $\mathbf{p}^{(k)}$ tenha direção de descida. A direção de Newton é dada por $\mathbf{p}^{(k)} = -H_f^{-1}(\mathbf{x}^{(k)})\nabla_f(\mathbf{x}^{(k)})$. Esta direção pode ser utilizada no método de busca em linha quando $H_f(\mathbf{x}^{(k)})$ é positiva definida. A maioria das implementações do método de Newton utiliza um tamanho de passo $t = 1$ quando possível e ajusta t quando não produz uma redução satisfatória na função f .

Quando $H_f(\mathbf{x})$ não é positiva definida, a direção de Newton pode não estar definida já que $H_f^{-1}(\mathbf{x}^{(k)})$ pode não existir. Neste caso, a direção de Newton é dada por $H_f(\mathbf{x}^{(k)})\mathbf{p}^{(k)} = -\nabla_f(\mathbf{x}^{(k)})$. Fazendo modificações adequadas, pode-se executar o método de busca em linha de Newton através do seguinte algoritmo:

Algoritmo 2 *Algoritmo de Busca em Linha de Newton*

- (i) Assuma um valor inicial $\mathbf{x}^{(0)}$ que pertence ao domínio da função objetivo. Para $k = 0, 1, 2, \dots$, repita os itens (ii)–(iv) até que uma dada condição de parada seja satisfeita.
- (ii) Fatorize a matriz $B^{(k)} = H_f(\mathbf{x}^{(k)}) + E^{(k)}$, em que $E^{(k)} = 0$ se $H_f(\mathbf{x}^{(k)})$ é positiva definida. Caso contrário, $E^{(k)}$ é definida para assegurar que $B^{(k)}$ é positiva definida.
- (iii) Resolva o sistema $B^{(k)}\mathbf{p}^{(k)} = -\nabla_f(\mathbf{x}^{(k)})$.
- (iv) Faça $\mathbf{x}^{(k+1)} \leftarrow \mathbf{x}^{(k)} + \alpha^{(k)}\mathbf{p}^{(k)}$ e $k = k + 1$.

Métodos que usam a direção de Newton tem uma rápida taxa de convergência local, tipicamente quadrática. Após alcançada uma vizinhança da solução, a convergência de alta acurácia é encontrada após poucas iterações. A desvantagem da direção de Newton é a necessidade de se avaliar a matriz hessiana $H_f(\mathbf{x}^{(k)})$ na k -ésima iteração do algoritmo. Este cálculo é sujeito a erros e custoso computacionalmente. Isto pode ser evitado obtendo a matriz hessiana via diferenças finitas.

A maior parte dos algoritmos de busca em linha utilizados na prática são inexatos no sentido de que, em geral, o tamanho do passo é escolhido para aproximadamente minimizar f no raio $\{\mathbf{x}^{(k)} + t^{(k)}\mathbf{p}^{(k)} \mid t^{(k)} \geq 0\}$ ou para suficientemente reduzir f . Um método inexato de busca em linha simples e eficiente é o método de busca em linha retroativa. Este método depende de duas constantes α e β , com $\alpha \in (0, 0.5)$ e $\beta \in (0, 1)$. O algoritmo para a execução deste método é o que segue.

Algoritmo 3 Algoritmo de Busca em Linha Retroativa Genérico

1. Assuma um valor inicial $\mathbf{x}^{(0)}$ que pertence ao domínio da função objetivo. Fixe $\alpha \in (0, 0.5)$ e $\beta \in (0, 1)$
2. Faça $t^{(0)} = 1$. Enquanto $f(\mathbf{x} + t\nabla_f(\mathbf{x})^T) > f(\mathbf{x}) + \alpha t\nabla_f(\mathbf{x})^T$, repita o item 3.
3. $t^{(k+1)} = \beta t^{(k)}$

O método acima é chamado de método de busca em linha retroativa pois inicia-se com um tamanho de passo unitário $t^{(0)} = 1$

2.3 Interpolação

Interpolar uma função $f(x)$ consiste em aproximar essa função por uma outra função $g(x)$, escolhida entre uma classe de funções definida *a priori* que satisfaça algumas propriedades, segundo [Ruggiero e Lopes \(1996\)](#). Uma função $g(x)$ é usada em substituição à função $f(x)$.

A necessidade de se fazer esta substituição pode ocorrer quando operações de diferenciação ou integração são difíceis ou impossíveis de serem realizadas através da função de estudo ou, ainda, quando são conhecidos apenas alguns valores da função e deseja-se obter o valor da função em um ponto que não é conhecido.

Através de amostragem ou experimentação, em diversas aplicações, obtém-se um conjunto de dados e tenta-se construir uma função que melhor se ajuste a este conjunto de dados. A interpolação é um caso específico de ajuste de curvas em que a função deve passar exatamente pelos pontos dados.

Dados os $(n + 1)$ pontos distintos em $[a, b]$, $a = x_0 < x_1 < x_2 < \dots < x_n = b$, a interpolação polinomial aproxima uma dada função $f(x)$ por um polinômio de $p_n(x)$ de grau menor ou igual a n , tal que $f(x_k) = p_n(x_k)$, $k = 0, 1, 2, \dots, n$. Pode-se mostrar que existe um único polinômio $p_n(x)$ de grau menor ou igual a n que satisfaz $f(x_k) = p_n(x_k)$, $k = 0, 1, 2, \dots, n$, desde que $x_k \neq x_j, k \neq j$.

Assim, o polinômio $p_n(x)$ é único, no entanto, há diversas formas de obtê-lo, como por exemplo as formas de Lagrange e de Newton. Sugere-se a leitura de [Ruggiero e Lopes \(1996\)](#) ou [Burden e Faires \(2001\)](#), que são as referências principais desta seção, para obter mais detalhes.

2.3.1 Forma de Lagrange

Considere novamente os $(n + 1)$ pontos distintos em $[a, b]$, $a = x_0 < x_1 < x_2 < \dots < x_n = b$ e que $y_k = f(x_k)$, para $k = 0, 1, 2, \dots, n$. Seja $p_n(x)$ o polinômio de grau n que interpola f em x_0, x_1, \dots, x_n . A forma de Lagrange para o polinômio interpolador é dada por

$$p_n(x) = \sum_{k=0}^n y_k L_k(x), \quad (2.6)$$

em que

$$L_k(x) = \frac{(x - x_0)(x - x_1) \dots (x - x_{k-1})(x - x_{k+1}) \dots (x - x_n)}{(x_k - x_0)(x_k - x_1) \dots (x_k - x_{k-1})(x_k - x_{k+1}) \dots (x_k - x_n)} \quad (2.7)$$

O erro em se aproximar $f(x)$ por $p_n(x)$ é dado por $E_n(x) = f(x) - p_n(x)$, para todo x no intervalo $[x_0, x_n]$.

Um dos métodos mais simples de interpolação é a interpolação linear. A interpolação linear é frequentemente usada para aproximar o valor de alguma função f em um dado ponto x usando dois valores conhecidos da função, por exemplo, x_0 e x_1 . Utilizando-se a forma de Lagrange, o polinômio interpolador de grau 1 é dado por:

$$p_1(x) = f(x_0) + \frac{f(x_1) - f(x_0)}{x_1 - x_0}(x - x_0). \quad (2.8)$$

De forma mais geral, a interpolação linear em um conjunto de pontos $(x_0, f(x_0)), (x_1, f(x_1)), \dots, (x_n, f(x_n))$ é obtida fazendo-se a interpolação linear entre cada par de pontos dados.

2.3.2 Spline Cúbica Interpolante

Através da forma de Lagrange, obtém-se um polinômio que interpola a função f em pontos dados. Trata-se, portanto, de uma aproximação em intervalos fechados usando polinômios.

Os polinômios de ordem elevada tem natureza oscilatória, no sentido de que uma flutuação em uma pequena parte do intervalo pode levar a grandes flutuações em toda a amplitude do intervalo.

Uma alternativa é interpolar $f(x)$ em grupos de poucos pontos, obtendo polinômios de graus menores, além de impor condições para que sejam contínuas tanto a função de aproximação como suas derivadas até uma determinada ordem.

Neste caso, pode-se considerar uma interpolação linear entre cada par de pontos dados. A função obtida é contínua porém não há diferenciação nos extremos dos subintervalos. Uma outra possibilidade é aproximar um polinômio de grau 2 em cada trio de pontos dados. No entanto, assim como no caso anterior, também só seria garantida a continuidade da função que aproxima $f(x)$.

Considere novamente uma função $f(x)$ nos seguintes pontos dados $a = x_0 < x_1 < \dots < x_n = b$. Denota-se por $S_p(x)$ uma função *spline* de grau p . Em cada subintervalo $[x_i, x_{i+1}]$, $i = 0, 1, \dots, n-1$, $S_p(x)$ é um polinômio de grau p , sendo esta função contínua e com derivada contínua até a ordem $(p-1)$ em $[a, b]$. Se além dessas condições a função for interpolante, ou seja, satisfizer a condição de que $S_p(x_i) = f(x_i)$, $i = 0, 1, 2, \dots, n$, então ela será chamada de *spline* interpolante.

A função *spline* linear possui derivada primeira descontínua nos pontos dados. Já a função *spline* quadrática possui derivada contínua até a ordem 1, o que pode levar o fato da curvatura de $S_2(x)$ mudar nos pontos dados. Pelo que foi dito, a função *spline* cúbica é mais utilizada na prática.

Supondo que se saiba os valores de $f(x)$ nos pontos x_i , $i = 0, 1, 2, \dots, n$, a função $S_3(x)$ é chamada *spline* cúbica interpolante de $f(x)$ nos pontos dados x_i , $i = 0, 1, 2, \dots, n$ se existem n polinômios de grau 3 $s_k(x)$, $k = 1, 2, \dots, n$ tais que são válidas as seguintes relações:

1. $S_3(x) = s_k(x)$, $x \in [x_{k-1}, x_k]$, $k = 1, 2, \dots, n$

2. $S_3(x_i) = f(x_i)$, $i = 0, 1, 2, \dots, n$

3. $s_k(x_k) = s_{k+1}(x_k)$, $k = 1, 2, \dots, n - 1$
4. $s'_k(x_k) = s'_{k+1}(x_k)$, $k = 1, 2, \dots, n - 1$
5. $s''_k(x_k) = s''_{k+1}(x_k)$, $k = 1, 2, \dots, n - 1$
6. $S''_3(x_0) = \alpha$ e $S''_3(x_n) = \beta$. Quando $\alpha = \beta = 0$, então a função *spline* cúbica é chamada de *spline* cúbica natural.

Assim, uma função *spline* cúbica é uma função polinomial por partes, contínua e com derivada contínua. Pode-se escrever $s_k(x)$ por

$$s_k(x) = a_k(x - x_k)^3 + b_k(x - x_k)^2 + c_k(x - x_k) + d_k, \quad (2.9)$$

$k = 1, 2, \dots, n$.

Para determinar a função *spline* cúbica, é necessário obter a_k , b_k , c_k e d_k , para cada $k = 1, 2, \dots, n$. Para isto, basta impor válidas as relações anteriores.

2.4 Integração Numérica

Suponha que o interesse é em resolver integrais do tipo

$$\int_a^b f(x)dx. \quad (2.10)$$

Por vezes, resolver tal integral pode não ser fácil ou, ainda, pode-se conhecer os valores da função $f(x)$ em apenas alguns pontos no intervalo $[a, b]$. Neste último caso, como não se tem a expressão analítica de $f(x)$, não é possível resolver (2.10).

Através da integração numérica, o objetivo será substituir a função $f(x)$ por um polinômio que a aproxime no intervalo $[a, b]$. Assim, tem-se que para $x_i \in [a, b]$, $i = 1, 2, \dots, n$

$$\int_a^b f(x)dx \approx \sum_{i=1}^n f(x_i)\Delta_i. \quad (2.11)$$

2.4.1 Integração por Quadratura

Uma regra de quadratura é uma aproximação de uma integral definida de uma função por uma soma ponderada da função em valores específicos dentro do domínio de integração.

Em particular, considere a regra de quadratura gaussiana de n pontos. O domínio de integração de tal regra é tomado como $[-1, 1]$, de modo que

$$\int_{-1}^1 f(x)dx \approx \sum_{i=1}^n w_i f(x_i). \quad (2.12)$$

Caso seja possível reescrever $f(x)$ tal que $f(x) = W(x)g(x)$, em que $W(x)$ é conhecida como função peso, então os pesos w_i são tais que

$$\int_{-1}^1 f(x)dx = \int_{-1}^1 W(x)g(x)dx \approx \sum_{i=1}^n w_i g(x_i). \quad (2.13)$$

Já o método de integração por quadratura conhecido por integração de Gauss-Hermite é um método de quadratura gaussiana no intervalo $(-\infty, +\infty)$, em que se considera a função peso por $W(x) = e^{-x^2}$.

Capítulo 3

Métodos Aproximados de Inferência

Seja \mathbf{Y} uma variável aleatória ou vetor aleatório com função de probabilidade ou função de densidade de probabilidade $p(\mathbf{Y}|\Psi)$, em que Ψ é um parâmetro ou vetor de parâmetros que caracteriza a distribuição de \mathbf{Y} . Considerando a abordagem bayesiana, antes da observação dos dados, assume-se uma distribuição de probabilidade para Ψ , chamada de distribuição *a priori*, referente à incerteza inicial a respeito de Ψ . Tendo observado os dados \mathbf{y} , através do Teorema de Bayes, encontra-se a distribuição *a posteriori* de Ψ , dada por

$$p(\Psi|\mathbf{y}) = \frac{p(\mathbf{y}|\Psi)p(\Psi)}{\int p(\mathbf{y}|\Psi)p(\Psi)d\Psi}. \quad (3.1)$$

A inferência estatística sob o enfoque bayesiano baseia-se na distribuição *a posteriori* de Ψ , que contém toda a informação probabilística a respeito do parâmetro de interesse. É possível resumir a informação contida na distribuição *a posteriori* através de estimação pontual ou por intervalo. [Migon e Gamerman \(1999\)](#) apresentam uma descrição detalhada sobre conceitos de inferência bayesiana.

Considere $\Psi = (\Psi_1, \dots, \Psi_k)'$. Por vezes, deseja-se obter distribuições *a posteriori* marginais, ou seja, $p(\Psi_{\mathbf{I}}|\mathbf{y})$, em que $\mathbf{I} \subseteq \{1, \dots, k\}$. Denotando $\bar{\mathbf{I}}$ o conjunto complementar a \mathbf{I} em relação a $\{1, \dots, k\}$, então pode-se obter a distribuição *a posteriori* marginal de $\Psi_{\mathbf{I}}$ da seguinte forma

$$p(\Psi_{\mathbf{I}}|\mathbf{y}) = \int p(\Psi|\mathbf{y})d\Psi_{\bar{\mathbf{I}}} \quad (3.2)$$

Um problema que surge é que pode não ser possível resolver a integração (3.2) de forma analítica. Muitas vezes, (3.1) também não pode ser resolvida analiticamente. Sendo assim, são necessários métodos de aproximação para que a distribuição de interesse possa ser obtida.

Os métodos de aproximação para a distribuição de interesse podem ser divididos em métodos determinísticos e métodos de simulação, em particular, de simulação estocástica.

As aproximações determinísticas foram desenvolvidas principalmente na década de 1980. Até então, os métodos computacionalmente intensivos estavam no início de seu desenvolvimento. A complexidade das técnicas determinísticas cresce à medida que a dimensão do espaço paramétrico cresce.

Os métodos de inferência baseados em simulação estocástica utilizam amostras da distribuição *a posteriori* para sumarizar informação. Esses métodos são utilizados para obter uma aproximação da distribuição *a posteriori* quando não é possível obtê-la analiticamente ou como alternativa aos métodos numéricos devido a altas dimensões paramétricas.

A partir década de 1990 com o avanço das técnicas computacionais e com o trabalho de Gelfand e Smith (1990), esses métodos começaram a ser extensamente explorados. Esta metodologia de aproximação pode ser excessivamente custosa computacionalmente em determinadas situações, o que foi uma das motivações para o estudo de métodos de aproximação alternativos.

Rue et al. (2009) propuseram o método de aproximação INLA como alternativa aos métodos baseados em simulação estocástica, permitindo realizar inferência bayesiana de forma aproximada em modelos latentes gaussianos de forma eficiente.

Nas seções que seguem, uma breve revisão de literatura é feita sobre os métodos de inferência determinísticos e de simulação estocástica, com o objetivo de entender o contexto histórico ao qual o desenvolvimento dos métodos aproximados de inferência se inserem.

3.1 Métodos Determinísticos

Como já foi comentado, em aplicações de inferência Bayesiana, muitas vezes as distribuições de interesse, como a distribuição *a posteriori* ou preditiva não podem ser obtidas de forma fechada, sendo necessários métodos de aproximação.

Desde a década de 1980, muitos trabalhos surgiram na literatura abordando aproximações numéricas. Pode-se citar os trabalhos de [Naylor e Smith \(1982\)](#), [Tierney e Kadane \(1986\)](#), [Smith et al. \(1987\)](#), [Tierney et al. \(1989\)](#). Na Seção 3.1.1, o método de Laplace proposto por [Tierney e Kadane \(1986\)](#) será descrito.

Nesta época, os avanços tecnológicos computacionais que viabilizariam a utilização de métodos de simulação estocástica estavam apenas começando. Como já foi dito, somente a partir da década de 1990 que os métodos baseados em simulação estocástica começaram a ser explorados.

[Rue et al. \(2009\)](#) propõem utilizar o método determinístico INLA de maneira rápida e precisa em abordagens inferência bayesiana para a classe de modelos latentes gaussianos. Este método é central nesta dissertação e, portanto, será descrito no Capítulo 5.

Por outro lado, outros métodos determinísticos vem sido explorados na literatura, principalmente na literatura de aprendizagem de máquina. Estas abordagens também podem ser utilizadas em problemas oriundos de aplicações em inferência bayesiana. Por exemplo, a abordagem através *variational Bayes* tem sido explorada na literatura e pode-se citar como referência os trabalhos de [Bernardo et al. \(2003\)](#) e [Beal \(2003\)](#). Esta abordagem será discutida na Seção 3.1.2, onde se faz uma breve revisão desta família de técnicas usada para aproximar integrais intratáveis. Além disso, a Seção 3.1.3 fala um pouco sobre o método proposto por [Minka \(2001\)](#), chamado *Expectation-Propagation*.

3.1.1 Método de Laplace

Considere g uma função suave e positiva em um espaço paramétrico. A média *a posteriori* de $g(\Psi)$ pode ser escrita por

$$E[g(\Psi)] = \int g(\Psi)p(\Psi|\mathbf{y})d\Psi = \frac{\int g(\Psi)l(\Psi;\mathbf{y})p(\Psi)d\Psi}{\int l(\Psi;\mathbf{y})p(\Psi)d\Psi}, \quad (3.3)$$

em que $p(\Psi|\mathbf{y})$ é a função de densidade *a posteriori* de Ψ , $l(\Psi;\mathbf{y})$ a função de verossimilhança e $p(\Psi)$ a função de densidade *a priori* de Ψ .

Em geral, as integrais do numerador e denominador de (3.3) não podem ser resolvidas de forma exata. Uma solução possível é aproximar o denominador de (3.3) pela integral de uma curva normal centrada na moda *a posteriori* e variância dada pelo inverso da segunda derivada

do logaritmo da distribuição *a posteriori* na sua moda. Essa aproximação pode ser vista como aplicação do método de Laplace para integrais, como descrito em [de Bruijn \(1961\)](#) *apud* [Tierney e Kadane \(1986\)](#). Os resultados serão razoáveis desde que a distribuição *a posteriori* seja unimodal ou dominado por uma única moda.

Além da aproximação feita no denominador de (3.3), [Tierney e Kadane \(1986\)](#) propuseram localizar a moda do numerador de (3.3), achar a segunda derivada nesta moda e então aproximar esta integral através de uma segunda aplicação do método de Laplace. Assim, aplicando a mesma técnica de aproximação no numerador e no denominador, os erros associados às aproximações serão similares e a razão dessas aproximações fará com que parte desses erros se cancelem.

Em particular, o método de Laplace pode ser utilizado para aproximar distribuições *a posteriori* marginais. Segundo [Tierney e Kadane \(1986\)](#), para obter a distribuição *a posteriori* marginal de Ψ_1 , particione o espaço paramétrico Ψ de tal forma que $\Psi = (\Psi_1, \Psi_2)$. Seja $\hat{\Psi} = (\hat{\Psi}_1, \hat{\Psi}_2)$ a moda *a posteriori* e Σ menos o inverso da matriz Hessiana de $(\log(p(\Psi)) + \log(l(\Psi; \mathbf{y}))) / n$ na moda $\hat{\Psi}$. Para um dado Ψ_1 , seja $\hat{\Psi}_2^* = \hat{\Psi}_2^*(\Psi_1)$ o argumento que maximiza a função $h(\cdot) = p(\Psi_1, \cdot)l(\Psi_1, \cdot; \mathbf{y})$, em que n é a dimensão dos dados, e seja $\Sigma^* = \Sigma^*(\Psi_1)$ menos o inverso da Hessiana de $\log(h(\cdot)) / n$. Aplicando o método de Laplace nas integrais do numerador e denominador da distribuição marginal de Ψ_1 ,

$$p(\Psi_1 | \mathbf{y}) = \frac{\int l((\Psi_1, \Psi_2); \mathbf{y})p(\Psi_1, \Psi_2)d\Psi_2}{\int l(\Psi; \mathbf{y})p(\Psi)d\Psi}, \quad (3.4)$$

tem-se a seguinte aproximação

$$\tilde{p}(\Psi_1 | \mathbf{y}) = \left(\frac{|\Sigma^*(\Psi_1)|}{2\pi |\Sigma|} \right)^{1/2} \frac{l((\Psi_1, \hat{\Psi}_2^*); \mathbf{y})p(\Psi_1, \hat{\Psi}_2^*)}{l(\hat{\Psi}; \mathbf{y})p(\hat{\Psi})} \quad (3.5)$$

Esta aproximação apresenta erro da ordem $O(n^{-3/2})$ e o tempo computacional necessário para obter essa aproximação cresce à medida que a dimensão do espaço paramétrico cresce.

3.1.2 Variational Bayes

O método VB (do inglês, *Variational Bayes*) é uma forma de realizar inferência de forma aproximada. Este método foi desenvolvido na literatura de aprendizagem de máquina (em inglês, *machine learning*). O método trata do problema de aproximar integrais intratáveis provenientes de problemas em estatística bayesiana e aprendizagem de máquina. A metodologia VB já produziu resultados em diversas áreas, tais como modelos ocultos de Markov (em inglês, *hidden Markov models*), modelos de mistura, modelos gráficos e modelos de espaço de estados.

Sendo \mathbf{y} o vetor de observações, \mathbf{x} o vetor de variáveis latentes e $\boldsymbol{\theta}$ os hiperparâmetros, considere a distribuição *a posteriori* $p(\mathbf{x}, \boldsymbol{\theta} \mid \mathbf{y})$ de um modelo genérico. O método VB considera a seguinte aproximação

$$p(\mathbf{x}, \boldsymbol{\theta} \mid \mathbf{y}) \approx q(\mathbf{x}, \boldsymbol{\theta}) \quad (3.6a)$$

$$= \arg \min_{q(\mathbf{x}, \boldsymbol{\theta})} KL[q(\mathbf{x}, \boldsymbol{\theta}) \parallel p(\mathbf{x}, \boldsymbol{\theta} \mid \mathbf{y})] \quad (3.6b)$$

A medida de divergência de *Kullback-Leibler* (Kullback e Leibler, 1951) mede a discrepância entre $p(\mathbf{x}, \boldsymbol{\theta} \mid \mathbf{y})$ e $q(\mathbf{x}, \boldsymbol{\theta})$. O objetivo, então, é achar a aproximação $q(\mathbf{x}, \boldsymbol{\theta})$ que minimiza $KL(q \parallel p)$, isto é, a divergência de *Kullback-Leibler* de $p(\mathbf{x}, \boldsymbol{\theta} \mid \mathbf{y})$ com respeito a $q(\mathbf{x}, \boldsymbol{\theta})$.

A minimização é feita com restrição e, em geral, assume-se que os estados e os hiperparâmetros são independentes, ou seja, $q(\mathbf{x}, \boldsymbol{\theta}) = q_x(\mathbf{x})q_\theta(\boldsymbol{\theta})$. Espera-se que as marginais q_x e q_θ aproximem bem as verdadeiras distribuições *a posteriori* marginais. Para o problema de minimização, pode ser utilizado algum algoritmo iterativo, tal como o algoritmo EM (Dempster et al., 1977). Para uma revisão detalhada sobre o método VB sugere-se Beal (2003).

Alguns problemas em se utilizar a abordagem via *variational Bayes* podem ser citados. O processo iterativo básico do método VB é tratável apenas sob conjugação de $p(\theta)$ com respeito a $p(\mathbf{x}, \mathbf{y} \mid \theta)$ e desde que $p(\mathbf{x}, \mathbf{y} \mid \theta)$ pertença à família exponencial (Beal, 2003). Rue et al. (2009) demonstraram que para modelos latentes gaussianos, ainda que o método VB aproxime bem a moda *a posteriori*, a variância *a posteriori* é subestimada.

Como o método é aproximado, o custo computacional é significativamente inferior às abordagens via MCMC e é também inferior à aproximação de Laplace, já que não é necessário o cálculo da matriz hessiana. Rue et al. (2009) mostraram, entretanto, que o método VB é muito mais lento do que o INLA.

3.1.3 Expectation-Propagation

Além do método *Variational Bayes* descrito na Seção 3.1.2, um outro esquema de aproximação comum em aplicações de aprendizado de máquina é a abordagem EP (do inglês, *Expectation-Propagation*), proposto por Minka (2001).

O método EP, similarmente ao VB, minimiza iterativamente a pseudo-distância entre $p(\mathbf{x}, \boldsymbol{\theta} \mid \mathbf{y})$ e a aproximação $q(\mathbf{x}, \boldsymbol{\theta})$, sujeito a uma fatorização de $q(\mathbf{x}, \boldsymbol{\theta})$ em que cada um dos fatores paramétricos envolva apenas um único componente de $(\mathbf{x}, \boldsymbol{\theta})$.

O método EP considera a seguinte aproximação

$$p(\mathbf{x}, \boldsymbol{\theta} \mid \mathbf{y}) \approx q(\mathbf{x}, \boldsymbol{\theta}) \quad (3.7a)$$

$$= \arg \min_{q(\mathbf{x}, \boldsymbol{\theta})} KL[p(\mathbf{x}, \boldsymbol{\theta} \mid \mathbf{y}) \parallel q(\mathbf{x}, \boldsymbol{\theta})] \quad (3.7b)$$

Neste caso, deseja-se achar a aproximação $q(\mathbf{x}, \boldsymbol{\theta})$ que minimiza a divergência de *Kullback-Leibler* de $q(\mathbf{x}, \boldsymbol{\theta})$ com respeito a $p(\mathbf{x}, \boldsymbol{\theta} \mid \mathbf{y})$ e não o contrário como é feito no método VB. Como visto em Bishop (2006) *apud* Rue et al. (2009), por conta disso, o método EP, em geral, superestima a variância *a posteriori*.

3.2 Métodos de Simulação Estocástica

No contexto bayesiano, os métodos de simulação estocástica utilizam amostras da distribuição *a posteriori* para sumarizar informação quando não é possível obter a distribuição *a posteriori* de forma analítica. [Gamerman e Lopes \(2006\)](#) apresentam uma descrição detalhada dos métodos de simulação estocástica, em especial, dos métodos de Monte Carlo via cadeias de Markov.

Os métodos de Monte Carlo via cadeias de Markov (MCMC) são métodos de simulação estocástica que consistem na construção de uma cadeia de Markov que, por meio de escolhas adequadas de núcleos de transição, tenha como distribuição estacionária a distribuição de interesse. No contexto de estimação bayesiana, a distribuição *a posteriori*. Mais especificamente, os métodos MCMC devem produzir uma cadeia de Markov homogênea, ergódica e irredutível cuja distribuição estacionária seja igual à distribuição de interesse.

Uma vez que a convergência da cadeia tenha sido atingida, as amostras estarão sendo geradas da distribuição estacionária. Para aproximar a distribuição *a posteriori*, utilizam-se amostras suficientemente grandes da distribuição estacionária.

Dentre os métodos mais utilizados para a construção da cadeia de Markov está o Amostrador de Gibbs proposto por [Geman e Geman \(1984\)](#), sendo subsequentemente popularizado [Gelfand e Smith \(1990\)](#), e o método de Metropolis-Hastings proposto por [Metropolis et al. \(1953\)](#) e estendido por [Hastings \(1970\)](#).

A difusão da aplicação destes métodos foi iniciada principalmente com o trabalho de [Gelfand e Smith \(1990\)](#), no qual foi feita uma comparação entre o amostrador de Gibbs, proposto inicialmente por [Geman e Geman \(1984\)](#), com outros esquemas de simulação estocástica. Até então, os trabalhos desenvolvidos eram baseados principalmente em aproximações numéricas e analíticas. O avanço computacional na década de 1990 facilitou a popularização de aplicações dos métodos bayesianos.

3.2.1 Algoritmo de Metropolis Hastings

Um dos métodos de Monte Carlo via cadeias de Markov é o algoritmo de Metropolis-Hastings, inicialmente proposto por [Metropolis et al. \(1953\)](#) e posteriormente estendido por [Hastings \(1970\)](#). Utiliza-se este método para gerar amostras de uma distribuição de interesse a qual não é possível de ser obtida analiticamente. Em sua aplicação, utiliza-se uma distribuição auxiliar, chamada de distribuição proposta, para gerar um valor proposto desta distribuição e aceitá-lo ou não de acordo com uma probabilidade. Este mecanismo de correção garante a convergência da cadeia para a distribuição de equilíbrio.

Denote $p(\cdot)$ como a função de probabilidade ou função de densidade de probabilidade da distribuição de interesse e $q(\cdot)$ a função de probabilidade ou função de densidade de probabilidade proposta. O algoritmo de Metropolis Hastings pode ser descrito através dos seguintes passos:

- (i) Determina-se um valor inicial arbitrário $\Psi^{(0)}$ para Ψ e inicializa-se o contador $j = 1$.
- (ii) Gera-se um valor ξ da distribuição proposta $q(\xi | \Psi^{(j-1)})$.
- (iii) Aceita-se o valor gerado em (ii) com probabilidade $\min \left\{ 1, \frac{p(\xi)q(\Psi^{(j-1)}|\xi)}{p(\Psi^{(j-1)})q(\xi|\Psi^{(j-1)})} \right\}$. Se o valor for aceito, $\Psi^{(j)} = \xi$. Caso contrário, a cadeia não se move e $\Psi^{(j)} = \Psi^{(j-1)}$.
- (iv) Atualiza-se o contador de j para $j + 1$ e retorna-se ao passo (ii) até que a convergência seja obtida.

Supondo que a partir da iteração J a cadeia atingiu a convergência para a distribuição de equilíbrio, os valores $\Psi^J, \Psi^{J+1}, \dots, \Psi^{j+M}$ constituem uma amostra aproximada da distribuição *a posteriori* de Ψ .

3.2.2 Amostrador de Gibbs

O amostrador de Gibbs é um esquema de simulação estocástica usando cadeias de Markov. Pode-se utilizar este esquema para amostrar de uma distribuição conjunta de interesse $p(\Psi)$ se soubermos as distribuições condicionais (completas) para cada parâmetro, já que neste método o núcleo de transição da cadeia é formado pelas distribuições condicionais (completas) das componentes Ψ_j de um vetor paramétrico $\Psi, \forall j$. Para cada parâmetro, a distribuição condicional completa é a distribuição deste parâmetro condicional à informação de todos os outros parâmetros, podendo ser denotada por $p(\Psi_j | \Psi_{-j})$, em que $\Psi_{-j} = (\Psi_1, \dots, \Psi_{j-1}, \Psi_{j+1}, \dots, \Psi_k)$.

O amostrador de Gibbs foi originalmente introduzido por [Geman e Geman \(1984\)](#) e posteriormente popularizado no trabalho de [Gelfand e Smith \(1990\)](#) que comparou este esquema com outros de simulação estocástica.

O algoritmo amostrador de Gibbs pode ser descrito através dos seguintes passos:

- (i) Determina-se um valor arbitrário inicial para $\Psi^{(0)}$ para Ψ e inicializa-se o contador $j = 1$;
- (ii) Obtém-se um novo valor $\Psi^{(j)}$ através de gerações sucessivas de

$$\begin{aligned}
 \Psi_1^{(j)} &\sim p(\Psi_1 | \Psi_2^{(j-1)}, \dots, \Psi_k^{(j-1)}) \\
 \Psi_2^{(j)} &\sim p(\Psi_2 | \Psi_1^{(j)}, \Psi_3^{(j-1)}, \dots, \Psi_k^{(j-1)}) \\
 \Psi_3^{(j)} &\sim p(\Psi_3 | \Psi_1^{(j)}, \Psi_2^{(j)}, \Psi_4^{(j-1)}, \dots, \Psi_k^{(j-1)}) \\
 &\vdots \\
 \Psi_k^{(j)} &\sim p(\Psi_k | \Psi_1^{(j)}, \Psi_2^{(j)}, \dots, \Psi_{k-1}^{(j)}).
 \end{aligned}$$

- (iii) Atualiza-se o contador de j para $j + 1$ e retorna-se ao passo (ii) até que a convergência seja obtida.

Capítulo 4

Modelos de Espaço de Estados

Uma série temporal é uma coleção de observações feitas sequencialmente ao longo do tempo, em que há dependência entre observações vizinhas. Dados de séries temporais surgem em várias áreas como na Economia com preços diários de ações, na Atuária com o número diário de sinistros registrados por uma seguradora, na Epidemiologia com o número mensal de novos casos de uma determinada doença, e em várias outras áreas do conhecimento.

A modelagem via espaço de estados é uma metodologia flexível em problemas de análises de séries temporais. Assume-se que o desenvolvimento do sistema em estudo ao longo do tempo pode ser determinado por componentes não-observáveis (campo latente) que se relacionam com a série de observações. Assim, o objetivo da análise via espaço de estados é inferir propriedades importantes sobre os componentes não-observáveis através do conhecimento das observações.

Na literatura Bayesiana, os modelos de espaço de estados são também conhecidos como modelos dinâmicos. Tais modelos consideram uma variação temporal em seus parâmetros. Distribuições de probabilidade são assumidas tanto para os dados como para a variação dos parâmetros (de estado).

Dentre os componentes não-observáveis mais comuns que uma série temporal pode ser decomposta, incluem *tendência*, que se refere à direção geral segundo a qual a série temporal se desenvolve ao longo de um intervalo de tempo; *sazonalidade*, que se refere a padrões semelhantes que uma série temporal pode apresentar durante intervalos sucessivos de tempo; *componente cíclica*, que se refere às oscilações em longo prazo ou aos desvios da tendência

ou, ainda, a *componente aleatória*.

Os modelos latentes gaussianos e os não-gaussianos são subclasses dos modelos de espaço de estados. Uma breve descrição desses modelos pode ser vista nas próximas seções.

4.1 Modelos Latentes Gaussianos

Os modelos latentes gaussianos são uma classe de modelos hierárquicos em que se assume uma distribuição normal multivariada para o campo latente \mathbf{x} de dimensão n , sendo esta controlada um vetor de parâmetros desconhecidos $\boldsymbol{\theta}$, denominados hiperparâmetros, parâmetros estáticos ou parâmetros fixos.

Muitas vezes, assume-se que as variáveis observadas $\{y_i : i \in I\}$ são condicionalmente independentes dado \mathbf{x} e $\boldsymbol{\theta}$. Neste caso, tem-se a seguinte estrutura hierárquica:

$$y_i | \mathbf{x}, \boldsymbol{\theta} \sim p(y_i | x_i, \boldsymbol{\theta}) \quad (4.1a)$$

$$\mathbf{x} | \boldsymbol{\theta} \sim p(\mathbf{x} | \boldsymbol{\theta}) \quad (4.1b)$$

$$\boldsymbol{\theta} \sim p(\boldsymbol{\theta}) \quad (4.1c)$$

em que $\mathbf{y} = (y_1, \dots, y_{n_d})'$ são os dados observados, $\mathbf{x} = (x_1, \dots, x_n)'$ os elementos do campo latente, $\boldsymbol{\theta}$ é o vetor m -dimensional de hiperparâmetros, a distribuição de $y_i | \mathbf{x}, \boldsymbol{\theta}$ pertence à família exponencial e a distribuição de $\mathbf{x} | \boldsymbol{\theta}$ é gaussiana.

A distribuição *a posteriori* é dada por:

$$p(\mathbf{x}, \boldsymbol{\theta} | \mathbf{y}) \propto p(\boldsymbol{\theta}) p(\mathbf{x} | \boldsymbol{\theta}) \prod_{i \in I} p(y_i | x_i, \boldsymbol{\theta}) \quad (4.2a)$$

$$\propto p(\boldsymbol{\theta}) |\mathbf{Q}(\boldsymbol{\theta})|^{1/2} \exp \left\{ -\frac{1}{2} \mathbf{x}^T \mathbf{Q}(\boldsymbol{\theta}) \mathbf{x} + \sum_{i \in I} \log(p(y_i | x_i, \boldsymbol{\theta})) \right\}, \quad (4.2b)$$

sendo a distribuição *a priori* de $\mathbf{x} | \boldsymbol{\theta}$ dada por uma normal com média zero e matriz de precisão $\mathbf{Q}(\boldsymbol{\theta})$.

Muitos modelos utilizados na literatura possuem estrutura hierárquica similar à descrita nesta seção, como por exemplo, os modelos de regressão, os modelos dinâmicos e os modelos espaço-temporais.

4.1.1 Modelos Lineares Dinâmicos Normais

Um caso particular dos modelos latentes gaussianos é o modelo linear dinâmico normal, em que se assume uma distribuição normal para as observações. Considere uma série temporal $\mathbf{Y}_1, \mathbf{Y}_2, \dots$ em que \mathbf{Y}_t , para $t = 1, 2, \dots$, é o vetor de observações r -dimensional. Um modelo linear dinâmico normal pode ser caracterizado através das seguintes equações

$$\mathbf{Y}_t = \mathbf{F}'_t \mathbf{x}_t + \mathbf{v}_t, \quad \mathbf{v}_t \sim N(\mathbf{0}, \mathbf{V}_t) \quad (4.3a)$$

$$\mathbf{x}_t = \mathbf{G}_t \mathbf{x}_{t-1} + \boldsymbol{\omega}_t, \quad \boldsymbol{\omega}_t \sim N(\mathbf{0}, \mathbf{W}_t), \quad (4.3b)$$

em que, para $t = 1, 2, \dots$, \mathbf{x}_t é um vetor n -dimensional denominado vetor de estados; \mathbf{F}_t é uma matriz $(n \times r)$ cujos elementos são valores conhecidos; \mathbf{G}_t é uma matriz quadrada de ordem n conhecida que descreve a evolução temporal dos parâmetros de estado; \mathbf{V}_t é uma matriz de covariância $r \times r$ conhecida associada ao erro observacional \mathbf{v}_t e \mathbf{W}_t é uma matriz de covariância $n \times n$ conhecida associada ao erro de evolução dos estados $\boldsymbol{\omega}_t$.

O modelo descrito em (4.3) é completamente especificado através da quádrupla $\{\mathbf{F}, \mathbf{G}, \mathbf{V}, \mathbf{W}\}_t$ e de uma distribuição *a priori* assumida para os parâmetros de estado. A equação (4.3a) é denominada equação da observação e relaciona o vetor de observações \mathbf{Y}_t aos estados $\boldsymbol{\theta}_t$, já a equação (4.3b) é denominada equação do sistema ou equação de estados, responsável pela evolução dos parâmetros de estado ao longo do tempo.

Assume-se que as observações \mathbf{Y}_t são condicionalmente independentes dado \mathbf{x}_t , para $t = 1, 2, \dots$. Ademais, assume-se que, para todo $t \neq s$, há independência entre os erros observacionais \mathbf{v}_t e \mathbf{v}_s , entre os erros de evolução $\boldsymbol{\omega}_t$ e $\boldsymbol{\omega}_s$ e entre os erros \mathbf{v}_t e $\boldsymbol{\omega}_s$. Note, ainda, que nos modelos dinâmicos lineares normais, supõe-se que os erros de observação e de evolução do sistema são gaussianos, sendo a interferência dos componentes latentes no modelo dada de forma linear em ambas as equações. Uma descrição detalhada desses modelos pode ser vista em [West e Harrison \(1997\)](#).

4.2 Modelos Não-Gaussianos

Os modelos de espaço de estados lineares gaussianos são adequados em diversas aplicações de séries temporais. Entretanto, em muitas situações práticas, alternativas são necessárias para o tratamento de séries temporais que não satisfazem a suposição de normalidade, como é o caso de séries temporais de contagem, por exemplo, em que a variável resposta é um inteiro não-negativo representando o número de ocorrências de determinado fenômeno. Neste caso, pode-se considerar uma distribuição na família exponencial para as observações e uma distribuição normal multivariada para o vetor de estados. Esta classe de modelos foi descrita na Seção 4.1.

Nelder e Wedderburn (1972) propuseram uma família denominada de modelos lineares generalizados, permitindo com que a distribuição da variável resposta pertença à família exponencial de distribuições. Sua aplicação em modelos de espaço de estados foi introduzida por West et al. (1985).

Outros tipos de modelos não-gaussianos são necessários para lidar com a presença de *outliers* na série ou mudanças estruturais na equação de estados.

Os modelos latentes não-gaussianos são uma classe de modelos hierárquicos em que não se assume uma distribuição normal multivariada para o campo latente \mathbf{x} de dimensão n .

Muitas vezes, assume-se que as variáveis observadas $\{y_i : i \in I\}$ são condicionalmente independentes dado \mathbf{x} e $\boldsymbol{\theta}$. Neste caso, tem-se a seguinte estrutura hierárquica:

$$y_i | \mathbf{x}, \boldsymbol{\theta} \sim p(y_i | x_i, \boldsymbol{\theta}) \quad (4.4a)$$

$$\mathbf{x} | \boldsymbol{\theta} \sim p(\mathbf{x} | \boldsymbol{\theta}) \quad (4.4b)$$

$$\boldsymbol{\theta} \sim p(\boldsymbol{\theta}) \quad (4.4c)$$

em que $\mathbf{y} = (y_1, \dots, y_{n_d})'$ são os dados observados, $\mathbf{x} = (x_1, \dots, x_n)'$ os elementos do campo latente, $\boldsymbol{\theta}$ é o vetor m -dimensional de hiperparâmetros, a distribuição de $y_i | \mathbf{x}, \boldsymbol{\theta}$ pertence à família exponencial e a distribuição de $\mathbf{x} | \boldsymbol{\theta}$ é não-gaussiana.

A distribuição *a posteriori* é dada por:

$$p(\mathbf{x}, \boldsymbol{\theta} | \mathbf{y}) \propto p(\boldsymbol{\theta}) p(\mathbf{x} | \boldsymbol{\theta}) \prod_{i \in I} p(y_i | x_i, \boldsymbol{\theta}) \quad (4.5)$$

Alguns trabalhos abordando modelos de espaço de estados não-gaussianos podem ser citados, como o de [Kitagawa \(1987\)](#), que trata da modelagem de séries não estacionárias, em que a equação de observação e sistema não necessariamente são gaussianas, o de [Durbin e Koopman \(2000\)](#) que discutem técnicas baseadas no Filtro de Kalman e na simulação de Monte Carlo usando amostragem por importância para estimação em modelos de espaço de estados não-gaussianos, além do trabalho de [Smith e Miller \(1986\)](#) em que é desenvolvida uma classe de modelos de espaço de estados para dados censurados, assumindo que as observações pertencem à distribuição exponencial e evolução não-gaussiana. [Harvey e Fernandes \(1989\)](#) tratam de modelos de séries temporais para dados de contagem em que a dinâmica do nível é inicializada com uma especificação *a priori* conjugada.

4.3 Família Gama de Modelos Dinâmicos

Segundo [Santos \(2009\)](#), uma variável aleatória Y possui distribuição na família gama de distribuições se sua densidade ou função de probabilidade é escrita na forma:

$$p(y | \mu, \varphi) = a(y, \varphi) \mu^{b(y, \varphi)} \exp(-\mu c(y, \varphi)) \quad (4.6)$$

em que $y \in H(\varphi) \subset \mathfrak{R}$ e 0 , caso contrário. As funções $a(\cdot)$, $b(\cdot)$, $c(\cdot)$ e $H(\cdot)$ são funções conhecidas.

Observe que se $b(y, \varphi) = b(\varphi)$ ou $c(y, \varphi) = c(\varphi)$ e $H(\varphi)$ não depende de φ , a família gama se torna um caso especial da família exponencial de distribuições.

Define-se a família gama de modelos dinâmicos da seguinte forma:

1. Relaciona-se o preditor com o parâmetro μ_t através da relação $\mu_t = \lambda_t g(r_t, \boldsymbol{\beta})$, em que g é a função de ligação, r_t é um vetor de covariáveis e $\boldsymbol{\beta}$ os coeficientes de regressão. Tais coeficientes são um dos componentes de φ . O parâmetro responsável pela descrição do nível dinâmico é dado por λ_t .

2. A dinâmica do nível λ_t evolui através da equação $\lambda_t = \frac{\lambda_{t-1}\eta_t}{w}$, em que

$$\eta_t \sim \text{Beta}(wa_{t-1}, (1-w)a_{t-1}) \quad (4.7)$$

3. Inicializa-se a dinâmica do nível através de uma especificação inicial *a priori*, dada por $\lambda_0|D_0 \sim \text{Gama}(a_0, b_0)$, em que D_0 denota o conjunto de informação inicial.

[Santos \(2009\)](#) mostrou que ao se definir o modelo da forma descrita anteriormente, os seguintes resultados são válidos, para cada $t = 1, 2, \dots$

1. $\lambda_t|D_{t-1} \sim \text{Gama}(wa_{t-1}, wb_{t-1})$. Este resultado pode ser encontrado resolvendo a seguinte integral:

$$p(\lambda_t|D_{t-1}) = \int p(\lambda_t|\lambda_{t-1}, D_{t-1})p(\lambda_{t-1}|D_{t-1})d\lambda_{t-1}$$

2. $\mu_t|D_{t-1} \sim \text{Gama}(wa_{t-1}, wb_{t-1}[g(r_t, \boldsymbol{\beta})]^{-1})$, pela propriedade de escala da distribuição Gama.

3. $\lambda_t|D_t \sim \text{Gama}(a_t, b_t)$, sendo $a_t = wa_{t-1} + b(y_t, \varphi)$ e $b_t = wb_{t-1} + c(y_t, \varphi)g(r_t, \boldsymbol{\beta})$.

4. $\mu_t|D_t \sim \text{Gama}(a_t^*, b_t^*)$, sendo $a_t^* = wa_{t-1} + b(y_t, \varphi)$ e $b_t^* = wb_{t-1}[g(r_t, \boldsymbol{\beta})]^{-1} + c(y_t, \varphi)$

em que, para cada $t = 1, 2, \dots$, $D_t = (D_0, y_1, \dots, y_t)$ e a_t e b_t são obtidos recursivamente.

Observe que

$$\text{Var}(\lambda_t | D_{t-1}) = \frac{a_{t-1}}{wb_{t-1}^2} = w^{-1}\text{Var}(\lambda_{t-1} | D_{t-1}). \quad (4.8)$$

Assim, devido à passagem do tempo, apenas $100w\%$ da informação é preservada.

A utilização desses modelos ao invés de modelos lineares dinâmicos generalizados ([West e Harrison, 1997](#)), cuja distribuição das observações da série pertence à família exponencial de distribuições, tem uma vantagem que é a possibilidade de obter a equação de evolução de forma exata. Como pode ser visto em [Franco et al. \(2009\)](#), alguns casos particulares da família gama de distribuições incluem os modelos Poisson, Gama, Weibull, Pareto, Beta e Normal com média conhecida.

4.4 Dengue na Região Metropolitana do Estado do Rio de Janeiro

4.4.1 Introdução

Um dos grandes problemas de saúde pública no mundo, especialmente em países de clima tropical como o Brasil, é a dengue. De acordo com a Organização Mundial de Saúde, a incidência de dengue tem aumentado significativamente nas últimas décadas, sendo que cerca de dois bilhões e meio de pessoas estão em exposição ao risco de adquirir a doença.

Na literatura, muitos autores sugerem diferentes formas de modelagem para casos de dengue. Entre eles, pode-se citar [Fernandes \(2006\)](#) e [Ferreira \(2004\)](#), cuja abordagem é baseada no número de notificações dos casos. O que geralmente ocorre nessas situações é que o número de notificações é bastante inferior aos números de casos. A isto chama-se subnotificação.

É possível que óbitos por dengue tenham sido classificados como óbitos por outras doenças, já que existem muitas doenças cujos sintomas se confundem com os sintomas de uma pessoa infectada com dengue. Por este motivo, neste trabalho, optou-se por analisar o número de óbitos causados pela dengue e por doenças infecciosas que fazem parte de seu diagnóstico diferencial na região metropolitana do Rio de Janeiro, diferente do que é feito usualmente, em que se modela o número de casos notificados de dengue. O objetivo, é investigar a ocorrência de sobremortalidade durante períodos epidêmicos, isto é, verificar excessos de óbitos não esperados. Tal análise pode ser útil para autoridades sanitárias a fim de uma possível intervenção em futuros processos epidêmicos.

4.4.2 Descrição do Modelo

Como já foi dito, o interesse é verificar se, em períodos epidêmicos de dengue, há um excesso de mortalidade. A motivação de se analisar os óbitos ao invés dos casos notificados é a ocorrência da subnotificação. Segundo a Organização Mundial de Saúde, apenas 10% dos casos de dengue são realmente notificados. Considerando, ainda, a dificuldade de diagnóstico, pode-se levantar a hipótese da ocorrência de óbitos não registrados como dengue.

Sendo assim, a análise da sobremortalidade de dengue na região metropolitana do Rio de Janeiro é feita utilizando não apenas o número de óbitos causados pela dengue, mas também os óbitos causados por doenças infecciosas que fazem parte de seu diagnóstico diferencial.

Desta forma, seja Y_t uma variável aleatória que representa o número de óbitos no instante t , $t = 1, \dots, T$. Suponha que

$$Y_t | \mu_t \sim \text{Poisson}(\mu_t N_t), \quad (4.9)$$

em que N_t é assumido conhecido para todo $t = 1, \dots, T$ e se refere à população de expostos ao risco de adquirir dengue ou doenças infecciosas de seu diagnóstico diferencial no instante t , tal que μ_t é interpretado como a taxa em que os óbitos ocorrem no tempo t .

É possível relacionar um vetor de covariáveis r_t ao parâmetro μ_t através da relação $\mu_t = \lambda_t g(r_t, \beta)$, em que g é uma função de ligação e λ_t é responsável pela descrição do nível dinâmico. No caso em questão, será considerado $g(r_t, \beta) = \exp\{r_t' \beta\}$ e, portanto

$$\mu_t = \lambda_t \exp\{r_t' \beta\}. \quad (4.10)$$

A utilização de modelos dinâmicos e a inclusão de covariáveis que representem características temporais são uma tentativa de captar estruturas presentes nos dados que modelos lineares generalizados poderiam não captar, já que há uma dependência temporal em séries de natureza epidemiológica, como as de estudo.

Considera-se que

$$\mu_t = \lambda_t \exp\{r_{t,1}\beta_1 + r_{t,2}\beta_2 + r_{t,3}\beta_3\}, \quad (4.11)$$

em que $r_{t,1}$ é uma variável *dummy* informada pelos epidemiologistas que indica os períodos epidêmicos, $r_{t,2}$ e $r_{t,3}$ correspondem às componentes sazonais $\text{sen}(2\pi t/12)$ e $\text{cos}(2\pi t/12)$, para $t = 1, 2, \dots, T$, respectivamente.

As componentes sazonais foram incluídas pois a incidência de dengue é influenciada pelas condições climáticas, que variam dependendo da estação do ano. A inclusão da variável indicadora dos períodos de epidemia de dengue se deve ao fato do interesse de ser verificado se há ocorrência de sobremortalidade durante os períodos epidêmicos.

Sendo assim, através da componente β_1 correspondente à variável indicadora, pode-se mensurar o efeito da ocorrência de epidemia na série de óbitos em análise.

A dinâmica do nível λ_t pode ser dada segundo a seguinte equação de evolução:

$$\lambda_t = \frac{\lambda_{t-1}\eta_t}{w}, \quad (4.12)$$

em que

$$\eta_t \sim \text{Beta}(wa_{t-1}, (1-w)a_{t-1}), \quad (4.13)$$

sendo a_t um parâmetro da distribuição *a posteriori* de $\lambda_t|D_t$, em que D_t denota o conjunto de informação no tempo $t, t = 1, 2, \dots, T$.

Neste caso, verifica-se que $\log(\lambda_t) = \log(\lambda_{t-1}) + \eta_t^*$, em que $\eta_t^* = \log(\frac{\eta_t}{w})$. Este resultado é similar à equação de evolução descrita em (4.3).

Como já foi dito na Seção 4.3, w , que varia entre 0 e 1, desempenha um papel de aumentar a variância devido à passagem de tempo.

Para completar a especificação do modelo, considera-se a seguinte distribuição *a priori*:

$$p(\boldsymbol{\lambda}, \boldsymbol{\beta}, w) = p(\boldsymbol{\lambda}|\boldsymbol{\beta}, w)p(\boldsymbol{\beta})p(w), \quad (4.14)$$

em que se assume

$$\boldsymbol{\beta} \sim N(\mathbf{0}, cI), \quad c \text{ grande.} \quad (4.15a)$$

$$w \sim U(0, 1] \quad (4.15b)$$

$$\lambda_1|w, D_0 \sim \text{Gama}(wa_0, wb_0) \quad (4.15c)$$

Nesta aplicação, considera-se as seguintes especificações *a priori*: $c = 1000$, $a_0 = 0.01$ e $b_0 = 0.01$.

Nas subseções a seguir, observa-se que essa escolha leva a uma análise conjugada para os parâmetros de estado, condicionalmente aos hiperparâmetros $\varphi = (\boldsymbol{\beta}, w)$.

4.4.3 Procedimento de Inferência

Todo o procedimento é realizado sob abordagem bayesiana. Dificuldades surgem já que a distribuição *a posteriori* $p(\boldsymbol{\lambda}, \boldsymbol{\beta}, w | \mathbf{y})$ não pode ser obtido analiticamente. Aqui, são utilizados os métodos MCMC para construir amostras da distribuição *a posteriori*.

Inferência para os Parâmetros Estáticos

Considere $\varphi = (\boldsymbol{\beta}, w)$. A distribuição *a posteriori* de φ é

$$p(\boldsymbol{\beta}, w; y) = \frac{l(\boldsymbol{\beta}, w; y)p(\boldsymbol{\beta})p(w)}{\int l(\boldsymbol{\beta}, w; y)p(\boldsymbol{\beta})p(w)d\boldsymbol{\beta}dw} \quad (4.16)$$

A verossimilhança marginal é dada por $l(\boldsymbol{\beta}, w; y) = \prod_{t=1}^T p(y_t | \boldsymbol{\beta}, w, D_{t-1})$, em que a distribuição preditiva $p(y_t | \boldsymbol{\beta}, w, D_{t-1})$ pertence a uma distribuição binomial negativa, o que pode ser visto resolvendo a seguinte integral:

$$p(y_t | \boldsymbol{\beta}, w, D_{t-1}) = \int \underbrace{p(y_t | \lambda_t, \boldsymbol{\beta})}_{\text{Poisson}} \underbrace{p(\lambda_t | \boldsymbol{\beta}, w, D_{t-1})}_{\text{Gama}} d\lambda_t$$

Neste caso, a distribuição *a posteriori* não está disponível em forma fechada, portanto são utilizados os métodos MCMC para aproximar $p(\boldsymbol{\beta}, w; y)$.

Inferência para os Parâmetros de Estado

Para a inferência sobre os parâmetros de estado, considere os seguintes resultados, condicionalmente a φ , embora não esteja explícito na notação.

Resultado 1: A distribuição *a priori* $\lambda_t | D_{t-1}$ segue uma distribuição Gama($a_{t|t-1}, b_{t|t-1}$) em que $a_{t|t-1} = wa_{t-1}$ e $b_{t|t-1} = wb_{t-1}$.

Resultado 2: A distribuição *a posteriori* $\lambda_t | D_t$ segue uma distribuição a Gama(a_t^*, b_t^*), em que $a_t^* = a_{t|t-1} + y_t$ e $b_t^* = b_{t|t-1} + N_t e^{\mathbf{r}_t' \boldsymbol{\beta}}$.

Resultado 3: $\lambda_{t-1} - w\lambda_t | \lambda_t, D_{t-1} \sim \text{Gama}((1-w)a_{t-1}^*, b_{t-1}^*)$, o que segue diretamente do Resultado 1.

Estes resultados são especialmente importantes. Já que

$$p(\lambda_1, \dots, \lambda_T | D_T) = p(\lambda_T | D_T) \prod_{t=1}^{T-1} p(\lambda_t | \lambda_{t+1}, D_t) \quad (4.17)$$

e que $p(\lambda_t | \lambda_{t+1}, D_t)$ é conhecido, é possível construir amostras de $(\lambda_1, \dots, \lambda_T | D_T)$ baseado no seguinte algoritmo:

1. Faça $t = T$ and amostre de $p(\lambda_T | D_T)$;
2. Faça $t \leftarrow t - 1$ e amostre de $p(\lambda_t | \lambda_{t+1}, D_t)$
3. Se $t > 1$, retorne ao item 2. Caso contrário, a amostra de $(\lambda_1, \dots, \lambda_T | D_T)$ está completa.

Previsões

Para se fazer previsões, para $h = 1, 2, \dots$, pode-se obter a densidade $y_{T+h} | D_T$ através de

$$p(y_{T+h} | D_T) = \int p(y_{T+h} | \varphi, \boldsymbol{\lambda}, D_T) p(\varphi, \boldsymbol{\lambda} | D_T) d\varphi d\boldsymbol{\lambda} \quad (4.18)$$

Esta distribuição pode ser aproximada por

$$p(y_{T+h} | D_T) \approx \frac{1}{L} \sum_{i=1}^L p(y_{T+h} | \varphi^{(i)}, \boldsymbol{\lambda}^{(i)}, D_T), \quad (4.19)$$

em que L é referente ao número de simulações de algoritmo MCMC, sendo $\varphi^{(i)}$ e $\boldsymbol{\lambda}^{(i)}$ gerados pela distribuição *a posteriori* $p(\varphi, \boldsymbol{\lambda} | D_T)$.

4.4.4 Descrição dos Dados

Uma série de dados de óbitos de dengue e doenças infecciosas que fazem parte de seu diagnóstico diferencial na região metropolitana do Rio de Janeiro no período de 1986 a 2008 é analisada sob uma perspectiva bayesiana após a especificação da distribuição *a priori* para o vetor de parâmetros. O objetivo geral é investigar a ocorrência de sobremortalidade durante a epidemia de dengue ocorrida na região metropolitana do Rio de Janeiro em 2008.

A região metropolitana do Estado do Rio de Janeiro é composta por dezesseis municípios: Belford Roxo, Duque de Caxias, Guapimirim, Itaboraí, Japeri, Magé, Nilópolis, Niterói, Nova Iguaçu, Paracambi, Queimados, Rio de Janeiro, São Gonçalo, São João de Meriti, Seropédica e Tanguá. Entretanto, os municípios de Itaguaí, Mangaratiba, Maricá e Mesquita foram mantidos na presente análise, tendo em vista a sua forte relação com a dinâmica metropolitana.

As doenças consideradas no diagnóstico diferencial do dengue foram: hepatites virais, salmonelose, escarlatina, febre purpúrica do Brasil, pneumonia, influenza, enterovirose, septicemia, infecção meningocócica, leptospirose, riquetsioses, febres recorrentes, malária, arboviroses (febre amarela, Mayaro, Oropuche e outras), hantaviroses, parvoviroses, mononucleose infecciosa, exantema súbito, sarampo, rubéola, citomegalovirose, outras doenças exantemáticas, encefalomielite, outras febres hemorrágicas transmitidas por mosquitos e carrapatos. Para mais detalhes, consultar [Malhao \(2010\)](#).

Por simplificação de notação, quando no texto houver referência para o número de óbitos, entenda que se trata dos óbitos de dengue e pelas doenças infecciosas que fazem parte de seu diagnóstico diferencial citadas acima.

Os dados são oriundos de duas fontes de informação: Sistema de Informação de Agravos de Notificação (Sinan), por ser o sistema oficial brasileiro de notificação, e o Sistema de Informação sobre Mortalidade (SIM), sistema de caráter universal de âmbito nacional, em que devem constar todos os óbitos ocorridos no país. Os dados populacionais são oriundos do Instituto Brasileiro de Geografia e Estatística (IBGE), com base nos censos demográficos de 1991 e 2000. Estas informações foram gentilmente cedidas por Thainá Alves Malhão.

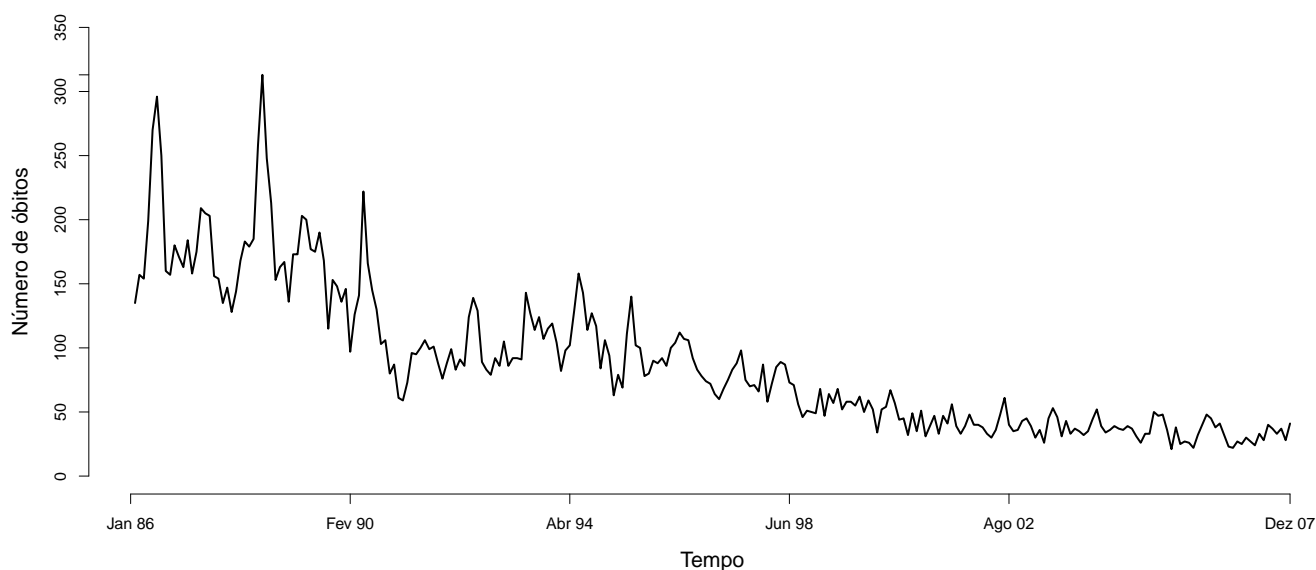
Devido à falta de consenso sobre o método mais adequado para identificar os meses epidêmicos, arbitrou-se como sendo o início e término da epidemia, quando houvesse um aumento ou decréscimo de no mínimo duas vezes o número de casos do mês de referência em relação ao mês anterior com evidência laboratorial de infecção. Estas informações foram gentilmente cedidas por Thainá Alves Malhão e Roberto de Andrade Medronho.

Os dados estão em frequência mensal de janeiro de 1986 a dezembro de 2008 e apenas são considerados grupos de indivíduos menores que 15 anos. As últimas 12 observações que correspondem ao ano de 2008 serão usadas para predição.

Com isso, é de interesse investigar a ocorrência de excesso de óbitos por dengue e doenças infecciosas que fazem parte do seu diagnóstico diferencial, na população residente com menos de 15 anos, durante a epidemia de dengue ocorrida na região metropolitana do Rio de Janeiro em 2008.

A Figura 4.1 mostra os número de óbitos para indivíduos menores que 15 anos de 1986 a 2007.

Figura 4.1: Número de óbitos para indivíduos menores que 15 anos de 1986 a 2007.



4.4.5 Resultados

Esta seção consta os resultados obtidos ao se analisar o número de óbitos por dengue e doenças infecciosas que fazem parte do seu diagnóstico diferencial, na população residente com menos de 15 anos.

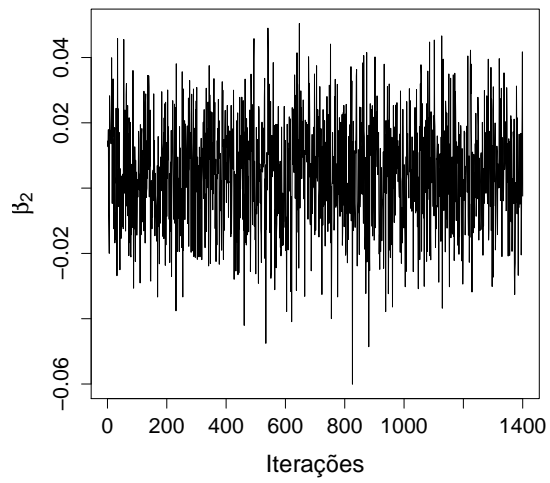
A especificação *a priori* consta na Seção 4.15.

Para os parâmetros estáticos, foi utilizado o algoritmo de Metropolis-Hastings com distribuição proposta normal passeio aleatório. Para os parâmetros de estado, a inferência foi baseada no esquema descrito no Resultado 3 da Seção 4.4.3.

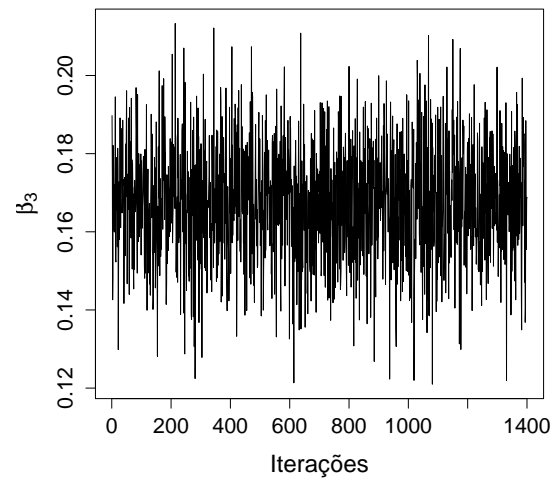
Foram simuladas 50 000 iterações, sendo as primeiras 8 000 descartadas, consideradas como período de aquecimento. Em seguida, foi considerado um espaçamento de 30 iterações. Com isso, a amostra final *a posteriori* possui tamanho 1 400.

A Figura 4.2 mostra os traços das cadeias de cada um dos parâmetros após o período de aquecimento.

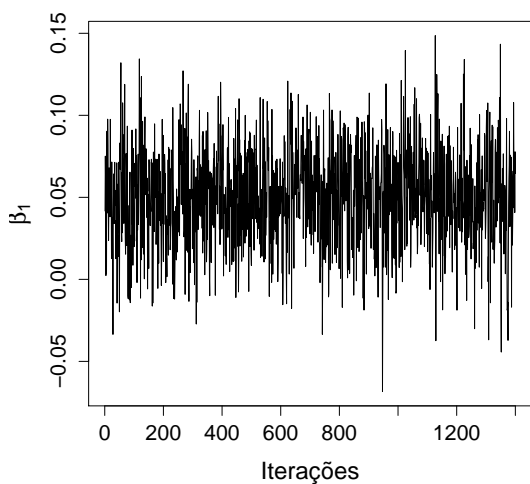
Figura 4.2: Traços das cadeias após período de aquecimento dos parâmetros fixos β_2 e β_3 , que são referentes às covariáveis que contém componentes sazonais, β_1 , que se refere à variável indicadora de epidemia de dengue e w .



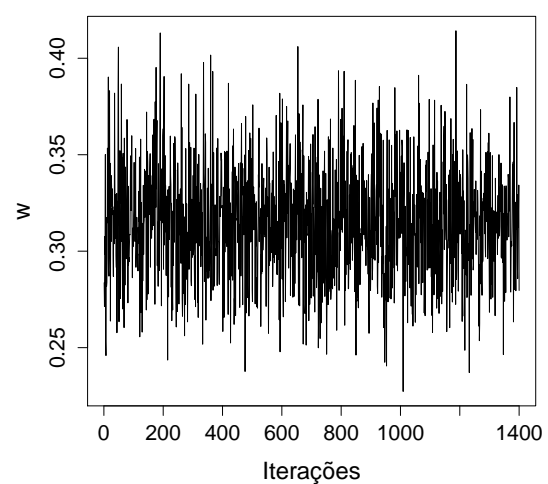
(a) β_2



(b) β_3



(c) β_1



(d) w

Na Tabela 4.1 consta o resumo *a posteriori* e as Figuras 4.3 e 4.4 o histograma dos parâmetros β_1 , β_2 , β_3 e w .

Observe que, nos períodos epidêmicos, estimou-se que há um aumento de apenas cerca de 5% no número de óbitos mensais por dengue e doenças infecciosas que fazem parte do seu diagnóstico diferencial em residentes da Região Metropolitana do Rio de Janeiro menores de 15 anos de idade, no ano de 2008.

Observa-se, entretanto, que os parâmetros β_1 , β_2 não se mostraram significativos ao nível de 5%. O parâmetro β_1 refere-se à variável indicadora de períodos epidêmicos. Assim sendo, verifica-se que o fato de um determinado período ser considerado epidêmico não influenciou na mortalidade do ano de 2008. Já o parâmetro β_2 é uma das componentes sazonais, o que indica que não foi necessário incluí-la no modelo.

A estimativa *a posteriori* de w é pequena, ou seja, devido à passagem de tempo, pouca informação está sendo passada, o que implica que λ varia significativamente de um instante para outro.

Tabela 4.1: Resumo *a posteriori* dos parâmetros estáticos.

Parâmetro	Média	Mediana	$q_{0.025}$	$q_{0.975}$	Desvio-Padrão
β_1	0.0504	0.0512	-0.0117	0.1075	0.0301
β_2	0.0041	0.0046	-0.0279	0.0359	0.0165
β_3	0.1674	0.1679	0.1382	0.1957	0.0152
w	0.3147	0.3138	0.2612	0.3767	0.0291

Figura 4.3: Histograma dos parâmetros estáticos *a posteriori* referentes às componentes sazonais para a série de dados de indivíduos menores que 15 anos de idade. A linha pontilhada representa o intervalo de credibilidade de 95%.

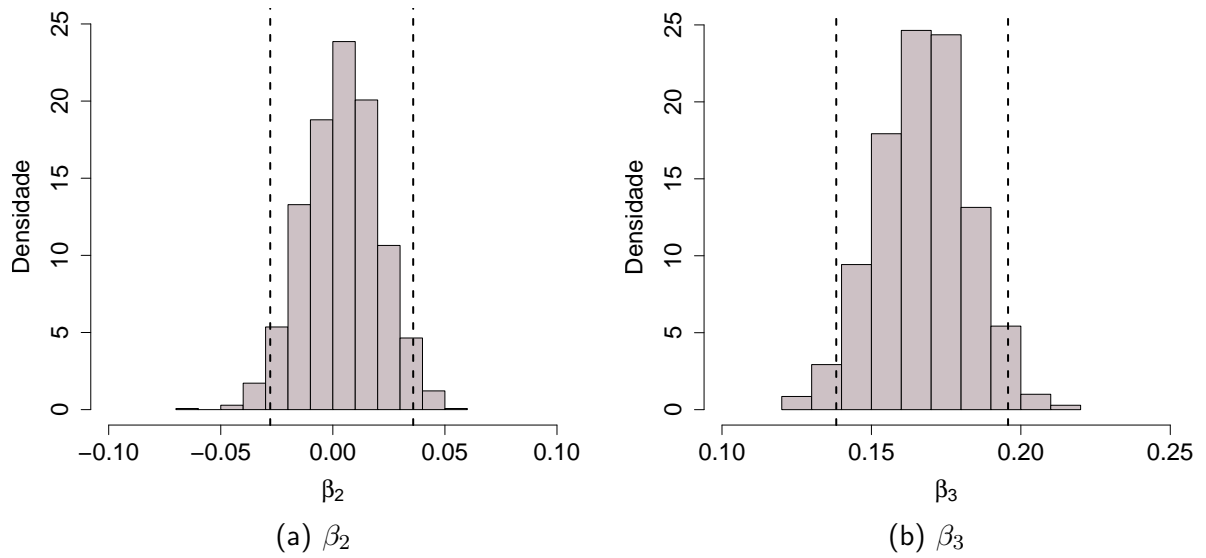
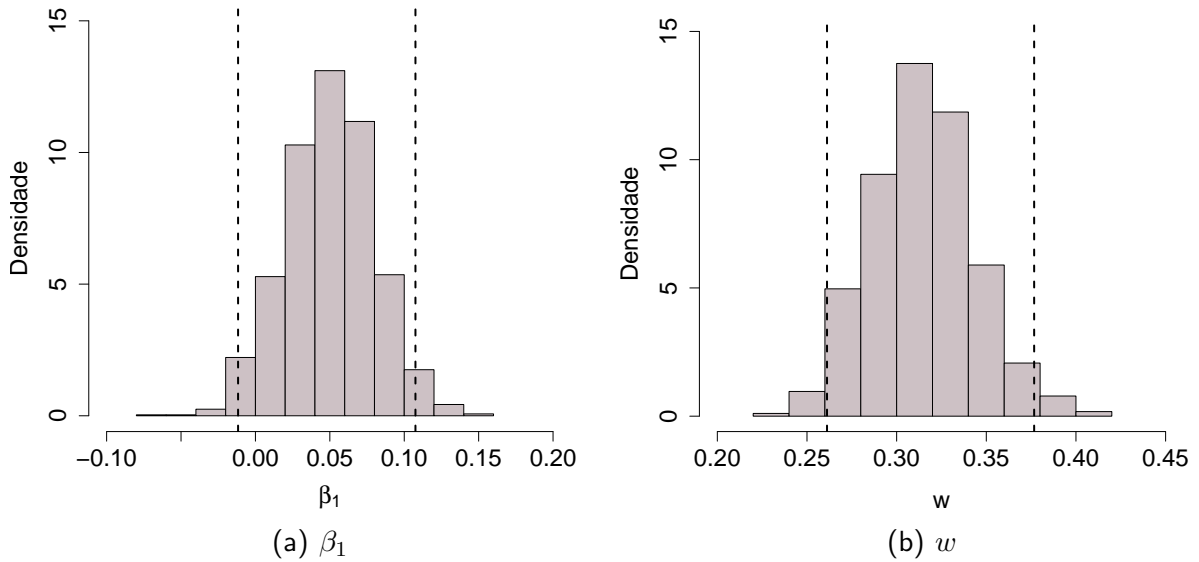
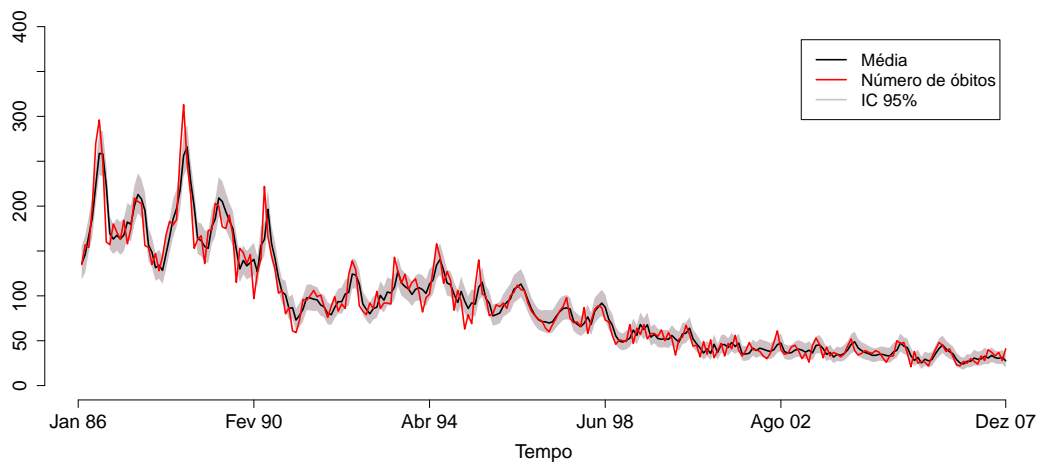


Figura 4.4: Histograma dos parâmetros estáticos *a posteriori* para a série de dados de indivíduos menores que 15 anos de idade. A linha pontilhada representa o intervalo de credibilidade de 95%.



Através da Figura 4.5, observa-se que o nível estimado $N_{t\mu_t}$ tem aproximadamente o mesmo comportamento do número de óbitos que realmente ocorreram.

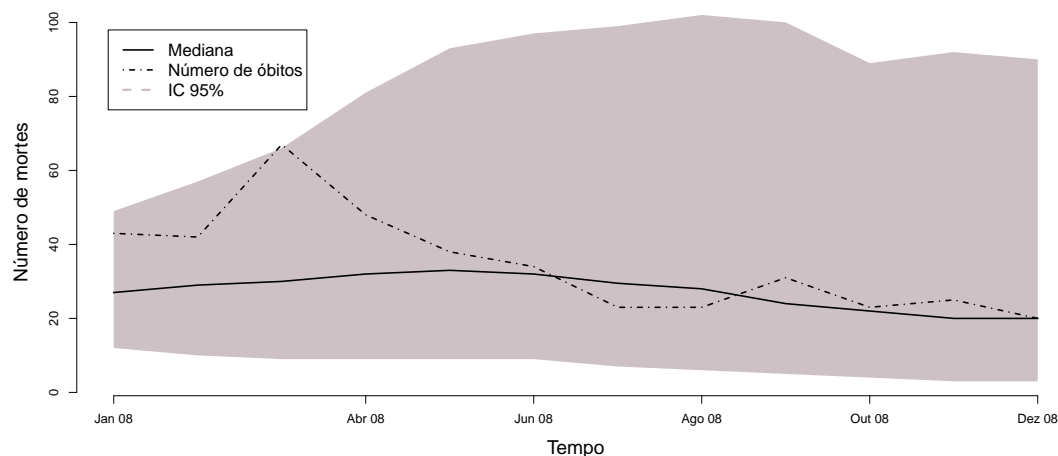
Figura 4.5: Média *a posteriori* de $N_{t\mu_t}$ e intervalo de credibilidade de 95%.



Através da Figura 4.6, observa-se que de janeiro a maio de 2008, o número de óbitos que realmente ocorreram está acima do esperado pelo modelo, apesar de não ter ultrapassado os limites do intervalo de credibilidade de 95%. De fato, neste período ocorreu uma epidemia de dengue no Rio de Janeiro que, diferentemente de outras epidemias, afetou bastante as crianças. A partir de julho de 2008, observa-se que o número de óbitos que realmente ocorreram seguem a mesma aproximadamente tendência do esperado pelo modelo.

Cabe observar que a amplitude do intervalo de credibilidade da previsão do número de óbitos para o ano de 2008 é alta. Devido a estimativa de w ser pequena, considera-se que, de um instante para outro, pouca informação é passada, levando a um aumento significativo na incerteza, sendo assim refletido na amplitude do intervalo de credibilidade.

Figura 4.6: Previsão do número de óbitos para indivíduos menores que 15 anos no ano de 2008.



De forma geral, entende-se que o modelo considerado não captou nenhuma sobremortalidade no ano de 2008. Mais ainda, conclui-se que, o modelo não captou aumentos da mortalidade em períodos epidêmicos de dengue de 1986 a 2007.

Capítulo 5

O Método de Aproximação INLA

O método de aproximação INLA (do inglês, *integrated nested Laplace approximation*) foi proposto por [Rue et al. \(2009\)](#) como forma de realizar inferência bayesiana de forma aproximada em uma subclasse de modelos de regressão estruturados aditivos chamados modelos latentes gaussianos. Os modelos latentes gaussianos foram discutidos na Seção (4.1).

Este capítulo explica detalhadamente toda a metodologia do INLA, além de apresentar ilustração do método. Em particular, a Seção 5.1 apresenta uma definição do método, explicitando qual a estrutura hierárquica que um modelo deve ter para poder ser aplicado o INLA, a Seção 5.2 explicita quais as distribuições são de interesse ao se aplicar o método INLA, e as demais seções explicam detalhadamente a metodologia que deve ser empregada.

5.1 Definição

O método de aproximação INLA ([Rue et al., 2009](#)) realiza inferência bayesiana de forma aproximada de forma rápida nos modelos latentes gaussianos (ver Seção 4.1). Este método foi proposto como uma alternativa aos métodos de Monte Carlo via cadeias de Markov, que têm sua forma de inferência baseada em simulações.

O método INLA, diferentemente, aproxima de forma direta as distribuições *a posteriori* de interesse. Com isso, é possível realizar inferência com um custo computacional consideravelmente inferior além de não ser preciso lidar com problemas de convergência.

Além da dificuldade no diagnóstico da convergência e o alto custo computacional, em se tratando de inferência em modelos latentes gaussianos, os métodos MCMC apresentam desvantagem do ponto de vista computacional devido à dependência dos elementos do campo latente entre si, além da dependência existente entre os hiperparâmetros e os parâmetros do campo latente. Com isso, o método INLA surge como uma alternativa para lidar com esses problemas.

Assume-se a seguinte estrutura hierárquica

$$y_i|\mathbf{x}, \boldsymbol{\theta} \sim p(y_i|x_i, \boldsymbol{\theta}) \quad (5.1a)$$

$$\mathbf{x}|\boldsymbol{\theta} \sim p(\mathbf{x}|\boldsymbol{\theta}) \quad (5.1b)$$

$$\boldsymbol{\theta} \sim p(\boldsymbol{\theta}), \quad (5.1c)$$

em que $\mathbf{y} = (y_1, \dots, y_{n_d})'$ são os dados observados, $\mathbf{x} = (x_1, \dots, x_n)'$ os elementos do campo latente, $\boldsymbol{\theta}$ é o vetor m -dimensional de hiperparâmetros, a distribuição de $y_i|\mathbf{x}, \boldsymbol{\theta}$ pertence à família exponencial e a distribuição de $\mathbf{x}|\boldsymbol{\theta}$ é gaussiana.

Assume-se que os dados observados são condicionalmente independentes dado o campo latente \mathbf{x} . Com isso, tem-se que:

$$p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}) = \prod_{i \in I} p(y_i|x_i, \boldsymbol{\theta}). \quad (5.2)$$

A distribuição *a posteriori* é dada por:

$$p(\mathbf{x}, \boldsymbol{\theta}|\mathbf{y}) \propto p(\boldsymbol{\theta})p(\mathbf{x}|\boldsymbol{\theta}) \prod_{i \in I} p(y_i|x_i, \boldsymbol{\theta}) \quad (5.3)$$

Rue et al. (2009) assumem que o campo latente \mathbf{x} possui grande dimensão, mas que admite propriedades de independência condicional. Considera-se modelos que satisfazem a propriedade de Markov, ou seja, $x_i|\mathbf{x}_{-i}$ só depende de um subconjunto de \mathbf{x}_{-i} , em que $\mathbf{x}_{-i} = (x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n)$. Neste caso, o campo latente é um campo aleatório markoviano gaussiano e a matriz de precisão referente ao campo latente é esparsa.

Além disso, assume-se que o número m de hiperparâmetros é pequeno, por exemplo, $m \leq 6$. Assim, é possível fazer uso de métodos numéricos para matrizes esparsas muito mais rapidamente do que para cálculos com matrizes densas (Rue e Held, 2005).

5.2 Distribuições de Interesse

As distribuições de interesse são as seguintes distribuições *a posteriori* marginais:

$$p(x_i|\mathbf{y}) = \int p(\boldsymbol{\theta}|\mathbf{y})p(x_i|\boldsymbol{\theta}, \mathbf{y})d\boldsymbol{\theta} \quad (5.4)$$

$$p(\theta_j|\mathbf{y}) = \int p(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta}_{-j}, \quad (5.5)$$

$\forall i = 1, \dots, n$ e $\forall j = 1, \dots, m$.

5.3 Parametrização Adequada do Vetor Paramétrico

Em geral, serão utilizados métodos numéricos para a obtenção das distribuições de interesse. Assim sendo, de acordo com Rue et al. (2009), é recomendado que seja feita uma reparametrização de $\boldsymbol{\Psi} = (\boldsymbol{\theta}, \mathbf{x})$ de modo que $\boldsymbol{\Psi}^* \in \mathbb{R}^k$, sendo k a dimensão do espaço paramétrico. Esta transformação facilitará a construção de grades otimizadas, em que as regiões com maior massa de probabilidade recebem mais pontos em relação às demais áreas. Do ponto de vista computacional, esta transformação permite que sejam utilizados algoritmos de otimização sem restrição, descritos na Seção 2.2.

Primeiramente, sugere-se a reparametrização dos parâmetros individuais, de tal forma que eles possuam suporte na reta real. Em seguida, deve ser encontrada a moda da distribuição *a posteriori* de $\boldsymbol{\Psi}^*$, cuja notação é $\boldsymbol{\Psi}_{Mo}^*$, além da matriz hessiana \mathbf{H} e avaliá-la nesta moda. Assim, define-se o vetor paramétrico reparametrizado da seguinte forma:

$$\boldsymbol{\Psi}^*(\mathbf{z}) = \boldsymbol{\Psi}_{Mo}^* + \mathbf{V}\boldsymbol{\Lambda}^{1/2}\mathbf{z}, \quad (5.6)$$

em que $\boldsymbol{\Sigma} = \mathbf{H}^{-1}$ e $\boldsymbol{\Sigma} = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}'$ a decomposição espectral de $\boldsymbol{\Sigma}$.

5.4 Esquema de Aproximação

Quando as integrações em (5.4) e (5.5) não podem ser resolvidas analiticamente, constrói-se aproximações para $p(x_i|\boldsymbol{\theta}, \mathbf{y})$ e $p(\boldsymbol{\theta}|\mathbf{y})$ dadas respectivamente por $\tilde{p}(x_i|\boldsymbol{\theta}, \mathbf{y})$ e $\tilde{p}(\boldsymbol{\theta}|\mathbf{y})$. Assim, é possível construir aproximações “aninhadas”:

$$\tilde{p}(x_i|\mathbf{y}) = \int \tilde{p}(\boldsymbol{\theta}|\mathbf{y})\tilde{p}(x_i|\boldsymbol{\theta}, \mathbf{y})d\boldsymbol{\theta} \quad (5.7)$$

$$\tilde{p}(\boldsymbol{\theta}_j|\mathbf{y}) = \int \tilde{p}(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta}_{-i}, \quad (5.8)$$

sendo as integrais resolvidas via integração numérica, ou seja,

$$\tilde{p}(x_i|\mathbf{y}) \approx \sum_j \tilde{p}(x_i|\boldsymbol{\theta}_j, \mathbf{y})\tilde{p}(\boldsymbol{\theta}_j|\mathbf{y})\Delta_j \quad (5.9)$$

$$\tilde{p}(\boldsymbol{\theta}_j|\mathbf{y}) \approx \sum_i p(\boldsymbol{\theta}_i|\mathbf{y})\Delta_{ji} \quad (5.10)$$

As demais seções descrevem como obter as aproximações necessárias.

5.5 Aproximação para $p(\boldsymbol{\theta}|\mathbf{y})$

A aproximação para $p(\boldsymbol{\theta}|\mathbf{y})$ é equivalente à aproximação de Laplace para a distribuição *a posteriori* marginal proposta por Tierney e Kadane (1986). Tal aproximação é dada por:

$$\tilde{p}(\boldsymbol{\theta}|\mathbf{y}) \propto \frac{p(\mathbf{x}, \boldsymbol{\theta}, \mathbf{y})}{\tilde{p}_G(\mathbf{x}|\boldsymbol{\theta}, \mathbf{y})} \Big|_{\mathbf{x}=\mathbf{x}^*(\boldsymbol{\theta})}, \quad (5.11)$$

em que $\tilde{p}_G(\mathbf{x}|\boldsymbol{\theta}, \mathbf{y})$ é a aproximação gaussiana da distribuição condicional completa de \mathbf{x} e $\mathbf{x}^*(\boldsymbol{\theta})$ é a moda da distribuição condicional completa de \mathbf{x} para um dado $\boldsymbol{\theta}$. Esta moda pode ser obtida através de algum método de otimização, como aqueles descritos na Seção 2.2.

A proporcionalidade vista em (5.11) é devido ao fato de não se conhecer a constante normalizadora de $p(\mathbf{x}, \boldsymbol{\theta}|\mathbf{y})$.

Tem-se que:

$$p(\mathbf{x}|\boldsymbol{\theta}, \mathbf{y}) \propto \exp \left\{ -\frac{1}{2} \mathbf{x}' \mathbf{Q}(\boldsymbol{\theta}) \mathbf{x} + \sum_i \log(p(y_i|x_i, \boldsymbol{\theta})) \right\}. \quad (5.12)$$

Obtendo a expansão de Taylor até a segunda ordem em torno de \mathbf{x}^* , encontra-se

$$\tilde{p}(\mathbf{x}|\boldsymbol{\theta}, \mathbf{y}) \propto \exp \left\{ -\frac{1}{2} \mathbf{x}' (\mathbf{Q}(\boldsymbol{\theta}) + \text{diag}(\mathbf{c})) \mathbf{x} + \mathbf{b}' \mathbf{x} \right\}, \quad (5.13)$$

sendo \mathbf{b} e \mathbf{c} os coeficientes da expansão.

O fato desta aproximação ser equivalente à aproximação de Laplace de [Tierney e Kadane \(1986\)](#) sugere o erro desta aproximação é da ordem de $O(n_d^{-3/2})$, após renormalização. Cabe ressaltar, entretanto, que a dimensão paramétrica não é fixa, já que depende de n_d , o que faz com que as suposições assintóticas utilizadas na aproximação de Laplace não sejam válidas.

[Rue e Martino \(2007\)](#) utilizaram a aproximação (5.11) em diversos modelos latentes gaussianos e concluíram que a aproximação dada em (5.11) foi bem precisa no sentido de que nem longas cadeias de MCMC foram capazes de detectar erros na aproximação.

5.5.1 Exploração da Grade de Pontos

Não é necessário representar $\tilde{p}(\boldsymbol{\theta}|\mathbf{y})$ parametricamente, mas sim explorá-la suficientemente bem a fim de selecionar bons pontos para a integração numérica.

Tal exploração de $p(\boldsymbol{\theta}|\mathbf{y})$, conforme sugerido em [Rue et al. \(2009\)](#), pode ser feito da seguinte forma:

1. Começar pela moda ($\mathbf{z} = \mathbf{0}$ em (5.6)) e ir na direção positiva de z_1 com espaçamento δ_z , por exemplo, $\delta_z = 1$;
2. Parar quando a condição $\log\{\tilde{p}(\mathbf{0}|\mathbf{y})\} - \log\{\tilde{p}(\boldsymbol{\theta}(\mathbf{z})|\mathbf{y})\} < \delta_\pi$ for satisfeita, sendo $\delta_\pi = 2.5$, por exemplo.
3. Repetir o procedimento na direção negativa de z_1 . As demais coordenadas seguem procedimento análogo, resultando nos chamados pontos pretos.

4. Completar a construção da grade com todas as combinações dos pontos pretos, verificando se a condição é satisfeita. Os novos pontos são chamados de pontos cinzas.

Tanto os pontos pretos como os pontos cinzas são utilizados para obter as distribuições marginais de $\theta|y$ via integração numérica.

5.6 Aproximação para $p(x_i|\theta, \mathbf{y})$

Rue et al. (2009) propõem três tipos de aproximação para $p(x_i|\theta, \mathbf{y})$, que variam em termos de custo computacional e/ou precisão. As subseções a seguir descrevem tais métodos.

5.6.1 Aproximação Gaussiana

A forma mais simples de obter uma aproximação para $p(x_i|\theta, \mathbf{y})$ é utilizar a distribuição marginal de $\tilde{p}_G(\mathbf{x}|\theta, \mathbf{y})$, ou seja, considerar que:

$$\tilde{p}_G(x_i|\theta, \mathbf{y}) \equiv N(\mu_i(\theta), \sigma_i^2(\theta)), \quad (5.14)$$

em que $\mu_i(\theta)$ e $\sigma_i^2(\theta)$ correspondem a média e variância marginais obtidas de $\tilde{p}_G(\mathbf{x}|\theta, \mathbf{y})$.

Apesar do fato de que esta aproximação pode trazer resultados satisfatórios em alguns casos, tal aproximação pode também trazer erros na locação ou erros relacionados à falta de simetria da aproximação (Rue e Martino, 2007).

5.6.2 Método de Laplace

A aproximação para $p(x_i|\theta, \mathbf{y})$ é dada por:

$$\tilde{p}_{LA}(x_i|\theta, \mathbf{y}) \propto \frac{p(\mathbf{x}, \theta, \mathbf{y})}{\tilde{p}_{GG}(\mathbf{x}_{-i}|x_i, \theta, \mathbf{y})} \Big|_{\mathbf{x}_{-i}=\mathbf{x}_{-i}^*(x_i, \theta)}, \quad (5.15)$$

em que $\tilde{p}_{GG}(\mathbf{x}_{-i}|x_i, \theta, \mathbf{y})$ é a aproximação gaussiana para $\mathbf{x}_{-i}|x_i, \theta, \mathbf{y}$ e $\mathbf{x}_{-i}^*(x_i, \theta)$ é a moda de $p(\mathbf{x}_{-i}|x_i, \theta, \mathbf{y})$.

Note que é necessário avaliar a expressão em (5.15) em diversos valores de x_i para obter a densidade, como por exemplo, $\{x_i^{(1)}, \dots, x_i^{(k)}\}$. Para selecionar estes pontos, considera-se $x_i^{(j)} = \mu_i(\boldsymbol{\theta}) + \sigma_i(\boldsymbol{\theta})x_{ab}^{(j)}$, em que $\mu_i(\boldsymbol{\theta})$ e $\sigma_i(\boldsymbol{\theta})$ são a média e desvio-padrão da aproximação gaussiana descrita em (5.11) e $x_{ab}^{(j)}$ é um ponto da abscissa dado pela quadratura de Gauss-Hermite (ver Seção 2.4.1).

Assim, a densidade $\tilde{p}_{LA}(x_i|\boldsymbol{\theta}, \mathbf{y})$ é representada por:

$$\tilde{p}_{LA}(x_i|\boldsymbol{\theta}, \mathbf{y}) \propto N\{\mu_i(\boldsymbol{\theta}), \sigma_i^2(\boldsymbol{\theta})\} \exp\{h(x_i)\}, \quad (5.16)$$

sendo a função $h(x)$ uma função *spline* cúbica ajustada à diferença da log-densidade de $\tilde{p}_{LA}(x_i|\boldsymbol{\theta}, \mathbf{y})$ e $\tilde{p}_G(x_i|\boldsymbol{\theta}, \mathbf{y})$, avaliada nos pontos selecionados. Em seguida, a densidade é normalizada utilizando integração por quadratura.

Um problema que surge é o fato de ser necessário avaliar \tilde{p}_{GG} para cada valor diferente de x_i e $\boldsymbol{\theta}$, o que leva a um alto custo computacional.

Uma alternativa é evitar o passo de otimização e computar $\tilde{p}_{GG}(\mathbf{x}_{-i}|x_i, \boldsymbol{\theta}, \mathbf{y})$ através de uma aproximação para a moda, dada por

$$\mathbf{x}_{-i}^*(x_i, \boldsymbol{\theta}) \approx E_{\tilde{p}_G}(\mathbf{x}_{-i}|x_i). \quad (5.17)$$

A outra modificação proposta por Rue et al. (2009) baseia-se no fato de que é razoável assumir que apenas os elementos do campo latente x_j que são de uma certa forma “próximos” de x_i devem, de fato, ter alguma influência da distribuição marginal de x_i .

Denote a região de interesse com relação à distribuição marginal de x_i por $R_i(\boldsymbol{\theta})$.

De fato, tem-se que:

$$\frac{E_{\tilde{p}_G}(x_j|x_i) - \mu_j(\boldsymbol{\theta})}{\sigma_j(\boldsymbol{\theta})} = a_{ij}(\boldsymbol{\theta}) \frac{x_i - \mu_i(\boldsymbol{\theta})}{\sigma_i(\boldsymbol{\theta})} \quad (5.18)$$

para algum $a_{ij}(\boldsymbol{\theta})$, $j \neq i$.

Pode-se considerar uma simples regra para construir $R_i(\boldsymbol{\theta})$ tal como:

$$R_i(\boldsymbol{\theta}) = \{j : |a_{ij}(\boldsymbol{\theta})| > 0,001\}. \quad (5.19)$$

Com isso, usando a região de interesse $R_i(\boldsymbol{\theta})$, apenas será necessário fatorizar matrizes esparsas de dimensão $|R_i(\boldsymbol{\theta})| \times |R_i(\boldsymbol{\theta})|$.

5.6.3 Método de Laplace Simplificado

O método de Laplace simplificado considera uma expansão de Taylor no numerador e no denominador da expressão (5.15) em torno de $x_i = \mu_i(\boldsymbol{\theta})$. Isso permitirá corrigir a aproximação gaussiana $\tilde{p}_G(x_i|\boldsymbol{\theta}, \mathbf{y})$ em locação e assimetria.

A expansão no numerador é feita até o termo de terceira ordem. Seja

$$x_i^s = (x_i - \mu_i(\boldsymbol{\theta}))/\sigma_i(\boldsymbol{\theta}). \quad (5.20)$$

Defina

$$d_j^{(3)}(x_i, \boldsymbol{\theta}) = \left. \frac{\partial^3}{\partial x_j^3} \log\{p(y_j|x_j, \boldsymbol{\theta})\} \right|_{x_j = E_{\tilde{p}_G}(x_j|x_i)}, \quad (5.21)$$

em que se assume que existe.

Expandindo o numerador da expressão (5.15) em torno de $x_i = \mu_i(\boldsymbol{\theta})$ até a terceira ordem, obtém-se

$$\begin{aligned} \log(p(\mathbf{x}, \boldsymbol{\theta}, \mathbf{y})|_{\mathbf{x}_{-i} = E_{\tilde{p}_G}(\mathbf{x}_{-1}|x_i)}) &= -\frac{1}{2}(x_i^{(s)})^2 + \frac{1}{6}(x_i^{(s)})^3 \\ &\times \sum_{j \in I \setminus i} d_j^{(3)}\{\mu_i(\boldsymbol{\theta}), \boldsymbol{\theta}\} \{\sigma_j(\boldsymbol{\theta}) a_{ij}(\boldsymbol{\theta})\}^3 + \dots \end{aligned} \quad (5.22)$$

Os termos de primeira e segunda ordem dão a aproximação gaussiana enquanto o termo de terceira ordem corrige a assimetria.

O denominador da expressão (5.15) se reduz a

$$\log\{\tilde{p}_{GG}\{\mathbf{x}_{-i}|x_i, \boldsymbol{\theta}, \mathbf{y}\}|_{\mathbf{x}_{-i} = E_{\tilde{p}_G}(\mathbf{x}_{-1}|x_i)} = cte + \frac{1}{2} \log |\mathbf{H} + \text{diag}\{\mathbf{c}(x_i, \boldsymbol{\theta})\}|, \quad (5.23)$$

em que \mathbf{H} é a matriz de precisão \mathbf{Q} em que a i -ésima linha e coluna são deletadas e $\mathbf{c}(x_i, \boldsymbol{\theta})$ é o vetor de menos a segunda derivada do termo da verossimilhança avaliada em $x_j = E_{\tilde{p}_G}(x_j|x_i)$.

Assim, obtém-se que

$$\begin{aligned} \log\{\tilde{p}_{GG}\{\mathbf{x}_{-i}|x_i, \boldsymbol{\theta}, \mathbf{y}\}|_{\mathbf{x}_{-i}=E_{\tilde{p}_G}(\mathbf{x}_{-1}|x_i)} &= cte - x_i^{(s)} \frac{1}{2} \sum_{j \in I \setminus i} var_{\tilde{p}_G}(x_j|x_i) \quad (5.24) \\ &\times d_j^{(3)}\{\mu_i(\boldsymbol{\theta}), \boldsymbol{\theta}\} \sigma_j(\boldsymbol{\theta}) a_{ij}(\boldsymbol{\theta}) + \dots \end{aligned}$$

Note que $var_{\tilde{p}_G}(x_j|x_i) = \sigma_j^2(\boldsymbol{\theta})\{1 - corr_{\tilde{p}_G}(x_i, x_j)^2\}$, sendo $corr_{\tilde{p}_G}(x_i, x_j)$ a correlação entre x_i e x_j sob \tilde{p}_G .

Definindo

$$\gamma_i^{(1)}(\boldsymbol{\theta}) = \frac{1}{2} \sum_{j \in I \setminus i} \sigma_j^2(\boldsymbol{\theta}) \{1 - corr_{\tilde{p}_G}(x_i, x_j)^2\} d_j^{(3)}\{\mu_i(\boldsymbol{\theta}), \boldsymbol{\theta}\} \sigma_j(\boldsymbol{\theta}) a_{ij}(\boldsymbol{\theta}) \quad (5.25)$$

e

$$\gamma_i^{(3)}(\boldsymbol{\theta}) = \sum_{j \in I \setminus i} d_j^{(3)}\{\mu_i(\boldsymbol{\theta}), \boldsymbol{\theta}\} \{\sigma_j(\boldsymbol{\theta}) a_{ij}(\boldsymbol{\theta})\}^3 \quad (5.26)$$

então

$$\log\{\tilde{p}_{SLA}(x_i^s|\boldsymbol{\theta}, \mathbf{y})\} = cte - \frac{1}{2}(x_i^{(s)})^2 + \gamma_i^{(1)}(\boldsymbol{\theta})x_i^{(s)} + \frac{1}{6}(x_i^{(s)})^3 \gamma_i^{(3)}(\boldsymbol{\theta}) + \dots \quad (5.27)$$

A equação em (5.27) não define uma densidade já que o termo de terceira ordem é ilimitado. Uma solução é ajustar uma densidade normal assimétrica (Azzalini e Capitanio, 1999) de modo que a terceira derivada avaliada na moda seja $\gamma_i^{(3)}$, a média seja $\gamma_i^{(1)}$ e a variância seja igual a 1. Dessa forma, $\gamma_i^{(3)}$ corrigirá a aproximação gaussiana na simetria e $\gamma_i^{(1)}$ ajustará a locação.

5.7 Algoritmo INLA

Pode-se resumir o método de aproximação INLA nos seguintes passos:

1. Explorar $\tilde{p}(\boldsymbol{\theta}|\mathbf{y})$.
 - Localizar a moda $\boldsymbol{\theta}^*$ de $\tilde{p}(\boldsymbol{\theta}|\mathbf{y})$, otimizando $\log\{\tilde{p}(\boldsymbol{\theta}|\mathbf{y})\}$ com respeito à $\boldsymbol{\theta}$.
 - Construir a transformação adequada para $\boldsymbol{\theta}$ tal como foi descrito em na Seção 5.3.
 - Explorar $\log\{\tilde{p}(\boldsymbol{\theta}|\mathbf{y})\}$ utilizando a nova parametrização.
 - Aproximar $p(\boldsymbol{\theta}_j|\mathbf{y})$ através de integração numérica, ou seja, $\sum_i p(\boldsymbol{\theta}_i|\mathbf{y})\Delta_{ji}$. Uma outra possibilidade é aproveitar os pontos calculados no item anterior e construir um interpolante para $\log\{p(\boldsymbol{\theta}|\mathbf{y})\}$ e fazer a integração numérica através deste interpolante.
2. Obter $\tilde{p}(x_i|\boldsymbol{\theta}, \mathbf{y})$. Para cada $\boldsymbol{\theta}_j$ da grade, encontrar a aproximação de Laplace para cada elemento do campo latente.
3. Obter $\tilde{p}(x_i|\mathbf{y})$ por meio de integração numérica, ou seja $\sum_j \tilde{p}(x_i|\boldsymbol{\theta}_j, \mathbf{y})\tilde{p}(\boldsymbol{\theta}_j|\mathbf{y})\Delta_j$.

5.8 Exemplo: Modelo Poisson Dinâmico

Suponha variáveis aleatórias Y_1, \dots, Y_T condicionalmente independentes seguindo uma distribuição Poisson com média μ_t . Assume-se que a média μ_t varia no tempo através de uma estrutura dinâmica, ou seja,

$$Y_t | \mu_t \sim Poi(\mu_t) \quad (5.28a)$$

$$\log(\mu_t) = \lambda_t \quad (5.28b)$$

$$\lambda_t = \lambda_{t-1} + \omega_t, \omega_t \sim N(0, W) \quad (5.28c)$$

O modelo é completado após uma especificação *a priori* a respeito do processo no instante inicial. Assume-se que

$$\lambda_0|D_0 \sim N(m_0, C_0), \quad (5.29)$$

sendo m_0 e C_0 valores conhecidos refletindo a incerteza *a priori* sobre o processo no instante inicial.

Assume-se que o valor de W seja desconhecido. Neste caso, atribui-se uma especificação *a priori* para W , a qual considera-se

$$W^{-1} \sim Gama(\nu_0, S_0), \quad (5.30)$$

sendo ν_0 e S_0 valores conhecidos.

Assim, as quantidades a serem estimadas são $\Psi = (W, \lambda_1, \dots, \lambda_T)$, em que $\theta = W$ é o parâmetro fixo e $(\lambda_1, \dots, \lambda_T)$ parâmetros do campo latente.

A distribuição conjunta dos parâmetros do campo latente (condicional a W , embora omitido na notação) é dada por:

$$p(\lambda_1, \dots, \lambda_T) = p(\lambda_1) \prod_{t=2}^T p(\lambda_t | \lambda_{t-1}) \quad (5.31a)$$

$$\begin{aligned} &\propto \exp \left\{ -\frac{1}{2(C_0 + W)} (\lambda_1 - m_0)^2 \right\} \times \quad (5.31b) \\ &\times \prod_{t=2}^T \exp \left\{ -\frac{W^{-1}}{2} (\lambda_t - \lambda_{t-1})^2 \right\}, \end{aligned}$$

devido à estrutura de independência condicional.

Assumindo $m_0 = 0$,

$$p(\lambda_1, \dots, \lambda_T) \propto \exp \left\{ -\frac{(C_0 + W)^{-1}}{2} \lambda_1^2 \right\} \prod_{t=2}^T \exp \left\{ -\frac{W^{-1}}{2} (\lambda_t - \lambda_{t-1})^2 \right\} \quad (5.32a)$$

$$\begin{aligned} &= \exp \left\{ -\frac{(C_0 + W)^{-1}}{2} \lambda_1^2 \right\} \times \quad (5.32b) \\ &\times \exp \left\{ \frac{W^{-1}}{2} \sum_{t=2}^T (\lambda_t^2 - 2\lambda_t \lambda_{t-1} + \lambda_{t-1}^2) \right\} \end{aligned}$$

Como $\sum_{t=2}^T \lambda_t^2 + \sum_{t=2}^T \lambda_{t-1}^2 = \lambda_1^2 + \lambda_T^2 + 2 \sum_{t=2}^{T-1} \lambda_t^2$, é possível reescrever a distribuição conjunta dos parâmetros do campo latente por

$$p(\lambda_1, \dots, \lambda_T) \propto \exp \left\{ -\frac{1}{2} [(W^{-1} + (C_0 + W)^{-1})\lambda_1^2 + \right. \quad (5.33)$$

$$\left. + W^{-1} \left(\lambda_T^2 + 2 \sum_{t=2}^{T-1} \lambda_t^2 - 2 \sum_{t=2}^T \lambda_t \lambda_{t-1} \right) \right\} \quad (5.34)$$

Assim, condicionalmente a W , a distribuição conjunta de $(\lambda_1, \dots, \lambda_T)$ é dada por

$$p(\lambda_1, \dots, \lambda_T | W) \propto \exp \left\{ -\frac{1}{2} \boldsymbol{\lambda}^T \Phi \boldsymbol{\lambda} \right\}, \quad (5.35)$$

em que

$$\Phi = \begin{pmatrix} W^{-1} + (C_0 + W)^{-1} & -W^{-1} & 0 & \dots & 0 & 0 \\ -W^{-1} & 2W^{-1} & -W^{-1} & \dots & 0 & 0 \\ 0 & -W^{-1} & 2W^{-1} & \dots & 0 & 0 \\ 0 & 0 & -W^{-1} & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & -W^{-1} & 0 \\ 0 & 0 & 0 & \dots & 2W^{-1} & -W^{-1} \\ 0 & 0 & 0 & \dots & -W^{-1} & W^{-1} \end{pmatrix}_{T \times T} \quad (5.36)$$

Considerando uma especificação *a priori* não informativa, $C_0 \rightarrow \infty$, pode-se reescrever Φ da seguinte forma

$$\Phi = W^{-1} \begin{pmatrix} 1 & -1 & 0 & \dots & 0 & 0 \\ -1 & 2 & -1 & \dots & 0 & 0 \\ 0 & -1 & 2 & \dots & 0 & 0 \\ 0 & 0 & -1 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & -1 & 0 \\ 0 & 0 & 0 & \dots & 2 & -1 \\ 0 & 0 & 0 & \dots & -1 & 1 \end{pmatrix}_{T \times T} \quad (5.37)$$

5.8.1 Resultados

Uma série temporal de dados artificiais de tamanho $T = 100$ foi gerada de um modelo Poisson dinâmico descrito anteriormente, fixando-se $W = 0.1$ e $\lambda_0 = 1$.

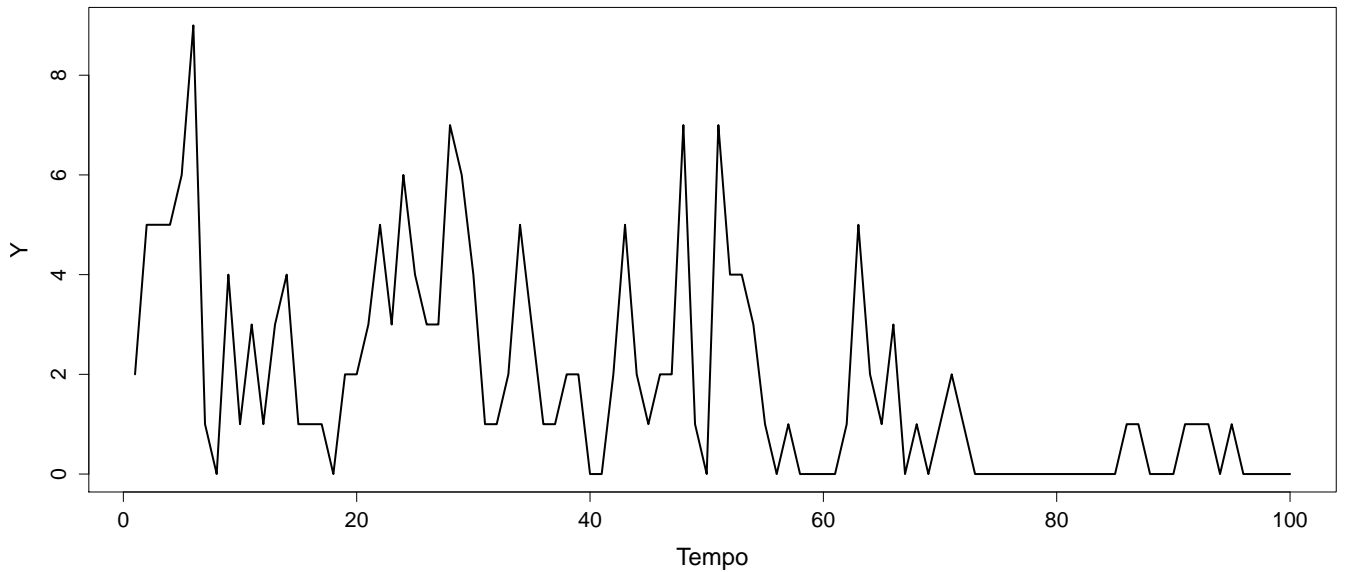


Figura 5.1: Série temporal artificial de tamanho $T = 100$.

Considerou-se a matriz de precisão *a priori* para o campo latente como descrita em (5.36) e assumiu-se que, *a priori*, $W^{-1} \sim Gama(1, 0.1)$.

A aproximação da distribuição *a posteriori* de W^{-1} pelo método INLA pode ser vista na Figura 5.2. Já a Figura 5.3 mostra os valores estimados para vetor de estados. Pode-se observar que os parâmetros foram bem estimados.

Do ponto de vista computacional, a estimação foi feita em segundos. Caso o mesmo problema fosse aplicado utilizando MCMC, o tempo computacional seria significativamente superior.

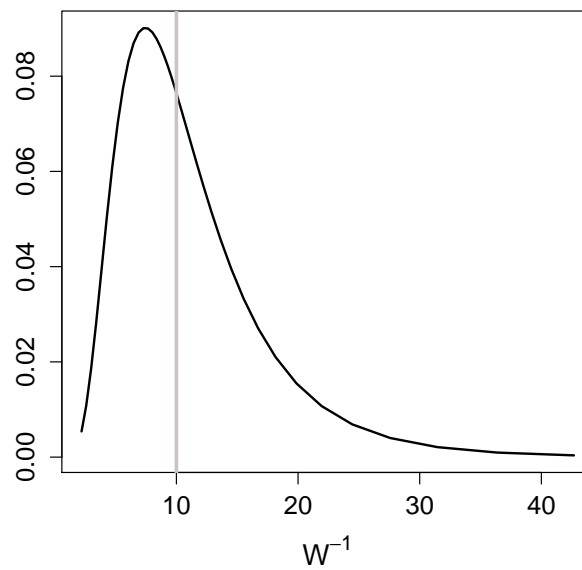


Figura 5.2: Distribuição *a posteriori* de W^{-1} aproximada pelo método INLA. A linha vertical refere-se ao valor verdadeiro.

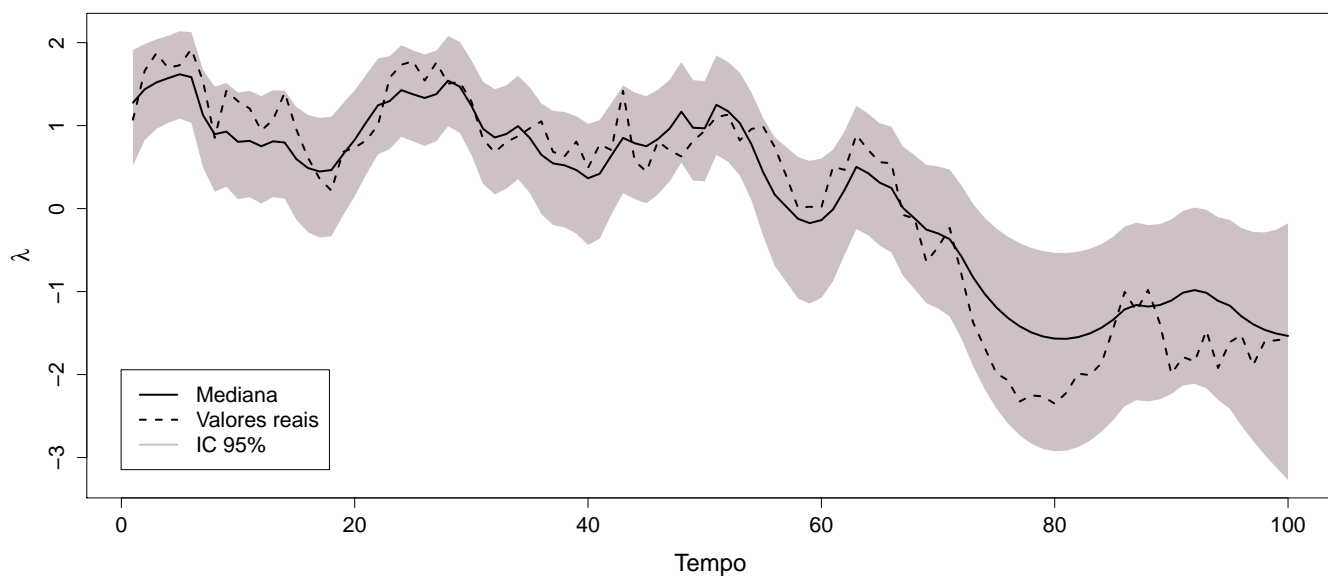


Figura 5.3: Valores estimados para $\lambda_1, \dots, \lambda_{100}$ pelo método INLA, valores reais (linha tracejada) e intervalo de 95% de credibilidade (área preenchida).

Capítulo 6

Extensões

Como visto no capítulo anterior, o método de inferência proposto por [Rue et al. \(2009\)](#) é aplicável em uma classe específica de modelos, os modelos latentes gaussianos. Neste capítulo, uma proposta de extensão para viabilizar a aplicação do método INLA em modelos latentes não-gaussianos é analisada.

6.1 INLA para Modelos Latentes Não-Gaussianos

Para lidar com situações em que se assume que a distribuição do campo latente não é gaussiana, é necessária uma modificação na abordagem proposta por [Rue et al. \(2009\)](#). Agora, considera-se a realização de inferência bayesiana aproximada via INLA em modelos com a seguinte estrutura hierárquica:

$$y_i | \mathbf{x}, \boldsymbol{\theta} \sim p(y_i | x_{\mathbf{I}_i}, \boldsymbol{\theta}) \quad (6.1a)$$

$$\mathbf{x} | \boldsymbol{\theta} \sim p(\mathbf{x} | \boldsymbol{\theta}) \quad (6.1b)$$

$$\boldsymbol{\theta} \sim p(\boldsymbol{\theta}), \quad (6.1c)$$

em que $\mathbf{y} = (y_1, \dots, y_{n_d})'$ são os dados observados, $\mathbf{x} = (x_1, \dots, x_n)'$ os elementos do campo latente, $\boldsymbol{\theta}$ é o vetor m -dimensional de hiperparâmetros, a distribuição de $y_i | \mathbf{x}, \boldsymbol{\theta}$ pertence à família exponencial e a distribuição de $\mathbf{x} | \boldsymbol{\theta}$ não é necessariamente gaussiana.

Note que, na estrutura acima, além de relaxar a hipótese de que o campo latente $\mathbf{x}|\boldsymbol{\theta}$ tenha que ser necessariamente gaussiano, considera-se a possibilidade de que cada y_i pode estar relacionado com um conjunto \mathbf{I}_i de elementos do campo latente.

Assim como nos modelos latentes gaussianos, o interesse aqui será nas seguintes distribuições marginais

$$p(x_i|\mathbf{y}) = \int p(\boldsymbol{\theta}|\mathbf{y})p(x_i|\boldsymbol{\theta}, \mathbf{y})d\boldsymbol{\theta} \quad (6.2a)$$

$$p(\theta_j|\mathbf{y}) = \int p(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta}_{-i}, \quad (6.2b)$$

$\forall i = 1, \dots, n$ e $\forall j = 1, \dots, m$.

Uma abordagem natural seria considerar uma aproximação gaussiana para o campo latente e trabalhar com um campo latente gaussiano aproximado. Este trabalho traz uma outra adaptação proposta à abordagem descrita em [Rue et al. \(2009\)](#), a qual pode ser vista nas seções a seguir.

6.2 Esquema de Aproximação

O esquema de aproximação se mantém similar ao proposto em [Rue et al. \(2009\)](#). Quando as integrações em (5.2) não podem ser resolvidas analiticamente, constrói-se aproximações para $p(x_i|\boldsymbol{\theta}, \mathbf{y})$ e $p(\boldsymbol{\theta}|\mathbf{y})$ dadas respectivamente por $\tilde{p}(x_i|\boldsymbol{\theta}, \mathbf{y})$ e $\tilde{p}(\boldsymbol{\theta}|\mathbf{y})$. Assim, é possível construir aproximações "aninhadas":

$$\tilde{p}(x_i|\mathbf{y}) = \int \tilde{p}(\boldsymbol{\theta}|\mathbf{y})\tilde{p}(x_i|\boldsymbol{\theta}, \mathbf{y})d\boldsymbol{\theta} \quad (6.3a)$$

$$\tilde{p}(\theta_j|\mathbf{y}) = \int \tilde{p}(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta}_{-i}, \quad (6.3b)$$

sendo as integrais resolvidas via integração numérica.

As próximas seções mostram uma proposta de como obter essas aproximações necessárias.

6.3 Caso com Independência Condicional

Considere a seguinte estrutura hierárquica:

$$y_i | \mathbf{x}, \boldsymbol{\theta} \sim p(y_i | x_i, \boldsymbol{\theta}) \quad (6.4a)$$

$$\mathbf{x} | \boldsymbol{\theta} \sim p(\mathbf{x} | \boldsymbol{\theta}) \quad (6.4b)$$

$$\boldsymbol{\theta} \sim p(\boldsymbol{\theta}), \quad (6.4c)$$

em que $\mathbf{y} = (y_1, \dots, y_{n_d})'$ são os dados observados, $\mathbf{x} = (x_1, \dots, x_n)'$ os elementos do campo latente, $\boldsymbol{\theta}$ é o vetor m -dimensional de hiperparâmetros, a distribuição de $y_i | \mathbf{x}, \boldsymbol{\theta}$ pertence à família exponencial e a distribuição de $\mathbf{x} | \boldsymbol{\theta}$ não é gaussiana.

Note que esta estrutura é um caso particular da estrutura anterior. Neste caso, um determinado y_i está conectado a apenas um elemento do campo latente. Assim sendo, este caso considera apenas o relaxamento da hipótese de normalidade do campo latente como proposto por [Rue et al. \(2009\)](#) e mantém a mesma estrutura de dependência.

As subseções que seguem, tratam da adaptação da abordagem do método INLA para modelos com esta estrutura hierárquica.

6.3.1 Aproximação para $p(\boldsymbol{\theta} | \mathbf{y})$

A aproximação para $p(\boldsymbol{\theta} | \mathbf{y})$ é dada por:

$$\tilde{p}(\boldsymbol{\theta} | \mathbf{y}) \propto \frac{p(\mathbf{x}, \boldsymbol{\theta}, \mathbf{y})}{\tilde{p}_G(\mathbf{x} | \boldsymbol{\theta}, \mathbf{y})} \Big|_{\mathbf{x}=\mathbf{x}^*(\boldsymbol{\theta})}, \quad (6.5)$$

em que $\tilde{p}_G(\mathbf{x} | \boldsymbol{\theta}, \mathbf{y})$ é a aproximação gaussiana da condicional completa de \mathbf{x} e $\mathbf{x}^*(\boldsymbol{\theta})$ é a moda da condicional completa de \mathbf{x} para um dado $\boldsymbol{\theta}$.

Tem-se que

$$p(\mathbf{x} | \boldsymbol{\theta}, \mathbf{y}) \propto \exp \left\{ \log(p(\mathbf{x} | \boldsymbol{\theta})) + \sum_i \log(p(y_i | x_i, \boldsymbol{\theta})) \right\}. \quad (6.6)$$

Diferentemente do que já foi visto, o núcleo de $\log(p(\mathbf{x}|\boldsymbol{\theta}))$ não é um termo quadrático. Para se encontrar $\tilde{p}_G(\mathbf{x}|\boldsymbol{\theta}, \mathbf{y})$, propõe-se obter a aproximação gaussiana de $\log(p(\mathbf{x}|\boldsymbol{\theta}))$, além de $\log(p(y_i|x_i, \boldsymbol{\theta}))$.

Assim,

$$\tilde{p}(\mathbf{x}|\boldsymbol{\theta}, \mathbf{y}) \propto \exp \left\{ -\frac{1}{2} \mathbf{x}' (\mathbf{Q}^* + \text{diag}(\mathbf{c})) \mathbf{x} + (\mathbf{b}^* + \mathbf{b})' \mathbf{x} \right\}, \quad (6.7)$$

em que \mathbf{b}^* e \mathbf{Q}^* vêm da expansão de Taylor de $\log(p(\mathbf{x}|\boldsymbol{\theta}))$, sendo \mathbf{Q}^* uma matriz esparsa banda-diagonal, devido à característica markoviana dos modelos considerados e \mathbf{b} e \mathbf{c} são coeficientes da expansão de $\log(p(y_i|x_i, \boldsymbol{\theta}))$.

O numerador da aproximação é dado por $p(\mathbf{x}, \boldsymbol{\theta}, \mathbf{y}) = p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})$. É importante destacar que $p(\mathbf{x}|\boldsymbol{\theta})$ se mantém em sua estrutura original não-gaussiana. Caso fosse escolhido obter uma aproximação gaussiana neste termo, estaríamos lidando com um campo latente gaussiano aproximado, porém isto não faz parte da abordagem sugerida.

Assim como em [Rue et al. \(2009\)](#), não é necessário representar $\tilde{p}(\boldsymbol{\theta}|\mathbf{y})$ parametricamente, mas sim explorá-la suficientemente bem a fim de selecionar bons pontos para a integração numérica. Com isso, procedimento análogo ao considerado na abordagem de [Rue et al. \(2009\)](#) para campos latentes gaussianos poderia ser utilizado.

6.3.2 Aproximação para $p(x_i|\boldsymbol{\theta}, \mathbf{y})$

Como forma rápida e simples de aproximar $p(x_i|\boldsymbol{\theta}, \mathbf{y})$, pode-se utilizar a distribuição marginal de $\tilde{p}_G(\mathbf{x}|\boldsymbol{\theta}, \mathbf{y})$, assim como foi sugerido em (5.14) para aplicação em campos latentes gaussianos.

Uma outra possibilidade é usar a Aproximação de Laplace. A aproximação para $p(x_i|\boldsymbol{\theta}, \mathbf{y})$ é dada por:

$$\tilde{p}_{LA}(x_i|\boldsymbol{\theta}, \mathbf{y}) \propto \frac{p(\mathbf{x}, \boldsymbol{\theta}, \mathbf{y})}{\tilde{p}_{GG}(\mathbf{x}_{-i}|x_i, \boldsymbol{\theta}, \mathbf{y})} \Big|_{\mathbf{x}_{-i}=\mathbf{x}_{-i}^*(x_i, \boldsymbol{\theta})}, \quad (6.8)$$

em que $\tilde{p}_{GG}(\mathbf{x}_{-i}|x_i, \boldsymbol{\theta}, \mathbf{y})$ é a aproximação gaussiana para $\mathbf{x}_{-i}|x_i, \boldsymbol{\theta}, \mathbf{y}$ e $\mathbf{x}_{-i}^*(x_i, \boldsymbol{\theta})$ é a moda de $p(\mathbf{x}_{-i}|x_i, \boldsymbol{\theta}, \mathbf{y})$.

O numerador da aproximação é dado por $p(\mathbf{x}, \boldsymbol{\theta}, \mathbf{y}) = p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})$. É importante destacar que $p(\mathbf{x}|\boldsymbol{\theta})$ se manterá em sua estrutura original não-gaussiana.

Pode-se, também, evitar o passo de otimização e computar $\tilde{p}_{GG}(\mathbf{x}_{-i}|x_i, \boldsymbol{\theta}, \mathbf{y})$ através de uma aproximação para a moda $\mathbf{x}_{-i}^*(x_i, \boldsymbol{\theta}) \approx E_{\tilde{p}_G}(\mathbf{x}_{-i}|x_i)$.

Acredita-se que esta modificação na abordagem não influenciará em termos de tempo computacional, já que, neste caso, também pode-se tirar proveito da esparsidade matriz de precisão da distribuição aproximada \tilde{p}_{GG} .

6.4 Caso Geral

Voltemos ao caso geral descrito no início do capítulo em que cada elemento y_i se relaciona com um conjunto \mathbf{I}_i de elementos do campo latente.

Neste caso, a matriz de precisão oriunda da aproximação gaussiana na distribuição condicional completa do campo latente não é mais esparsa. Esta situação pode trazer problemas computacionais, uma vez que é mais difícil lidar com matrizes densas.

Por vezes, ela pode vir a não ser positiva definida, o que será necessário fazer adaptações no algoritmo para lidar com essas situações. Mais ainda, pode-se verificar que o tempo computacional dependerá diretamente da dimensão paramétrica. Em situações em que se há muitos parâmetros, a inferência pode ser custosa computacionalmente, porém ainda menos custoso que utilizar métodos MCMC.

Esta situação foi considerada em [Martins \(2010\)](#) para modelos latentes gaussianos. Em outras palavras, [Martins \(2010\)](#) considerou o caso em que y_i se relaciona com um conjunto \mathbf{I}_i de elementos de um campo latente gaussiano. Note que, neste trabalho, além de relaxar esta hipótese, o campo latente não é mais gaussiano.

6.4.1 Aproximação para $p(\boldsymbol{\theta}|\mathbf{y})$

A aproximação para $p(\boldsymbol{\theta}|\mathbf{y})$ é dada por:

$$\tilde{p}(\boldsymbol{\theta}|\mathbf{y}) \propto \frac{p(\mathbf{x}, \boldsymbol{\theta}, \mathbf{y})}{\tilde{p}_G(\mathbf{x}|\boldsymbol{\theta}, \mathbf{y})} \Big|_{\mathbf{x}=\mathbf{x}^*(\boldsymbol{\theta})}, \quad (6.9)$$

em que $\tilde{p}_G(\mathbf{x}|\boldsymbol{\theta}, \mathbf{y})$ é a aproximação gaussiana da condicional completa de \mathbf{x} e $\mathbf{x}^*(\boldsymbol{\theta})$ é a moda da condicional completa de \mathbf{x} para um dado $\boldsymbol{\theta}$.

A distribuição $p(\mathbf{x}|\boldsymbol{\theta}, \mathbf{y})$ é dada por

$$p(\mathbf{x}|\boldsymbol{\theta}, \mathbf{y}) \propto \exp\{\log(p(\mathbf{x}|\boldsymbol{\theta})) + \sum_i \log(p(y_i|x_{I_i}, \boldsymbol{\theta}))\}. \quad (6.10)$$

A proposta aqui é realizar uma aproximação gaussiana, em $\log(p(\mathbf{x}|\boldsymbol{\theta}))$ além de $\log(p(y_i|x_{I_i}, \boldsymbol{\theta}))$, para cada $i = 1, \dots, n$. Neste caso, assume-se que

$$\tilde{p}(\mathbf{x}|\boldsymbol{\theta}, \mathbf{y}) \propto \exp\left\{-\frac{1}{2}\mathbf{x}'(\mathbf{Q}^* + C(x^*))\mathbf{x} + (\mathbf{b}^* + \mathbf{b})'\mathbf{x}\right\}, \quad (6.11)$$

em que \mathbf{b}^* e \mathbf{Q}^* vêm da expansão de Taylor de $\log(p(\mathbf{x}|\boldsymbol{\theta}))$, sendo \mathbf{Q}^* uma matriz densa e \mathbf{b} e \mathbf{c} são coeficientes da expansão de $\log(p(y_i|x_{I_i}, \boldsymbol{\theta}))$.

Assim como no caso anterior, o numerador da aproximação é dado por $p(\mathbf{x}, \boldsymbol{\theta}, \mathbf{y}) = p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})$ e $p(\mathbf{x}|\boldsymbol{\theta})$ se mantém em sua estrutura original não-gaussiana.

A moda $\mathbf{x}^*(\boldsymbol{\theta})$ pode ser obtida usando um dos métodos descritos na Seção 2.2. Para isso, será necessário avaliar a matriz de precisão \mathbf{Q}^* em diversos valores $\mathbf{x}_0, \mathbf{x}_1, \dots, \mathbf{x}_k$, em cada uma das k -ésimas iterações até o algoritmo escolhido alcançar convergência. É importante ressaltar, entretanto, que a matriz de precisão \mathbf{Q}^* para certos valores não é positiva definida. Motivo disto pode ser o fato de que a matriz $C(\cdot)$ é uma matriz densa.

Existem diversos métodos na literatura que propõem transformar uma matriz que não é positiva definida, como por exemplo, [Knol e ten Berge \(1989\)](#), [Cheng e Higham \(1998\)](#) e [Higham \(2002\)](#). [Martins \(2010\)](#) sugere uma abordagem simples, em que se aumenta os valores da diagonal principal de $C(\cdot)$ até que a matriz seja positiva definida. Ou ainda, pode-se reduzir os valores fora da diagonal principal. Note que aumentar os valores da diagonal principal não altera as relações de dependência entre os elementos. O problema da matriz não ser positiva definida, em geral, ocorre quando o algoritmo ainda se encontra distante da moda.

6.4.2 Aproximação para $p(x_i|\boldsymbol{\theta}, \mathbf{y})$

Assim como nos casos considerados anteriormente, como forma rápida e simples de aproximar $p(x_i|\boldsymbol{\theta}, \mathbf{y})$, pode-se utilizar a distribuição marginal de $\tilde{p}_G(\mathbf{x}|\boldsymbol{\theta}, \mathbf{y})$, como sugerido em (5.14).

Uma outra possibilidade é usar a Aproximação de Laplace. A aproximação para $p(x_i|\boldsymbol{\theta}, \mathbf{y})$ é dada por:

$$\tilde{p}_{LA}(x_i|\boldsymbol{\theta}, \mathbf{y}) \propto \frac{p(\mathbf{x}, \boldsymbol{\theta}, \mathbf{y})}{\tilde{p}_{GG}(\mathbf{x}_{-i}|x_i, \boldsymbol{\theta}, \mathbf{y})} \Big|_{\mathbf{x}_{-i}=\mathbf{x}_{-i}^*(x_i, \boldsymbol{\theta})}, \quad (6.12)$$

em que $\tilde{p}_{GG}(\mathbf{x}_{-i}|x_i, \boldsymbol{\theta}, \mathbf{y})$ é a aproximação gaussiana para $\mathbf{x}_{-i}|x_i, \boldsymbol{\theta}, \mathbf{y}$ e $\mathbf{x}_{-i}^*(x_i, \boldsymbol{\theta})$ é a moda de $p(\mathbf{x}_{-i}|x_i, \boldsymbol{\theta}, \mathbf{y})$.

O numerador da aproximação é dado por $p(\mathbf{x}, \boldsymbol{\theta}, \mathbf{y}) = p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})$. Considere que $p(\mathbf{x}|\boldsymbol{\theta})$ se mantém em sua estrutura original não-gaussiana.

Pode-se, também, evitar o passo de otimização e computar $\tilde{p}_{GG}(\mathbf{x}_{-i}|x_i, \boldsymbol{\theta}, \mathbf{y})$ através de uma aproximação para a moda $\mathbf{x}_{-i}^*(x_i, \boldsymbol{\theta}) \approx E_{\tilde{p}_G}(\mathbf{x}_{-i}|x_i)$.

6.5 Exemplo: Modelo Poisson não-Gaussiano

6.5.1 Descrição do Modelo

Suponha variáveis aleatórias Y_i , para $i = 1, \dots, n$ condicionalmente independentes seguindo uma distribuição Poisson com média $\lambda_i > 0$. Assume-se que a média λ_i varia no tempo através de uma estrutura dinâmica, ou seja,

$$\begin{aligned} Y_i|\lambda_i &\sim \text{Poisson}(\lambda_i) \\ \lambda_i &= \frac{\lambda_{i-1}\eta_i}{w} \end{aligned}$$

em que

$$\begin{aligned}\eta_i|w, D_{i-1} &\sim \text{Beta}(wa_{i-1}, (1-w)a_{i-1}), \\ \lambda_0|D_0 &\sim \text{Gamma}(a_0, b_0) \\ w &\sim p(w),\end{aligned}$$

o parâmetro a_i é oriundo da distribuição *a posteriori* de $\lambda_i|D_i$, para $i = 1, \dots, n$, D_i denota o conjunto de informação no tempo i , $i = 1, 2, \dots, n$ e a_0 e b_0 são constantes conhecidas.

Mais ainda, é valido que

$$\eta_i|w, D_0 \sim \text{Beta}(w^i a_0, (1-w)w^{i-1} a_0), \quad (6.13)$$

em que D_0 denota o conjunto de informação inicial, anterior à observação dos dados.

De fato, tem-se que

$$p(\boldsymbol{\lambda} | w, D_0) = \prod_i p(\lambda_i | \lambda_1, \dots, \lambda_{i-1}, D_0) \quad (6.14a)$$

$$= \prod_i p(\lambda_i | \lambda_{i-1}, D_0) \quad (6.14b)$$

Sabendo-se que

$$\frac{w\lambda_1}{\lambda_0} | w, \lambda_0, D_0 \sim \text{Beta}(wa_0, (1-w)a_0) \quad (6.15)$$

e

$$\lambda_0|D_0 \sim \text{Gama}(a_0, b_0) \quad (6.16)$$

conclui-se que

$$\lambda_1|w, D_0 \sim \text{Gama}(wa_0, wb_0). \quad (6.17)$$

De forma análoga,

$$\frac{w\lambda_2}{\lambda_1} | w, \lambda_1, D_1 \sim \text{Beta}(w a_1, (1-w)a_1) \quad (6.18)$$

e

$$\lambda_1 | D_1 \sim \text{Gama}(a_1, b_1) \quad (6.19)$$

conclui-se que

$$\lambda_2 | w, D_1 \sim \text{Gama}(w a_1, w b_1). \quad (6.20)$$

Se $D_i = D_{i-1} = \dots = D_1 = D_0$ tem-se que

$$\lambda_1 | w, D_0 \sim \text{Gama}(w a_0, w b_0). \quad (6.21)$$

Como (6.18) vale, consequentemente,

$$\lambda_2 | w, D_0 \sim \text{Gama}(w^2 a_0, w^2 b_0). \quad (6.22)$$

De forma mais geral, tem-se que:

$$\lambda_i | w, D_0 \sim \text{Gama}(w^i a_0, w^i b_0). \quad (6.23)$$

Também é válido que

$$\frac{w\lambda_i}{\lambda_{i-1}} | w, \lambda_{i-1}, D_{i-1} \sim \text{Beta}(w a_{i-1}, (1-w)a_{i-1}). \quad (6.24)$$

Ora, através da relação (6.23), como o caso considerado é aquele em que $D_{i-1} = \dots = D_1 = D_0$ é válido, pode-se reescrever (6.24) tal que

$$(\eta_i | w, D_0) \equiv \left(\frac{w\lambda_i}{\lambda_{i-1}} | w, \lambda_{i-1}, D_0 \right) \sim \text{Beta}(w^i a_0, (1-w)w^{i-1} a_0), \quad (6.25)$$

pois $a_{i-1} = w^{i-1} a_0$.

6.5.2 Procedimento de Inferência

Caso 1

Considere nesta seção o procedimento de inferência para $\Psi = (\theta, \lambda)$.

A verossimilhança é dada por:

$$p(y_i | \lambda_i, \theta) = \frac{e^{-\lambda_i} \lambda_i^{y_i}}{y_i!}, \quad (6.26)$$

para $y_i = 0, 1, 2, \dots$ e $\lambda_i > 0$, $i = 0, 1, \dots, n$.

A distribuição *a priori* é dada por:

$$p(\Psi) = p(\lambda | \theta)p(\theta), \quad (6.27)$$

em que

$$\begin{aligned} p(\lambda | \theta) &\propto \lambda_1^{a_0-1} e^{-b_0 \lambda_1} \times \\ &\times \prod_{i=2}^n \left[\frac{1}{\lambda_{i-1}} \left(\frac{w \lambda_i}{\lambda_{i-1}} \right)^{w^i a_0 - 1} \left(1 - \frac{w \lambda_i}{\lambda_{i-1}} \right)^{w^{i-1} (1-w) a_0 - 1} \right] \end{aligned} \quad (6.28)$$

Pelo Teorema de Bayes, a distribuição *a posteriori* é dada por:

$$p(\Psi | \mathbf{y}) \propto p(\theta) p(\lambda | \theta) \prod_i p(y_i | \lambda_i, \theta) \quad (6.29)$$

Considere o método de aproximação INLA.

A distribuição condicional completa de λ é dada por:

$$p(\lambda | \theta, \mathbf{y}) \propto \exp\{\log(p(\lambda | \theta)) + \sum_i \log(p(y_i | \lambda_i, \theta))\} \quad (6.30)$$

Denote $g_i(\lambda_i) = \log(p(y_i | \lambda_i, \theta))$. Segundo [Martins \(2010\)](#), a expansão de Taylor realizada ao redor da moda melhora a aproximação gaussiana. Assim sendo, seja $\lambda^{(0)}$ um vetor de

valores iniciais a serem utilizados para inicialização do algoritmo de busca da moda de $p(\boldsymbol{\theta} | \mathbf{y})$.

Expandindo $g_i(\lambda_i)$ em torno de $\boldsymbol{\lambda}^{(0)}$, obtém-se

$$\sum_i \log(p(y_i | \lambda_i, \boldsymbol{\theta})) \approx -\frac{1}{2} \boldsymbol{\lambda}' C(\boldsymbol{\lambda}^{(0)}) \boldsymbol{\lambda} + b'(\boldsymbol{\lambda}^{(0)}) \boldsymbol{\lambda}, \quad (6.31)$$

em que

$$b(\boldsymbol{\lambda}^{(0)}) = \begin{bmatrix} \frac{2y_1}{\lambda_1^{(0)}} \\ \frac{2y_2}{\lambda_2^{(0)}} \\ \vdots \\ \frac{2y_n}{\lambda_n^{(0)}} \end{bmatrix} \quad (6.32)$$

e

$$C(\boldsymbol{\lambda}^{(0)}) = \text{diag} \left(\frac{y_1}{\lambda_1^{(0)}}, \frac{y_2}{\lambda_2^{(0)}}, \dots, \frac{y_n}{\lambda_n^{(0)}} \right) \quad (6.33)$$

Já a aproximação gaussiana de $\log(p(\boldsymbol{\lambda} | \boldsymbol{\theta}))$ é dada por

$$\log(p(\boldsymbol{\lambda} | \boldsymbol{\theta})) \approx -\frac{1}{2} \boldsymbol{\lambda}' Q(\boldsymbol{\lambda}^{(0)}) \boldsymbol{\lambda} + b^*(\boldsymbol{\lambda}^{(0)}) \boldsymbol{\lambda}, \quad (6.34)$$

em que

$$b^*(\boldsymbol{\lambda}^{(0)}) = \nabla_f(\boldsymbol{\lambda}^{(0)}) - H_f(\boldsymbol{\lambda}^{(0)}) \boldsymbol{\lambda}^{(0)} \quad (6.35)$$

e

$$Q(\boldsymbol{\lambda}^{(0)}) = -H_f(\boldsymbol{\lambda}^{(0)}) \quad (6.36)$$

denotando $f = \log(p(\boldsymbol{\lambda} | \boldsymbol{\theta}))$.

Abaixo seguem as derivadas necessárias para obter ∇_f e H_f .

$$\frac{\partial \log(p(\boldsymbol{\lambda} | \boldsymbol{\theta}))}{\partial \lambda_1} = -b_0 + \frac{-2w^2a_0 + w^3a_0 - 1}{\lambda_1} + \frac{a_0 - 1}{\lambda_1} + \frac{(1-w)w^2a_0 - 1}{\lambda_1 - w\lambda_2}$$

$$\frac{\partial \log(p(\boldsymbol{\lambda} | \boldsymbol{\theta}))}{\partial \lambda_i} = \frac{w^i a_0 - 1}{\lambda_i} - \frac{w((1-w)w^i a_0 - 1)}{\lambda_{i-1} - w\lambda_i} + \frac{-2w^{i+1}a_0 - w^{i+2}a_0 + 1}{\lambda_i} + \frac{(1-w)w^{i+1}a_0 - 1}{\lambda_i - w\lambda_{i+1}},$$

para $i = 2, \dots, n-1$.

$$\frac{\partial \log(p(\boldsymbol{\lambda} | \boldsymbol{\theta}))}{\partial \lambda_n} = \frac{w^n a_0 - 1}{\lambda_n} - \frac{w((1-w)w^n a_0 - 1)}{\lambda_{n-1} - w\lambda_n}$$

$$\frac{\partial^2 \log(p(\boldsymbol{\lambda} | \boldsymbol{\theta}))}{\partial \lambda_1^2} = -\frac{-2w^2a_0 + w^3a_0 - 1}{\lambda_1^2} - \frac{wa_0 - 1}{\lambda_1^2} - \frac{(1-w)w^2a_0 - 1}{(\lambda_1 - w\lambda_2)^2}$$

$$\frac{\partial^2 \log(p(\boldsymbol{\lambda} | \boldsymbol{\theta}))}{\partial \lambda_i^2} = -\frac{w^i a_0 - 1}{\lambda_i^2} - \frac{w^2[(1-w)w^i a_0 - 1]}{(\lambda_{i-1} - w\lambda_i)^2} - \frac{-2w^{i+1}a_0 - w^{i+2}a_0 + 1}{\lambda_i^2} - \frac{(1-w)w^{i+1}a_0 - 1}{(\lambda_i - w\lambda_{i+1})^2},$$

para $i = 2, \dots, n$.

$$\frac{\partial^2 \log(p(\boldsymbol{\lambda} | \boldsymbol{\theta}))}{\partial \lambda_1 \partial \lambda_2} = \frac{w((1-w)w^2a_0 - 1)}{\lambda_1 - w\lambda_2}$$

$$\frac{\partial^2 \log(p(\boldsymbol{\lambda} | \boldsymbol{\theta}))}{\partial \lambda_{i+1} \partial \lambda_i} = \frac{w((1-w)w^{i+1}a_0 - 1)}{(\lambda_i - w\lambda_{i+1})^2},$$

para $i = 2, \dots, n-1$.

As derivadas necessárias para as aproximações acima podem ser feitas de forma automatizada em programas estatísticos. Em geral, estas as derivadas são obtidas numericamente via diferenças finitas. Ver [Burden e Faires \(2001\)](#) para mais detalhes.

Como o vetor de estados $\boldsymbol{\lambda}$ só assume valores positivos, conforme discutido na Seção 5.3, é recomendável transformar o vetor de estados para que ele varie na reta real. O motivo disto é que, do ponto de vista computacional, esta transformação permite que sejam utilizados algoritmos de otimização sem restrição, descritos na Seção 2.2.

Os Casos 2 e 3 citados a seguir consideram transformações no vetor de estados para que este varie na reta real.

Caso 2

Uma transformação razoável para o vetor de estados que só assume valores positivos é a transformação logarítmica no vetor de estados. Agora, o interesse é em $x_i = \log(\lambda_i)$, $\forall i = 1, \dots, n$. Com isso, este caso considera que $\Psi = (\mathbf{x}, \boldsymbol{\theta})$.

A verossimilhança é dada por:

$$p(y_i | x_i, \boldsymbol{\theta}) = \frac{e^{-e^{x_i}} e^{x_i y_i}}{y_i!}, \quad (6.37)$$

para $y_i = 0, 1, 2, \dots$ e $i = 1, \dots, n$.

A distribuição *a priori* é dada por:

$$p(\Psi) = p(\mathbf{x} | \boldsymbol{\theta})p(\boldsymbol{\theta}). \quad (6.38)$$

Neste caso, tem-se que

$$p(\mathbf{x} | \boldsymbol{\theta}) \propto e^{x_1(a_0-1)} e^{-b_0 e^{x_1}} e^{x_1} \prod_{i=2}^n \left[\frac{e^{x_i}}{e^{x_{i-1}}} \left(\frac{w e^{x_i}}{e^{x_{i-1}}} \right)^{w^i a_0 - 1} \left(1 - \frac{w e^{x_i}}{e^{x_{i-1}}} \right)^{(1-w)w^i a_0 - 1} \right] \quad (6.39)$$

Pelo Teorema de Bayes, a distribuição *a posteriori* é dada por:

$$p(\Psi | \mathbf{y}) \propto p(\boldsymbol{\theta})p(\mathbf{x} | \boldsymbol{\theta}) \prod_i p(y_i | x_i, \boldsymbol{\theta}) \quad (6.40)$$

Considere o método de aproximação INLA.

A distribuição condicional completa de \mathbf{x} é dada por:

$$p(\mathbf{x} | \boldsymbol{\theta}, \mathbf{y}) \propto \exp\{\log(p(\mathbf{x} | \boldsymbol{\theta})) + \sum_i \log(p(y_i | x_i, \boldsymbol{\theta}))\} \quad (6.41)$$

Considerando a transformação logarítmica, indiretamente assume-se que deva ser verdade que $e^{x_{i-1}} > w e^{x_i - 1}$. Isto vem do fato de que, ao se aplicar o logaritmo em $p(\mathbf{x} | \boldsymbol{\theta})$ para se obter a aproximação gaussiana da distribuição condicional completa de \mathbf{x} , tal relação tem que ser válida.

Em muitos casos, é possível que essa relação não seja válida. Problemas decorrentes da transformação logarítmica neste exemplo específico impedem a aplicação do método INLA, uma vez que seria necessário um algoritmo de otimização de buscas de moda em que é necessário avaliar $p(\mathbf{x} | \boldsymbol{\theta}, \mathbf{y})$ para vários valores $\mathbf{x}^{(0)}, \mathbf{x}^{(1)}, \dots, \mathbf{x}^{(k)}$ em cada uma da k -ésima iteração.

Muitos erros numéricos podem vir a surgir ao se considerar a transformação logarítmica para este modelo específico.

Com isso, nota-se que muito cuidado deve ser tomado ao escolher as transformações adequadas para o vetor paramétrico.

Caso 3

Considere agora uma nova transformação para o vetor de estados:

$$z_1 = \log(\lambda_1) \quad (6.42a)$$

$$z_i = \text{logit}(\eta_i) = \log\left(\frac{w\lambda_i}{\lambda_{i-1} - w\lambda_i}\right), i = 2, \dots, n \quad (6.42b)$$

Com isso, tem-se que

$$\lambda_1 = e^{z_1} \quad (6.43a)$$

$$\lambda_i = \frac{e^{z_1}/w^{i-1}}{\prod_{j=2}^i (1 + e^{-z_j})}, i = 2, \dots, n \quad (6.43b)$$

A verossimilhança é dada por:

$$p(\mathbf{y} | \mathbf{z}, \boldsymbol{\theta}) = \prod_i p(y_i | z_{1:i}, \boldsymbol{\theta}) \quad (6.44)$$

A distribuição *a priori* é dada por:

$$p(\mathbf{z}, \boldsymbol{\theta}) = p(\mathbf{z} | \boldsymbol{\theta})p(\boldsymbol{\theta}), \quad (6.45)$$

em que

$$p(\mathbf{z} | \boldsymbol{\theta}) \propto e^{z_1(wa_0-1)} e^{-wb_0e^{z_1}} e^{z_1} \prod_{i=2}^n \left[(1 + e^{-z_i})^{-w^{i-1}a_0+1} \right. \\ \left. \times \left(\frac{e^{-z_i}}{1 + e^{-z_i}} \right)^{(1-w)w^{i-1}a_0-1} \frac{e^{-z_i}}{(1 + e^{-z_i})^2} \right] \quad (6.46)$$

Considere novamente o método de aproximação INLA.

A distribuição condicional completa de \mathbf{z} é dada por:

$$p(\mathbf{z} | \boldsymbol{\theta}, \mathbf{y}) \propto \exp\{\log(p(\mathbf{z} | \boldsymbol{\theta})) + \sum_i \log(p(y_i | z_{1:i}, \boldsymbol{\theta}))\} \quad (6.47)$$

A aproximação gaussiana de $\log p(\mathbf{z} | \boldsymbol{\theta})$ é dada por:

$$\log p(\mathbf{z} | \boldsymbol{\theta}) \approx -\frac{1}{2} \mathbf{z}' Q(\mathbf{z}^{(0)}) \mathbf{z} + b^*(\mathbf{z}^{(0)})' \mathbf{z} \quad (6.48)$$

em que

$$Q(\mathbf{z}^{(0)}) = -H_f(\mathbf{z}^{(0)}) \quad (6.49)$$

e

$$b^*(\mathbf{z}^{(0)}) = \nabla_f(\mathbf{z}^{(0)}) - H_f(\mathbf{z}^{(0)}) \mathbf{z}^{(0)} \quad (6.50)$$

denotando $f = \log p(\mathbf{z} | \boldsymbol{\theta})$.

As derivadas necessárias para obter $Q(\mathbf{z}^{(0)})$ e $b^*(\mathbf{z}^{(0)})$ são dadas por

$$\frac{\partial \log(p(\mathbf{z} | \boldsymbol{\theta}))}{\partial z_1} = wa_0 - wb_0e^{z_1}$$

$$\frac{\partial \log(p(\mathbf{z} | \boldsymbol{\theta}))}{\partial z_i} = -(w^{i-1}a_0) \frac{e^{-z_i}}{1 + e^{-z_i}} - ((1-w)w^{i-1}a_0),$$

para $i = 2, \dots, n$

$$\frac{\partial^2 \log(p(\mathbf{z} | \boldsymbol{\theta}))}{\partial z_1^2} = -wb_0 e^{z_1}$$

$$\frac{\partial^2 \log(p(\mathbf{z} | \boldsymbol{\theta}))}{\partial z_i^2} = (-w^{i-1} a_0) \left[\frac{e^{-z_i}}{1 + e^{-z_i}} - \frac{e^{-2z_i}}{(1 + e^{-z_i})^2} \right],$$

para $i = 2, \dots, n$

Para obter a aproximação gaussiana de $\log(p(\mathbf{y} | \mathbf{z}, \boldsymbol{\theta}))$, observe que este caso é um caso particular do descrito na Seção 6.4, em que $I_i = \{z_1, \dots, z_i\}$, $i = 1, \dots, n$, defina $g_i(z_1, \dots, z_i)$ como sendo $g_i(z_1, \dots, z_i) = \log(p(y_i | z_{1:i}, \boldsymbol{\theta}))$, em que por simplificação, (z_1, \dots, z_i) foi denotado por $z_{1:i}$.

Assim sendo, $\log(p(\mathbf{y} | \mathbf{z}, \boldsymbol{\theta})) = \sum_{i=1}^n g_i(z_{1:i})$.

A seguir, descreve-se como obter a aproximação gaussiana de $\log(p(\mathbf{y} | \mathbf{z}, \boldsymbol{\theta}))$.

Através da aproximação de Taylor de segunda ordem, obtém-se

$$g_i(z_{1:i}) = g_i(z_{1:i}^{(0)}) + (z_{1:i} - z_{1:i}^{(0)})' \nabla_{g_i} + \frac{1}{2} (z_{1:i} - z_{1:i}^{(0)})' H_{g_i} (z_{1:i} - z_{1:i}^{(0)}) \quad (6.51a)$$

$$= cte + b'_i(z_{1:i}^{(0)}) z_{1:i} - \frac{1}{2} z_{1:i}' C_i(z_{1:i}^{(0)}) z_{1:i}, \quad (6.51b)$$

em que

$$b_i(z_{1:i}^{(0)}) = \nabla_{g_i}(z_{1:i}^{(0)}) - H_{g_i}(z_{1:i}^{(0)}) z_{1:i}^{(0)} \quad (6.52)$$

e

$$C_i(z_{1:i}^{(0)}) = -H_{g_i}(z_{1:i}^{(0)}). \quad (6.53)$$

Com isso, a aproximação para $\sum_{i=1}^n g_i(z_{1:i})$ é dada por

$$\sum_{i=1}^n g_i(z_{1:i}) \approx \sum_{i=1}^n \left[b'_i(z_{1:i}^{(0)}) z_{1:i} - \frac{1}{2} z_{1:i}' C_i(z_{1:i}^{(0)}) z_{1:i} \right] = -\frac{1}{2} z_{1:n}' (C^+) z_{1:n} + (b^+)' z_{1:n}. \quad (6.54)$$

Esta aproximação resulta em uma distribuição normal com média $(C^+)^{-1}b^+$ e matriz de precisão C^+ , em que

$$C^+ = \sum_{i=1}^n \begin{bmatrix} C_i & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}_{n \times n} \quad (6.55)$$

e

$$b^+ = \sum_{i=1}^n \begin{bmatrix} b_i \\ \mathbf{0} \end{bmatrix}_{n \times 1}, \quad (6.56)$$

sendo $\mathbf{0}$ um vetor de zeros com dimensão apropriada.

Cabe ressaltar que a matriz de precisão da aproximação gaussiana da distribuição condicional completa de \mathbf{z} é uma matriz densa. Sendo assim, o método de aproximação INLA pode não se mostrar tão eficiente do ponto de vista computacional quanto no caso em que é necessário lidar com matrizes esparsas.

Pode-se verificar que a transformação feita no campo latente transformou o problema em algo mais complexo e de implementação computacional menos eficiente. Por outro lado, esta transformação facilita a execução dos algoritmos de otimização.

Tudo que foi visto corresponde ao necessário para implementar o método INLA no caso não-gaussiano, sob uma abordagem proposta. O intuito é se fazer uma verificação da qualidade da aproximação sugerida empiricamente através de dados artificiais e dados reais usando o conjunto de dados da mortalidade de dengue descrito no Capítulo 4.4. Entretanto, problemas computacionais impediram com que o método fosse exemplificado nesta dissertação. Como trabalho futuro, tem-se a exemplificação do método e a verificação empírica da qualidade da aproximação.

Capítulo 7

Considerações Finais

Esta dissertação iniciou com um capítulo com conceitos preliminares de métodos numéricos. Em seguida, foram descritos algumas formas de se realizar inferência bayesiana de forma aproximada.

Uma aplicação de métodos de Monte Carlo via cadeias de Markov a dados reais de mortalidade dengue na região metropolitana do Rio de Janeiro foi apresentada. O objetivo foi verificar se houve um excesso de mortalidade nos períodos epidêmicos de dengue. Não foi verificado tal excesso de mortalidade.

Dentre diversos métodos aproximados de inferência citados, esta dissertação focou no método de aproximação INLA para uma classe específica de modelos, os modelos de espaço de estados. Este método foi descrito de forma detalhada no Capítulo 5 da forma que foi proposto por [Rue et al. \(2009\)](#).

Uma ilustração de aplicação do método INLA em um modelo Poisson dinâmico com dados artificiais foi apresentada. O método INLA se mostrou eficiente do ponto de vista computacional e o motivo disto é que método usufrui de vantagens que a estrutura markoviana traz nesta classe de modelos latentes gaussianos. Entretanto, cabe ressaltar que o método de aproximação INLA tem a desvantagem de ter seu custo computacional crescente à medida que cresce a dimensão paramétrica.

O Capítulo 6 tratou de uma proposta para a extensão do método INLA do capítulo anterior, a fim de possibilitar sua aplicação em modelos não-gaussianos e em modelos com uma estrutura hierárquica mais geral do que a inicialmente proposta por [Rue et al. \(2009\)](#).

Foi apresentado um exemplo ilustrativo teórico de como a abordagem deve ser executada.

Futuramente, há um interesse em se verificar empiricamente se a modelagem proposta no Capítulo 6 se mostra eficiente e precisa, tal como o resultado encontrado para o exemplo do modelo Poisson descrito no Capítulo 5.

O método proposto por [Rue et al. \(2009\)](#) trouxe de volta a atenção para o desenvolvimento de metodologias estatísticas determinísticas para a obtenção de distribuições *a posteriori* marginais. Isto pode ser considerado um ponto de partida para trabalhos posteriores. Sobre o método INLA, por exemplo, é de extrema importância que mais pesquisa seja feita a fim de se saber mais sobre a acurácia de sua aproximação.

Pode-se perceber que o método INLA é de difícil implementação, requer cálculos analíticos trabalhosos ou a utilização de diferenciação numérica, que não se pode confiar inteiramente, além de otimizações numéricas complexas. O tempo de análise do método por parte do usuário, em geral, é consideravelmente alto. Contudo, a implementação computacional de alguns modelos latentes gaussianos já está disponível no pacote R-INLA ([Rue e Martino, 2009](#)) que deve ser instalado no software R ([R Development Core Team, 2010](#)).

O tempo de análise necessário para a implementação do INLA contrasta com a rápida execução computacional em alguns casos. Casos estes em que se assume independência condicional nos elementos do campo latente.

Cabe ressaltar, entretanto, que essa suposição é de certa forma restritiva. Violações desta suposição fazem com que o método INLA perca em eficiência, uma vez que ela é crucial para que seja possível acelerar o tempo computacional, em que se tira proveito da esparsidade de matrizes envolvidas.

Como trabalhos futuros, do ponto de vista teórico, pode-se considerar adaptações na metodologia do INLA para viabilizar sua aplicação em modelos não-lineares.

Do ponto de vista prático, há uma vasta possibilidade de utilização do método em problemas reais cuja modelagem considerada possui a mesma estrutura hierárquica dos modelos descritos nesta dissertação.

Referências Bibliográficas

- Azzalini, A. e Capitanio, A. (1999) Statistical applications for the multivariate skew normal distribution. *Journal of the Royal Statistical Society Series B*, **61**, 579–602.
- Beal, M. J. (2003) *Variational Algorithms for Approximate Bayesian Inference*. Tese de Doutorado, University College London, Londres.
- Bernardo, J. M., Bayarri, M. J., Berger, J. O., Dawid, A. P., Heckerman, D., Smith, A. F. M., West, M. e., Beal, M. J. e Ghahramani, Z. (2003) The variational bayesian em algorithm for incomplete data: with application to scoring graphical model structures. *Bayesian Statistics 7*, Oxford University Press.
- Bishop, C. M. (2006) *Pattern Recognition and Machine Learning*. New York: Springer.
- Boyd, S. P. e Vandenberghe, L. (2004) *Convex optimization*. Cambridge University Press.
- de Bruijn, N. G. (1961) *Asymptotic Methods in Analysis*. Amsterdam: North-Holland.
- Burden, R. L. e Faires, J. D. (2001) *Numerical analysis*. Brooks/Cole Publishing Company.
- Cheng, S. H. e Higham, N. J. (1998) A modified cholesky algorithm based on a symmetric indefinite factorization. *SIAM Journal On Matrix Analysis And Applications*, **19**.
- Dempster, A. P., Laird, N. M. e Rubin, D. B. (1977) Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society Series B*, **39**.
- Durbin, J. e Koopman, S. J. (2000) Time series analysis of non-gaussian observations based on state space models from both classical and bayesian perspectives. *Journal of the Royal Statistical Society Series B*, **62(1)**, 3–56.

- Fernandes, M. V. M. (2006) *Modelos para Processos Espaços-Temporais Inflacionados de Zeros*. Dissertação de Mestrado, Universidade Federal do Rio de Janeiro, Rio de Janeiro.
- Ferreira, G. S. (2004) *Análise Espaço-Temporal da Distribuição dos Casos de Dengue na Cidade do Rio de Janeiro no Período de 1986 a 2002*. Dissertação de Mestrado, Universidade Federal do Rio de Janeiro, Rio de Janeiro.
- Franco, G. C., Gamerman, D. e Santos, T. R. (2009) *Modelos de Espaços de Estados: Abordagens Clássica e Bayesiana*. Associação Brasileira de Estatística.
- Gamerman, D. e Lopes, H. (2006) *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*. New York: Chapman & Hall / CRC.
- Gelfand, A. E. e Smith, A. F. M. (1990) Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, **85**, 398–409.
- Geman, S. e Geman, D. (1984) Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **6**, 721–741.
- Harvey, A. C. e Fernandes, C. (1989) A non-gaussian state space model and application to prediction of records. *Journal of Business and Economic Statistics*, **7**, 407–417.
- Hastings, H. (1970) Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, **57**, 97–109.
- Higham, N. J. (2002) Computing the nearest correlation matrix - a problem from finance. *IMA Journal of Numerical Analysis*, **22**, 329–343.
- Kitagawa, G. (1987) Non-gaussian state-space modeling of nonstationary time series. *Journal of the American Statistical Association*, **82**, 400, 1032–1041.
- Knol, D. e ten Berge, J. (1989) Least-squares approximation of an improper correlation matrix by a proper one. *Psychometrika*, **54**, 53–61.
- Kullback, S. e Leibler, R. A. (1951) On information and sufficiency. *The Annals of Mathematical Statistics*, **22**.

- Malhao, T. A. (2010) *Sobremortalidade durante epidemia de dengue. Região Metropolitana do Rio de Janeiro, 2007-2008*. Dissertação de Mestrado, Universidade Federal do Rio de Janeiro, Rio de Janeiro.
- Martins, T. G. (2010) *Aproximações Determinísticas para Distribuições a Posteriori Marginais*. Dissertação de Mestrado, Universidade Federal do Rio de Janeiro, Rio de Janeiro.
- Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A. e Teller, E. (1953) Equation of state calculations by fast computing machine. *Journal of Chemical Physics*, **21**, 1087–1091.
- Migon, H. e Gamerman, D. (1999) *Statistical Inference: an Integrated Approach*. London: Arnold.
- Minka, T. P. (2001) *A family of algorithms for approximate Bayesian inference*. Tese de Doutorado, Massachusetts Institute of Technology, Cambridge.
- Naylor, J. C. e Smith, A. F. M. (1982) Applications of a method for the efficient computation of posterior distributions. *Journal of the Royal Statistical Society Series C*, **31**, 214–225.
- Nelder, J. A. e Wedderburn, R. W. M. (1972) Generalized linear models. *Journal of the Royal Statistical Society A*, **135**, 370–384.
- Nocedal, J. e Wright, S. J. (2006) *Numerical Optimization*. Springer.
- Pereira, J. B. M. (2010) *Modelos para Dados de Contagem com Estrutura Temporal*. Dissertação de Mestrado, Universidade Federal do Rio de Janeiro.
- R Development Core Team (2010) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL <http://www.R-project.org/>. ISBN 3-900051-07-0.
- Rue, H. e Held, L. (2005) *Gaussian Markov Random Fields: Theory and Applications*. London: Chapman and Hall/CRC Press.
- Rue, H. e Martino, S. (2007) Approximate bayesian inference for hierarchical gaussian markov random field models. *Journal of statistical planning and inference*, **137(10)**, 3177–3192.

- (2009) *INLA: Functions which allow to perform a full Bayesian analysis of structured additive models using Integrated Nested Laplace Approximation*. R package version 0.0.
- Rue, H., Martino, S. e Chopin, N. (2009) Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society Series B*, **71**, 319–392.
- Ruggiero, M. A. G. e Lopes, V. L. R. (1996) *Cálculo Numérico: Aspectos Teóricos e Computacionais*. MAKRON Books.
- Ruiz-Cárdenas, R., Krainski, E. T. e Rue, H. (2010) Fitting dynamic models using integrated nested laplace approximations - inla. *Preprint Statistics No 12/2010, Department of Mathematical Sciences, Norwegian University of Science and Technology, Trondheim, Norway*. URL http://www.math.ntnu.no/inla/r-inla.org/papers/Ruiz_Krainski_Rue_TR_23112010.pdf.
- Santos, T. R. (2009) *Inferência sobre os Hiperparâmetros em Modelos Estruturais sob as Perspectivas Clássica e Bayesiana*. Dissertação de Mestrado, Universidade Federal de Minas Gerais, Belo Horizonte.
- Smith, A. F. M., Skene, A. M., Shaw, J. E. H. e Naylor, J. C. (1987) Progress with numerical and graphical methods for practical bayesian statistics. *Journal of the Royal Statistical Society. Series D (The Statistician)*, **36**, 75–82.
- Smith, R. L. e Miller, J. E. (1986) A non-gaussian state space model and application to prediction of records. *Journal of the Royal Statistical Society B*, **48**, 79–88.
- Tierney, L. e Kadane, J. B. (1986) Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistics Association*, **81**, 82–86.
- Tierney, L., Kass, R. E. e Kadane, J. B. (1989) Approximate marginal densities of nonlinear functions. *Biometrika*, **76**, 425–433.
- West, M. e Harrison, J. (1997) *Bayesian Forecasting and Dynamic Models*. New York: Springer-Verlag, 2a edn.

West, M., Harrison, J. e Migon, H. S. (1985) Dynamic generalized linear models and bayesian forecasting. *Journal of the Americal Statistical Association*, **80**, 387, 73–97.