Análise bayesiana de densidades aleatórias simples

Paulo Cilas Marques Filho

Tese apresentada
ao
Instituto de Matemática e Estatística
da
Universidade de São Paulo
para
obtenção do título
de
Doutor em Ciências

Programa: Estatística

Orientador: Prof. Dr. Carlos Alberto de Bragança Pereira

São Paulo, Dezembro de 2011

Análise bayesiana de densidades aleatórias simples

Este exemplar corresponde à redação final da tese devidamente corrigida e defendida por Paulo Cilas Marques Filho e aprovada pela Comissão Julgadora em 19/12/2011.

Comissão Julgadora:

Prof. Dr. Carlos Alberto de Bragança Pereira (orientador)	IME – USP
Prof. Dr. Sergio Wechsler	IME – USP
Prof. Dr. Antonio Luiz Pereira	IME – USP
Prof. Dr. Marcelo de Souza Lauretto	EACH – USP
Prof. Dr. Adriano Polpo de Campos	UFSCar

Análise Bayesiana de Densidades Aleatórias Simples

Paulo Cilas Marques Filho

São Paulo, Dezembro de 2011

à Luciana e ao Lucas

Resumo

Definimos, a partir de uma partição de um intervalo limitado da reta real formada por subintervalos, uma distribuição *a priori* sobre uma classe de densidades em relação à medida de Lebesgue construindo uma densidade aleatória cujas realizações são funções simples não negativas que assumem um valor constante em cada subintervalo da partição e possuem integral unitária. Utilizamos tais *densidades aleatórias simples* na análise bayesiana de um conjunto de observáveis absolutamente contínuos e provamos que a distribuição *a priori* é fechada sob amostragem. Exploramos as distribuições *a priori* e *a posteriori* via simulações estocásticas e obtemos soluções bayesianas para o problema de estimação de densidade. Os resultados das simulações exibem o comportamento assintótico da distribuição *a posteriori* quando crescemos o tamanho das amostras dos dados analisados. Quando a partição não é conhecida *a priori*, propomos um critério de escolha a partir da informação contida na amostra. Apesar de a esperança de uma densidade aleatória simples ser sempre uma densidade descontínua, obtemos estimativas suaves resolvendo um problema de decisão em que os estados da natureza são realizações da densidade aleatória simples e as ações são densidades suaves de uma classe adequada.

Abstract

We define, from a known partition in subintervals of a bounded interval of the real line, a prior distribution over a class of densities with respect to Lebesgue measure constructing a random density whose realizations are nonnegative simple functions that integrate to one and have a constant value on each subinterval of the partition. These *simple random densities* are used in the Bayesian analysis of a set of absolutely continuous observables and the prior distribution is proved to be closed under sampling. We explore the prior and posterior distributions through stochastic simulations and find Bayesian solutions to the problem of density estimation. Simulations results show the asymptotic behavior of the posterior distribution as we increase the size of the analyzed data samples. When the partition is unknown, we propose a choice criterion based on the information contained in the sample. In spite of the fact that the expectation of a simple random density is always a discontinuous density, we get smooth estimates solving a decision problem where the states of nature are realizations of the simple random density and the actions are smooth densities of a suitable class.

Agradecimentos

Durante todos os anos deste doutorado, meu orientador, Professor Carlinhos, manteve um entusiasmo incondicional, acreditando nas ideias da tese e dando todo o apoio acadêmico necessário, além de ter me ensinado muito sobre Estatística em geral. Deixo aqui meus mais sinceros agradecimentos ao Professor e espero que este trabalho, de algum modo, represente um pouco o seu espírito de que não devemos complicar o que deve ser simples.

O Professor Sergio Wechsler, com seu conhecimento enciclopédico sobre inferência, tem sido uma fonte constante de sabedoria sobre "todas as coisas bayesianas". Em particular, a discussão sobre permutabilidade no início do segundo capítulo é fruto de uma generosa e profunda explicação que me foi dada pelo Professor no segundo ano do doutorado.

Assim que entrei no instituto, tive a boa sorte de assistir às preciosas aulas do Professor Luis Gustavo Esteves. A marca indelével deixada por estas aulas em minha memória compõe boa parte do que posso chamar de minha formação acadêmica em Estatística.

Meu amigo, Dr. Luiz Eugênio Barboza de Oliveira, com sua sabedoria abissal sobre tantas coisas desta vida, é um dos responsáveis pela existência desta tese. Espero poder retribuir a contento sua amizade nos muitos anos vindouros.

Meus pais e irmãs tornaram este trabalho possível, apoiando desde sempre minhas escolhas e intenções. Meu caráter eu devo principalmente a vocês. Obrigado por serem quem vocês são.

Finalmente, só posso agradecer aos meus dois amores: minha esposa Luciana e meu filho Lucas. Foi Luciana que, me vendo fazer inferência bayesiana em meu trabalho durante muitos anos, descobriu que, sim, havia bayesianos na USP e sugeriu que eu trilhasse este caminho. Para o Lucas, que, por tanto tempo, interessadíssimo, me observou rabiscar símbolos estranhos no papel, quero dizer que o "nosso livro" ficou pronto. É com o mais puro prazer que eu dedico esta tese a vocês dois.

Conteúdo

Introdução

Até o início da década de setenta do século passado, a Estatística bayesiana [8] se concentrava na análise de modelos estatísticos que compreendem famílias de distribuições de probabilidade de dimensão finita. O trabalho de Ferguson [9], que introduziu o Processo Dirichlet, foi uma das primeiras extensões para a situação não paramétrica, na qual é construída uma distribuição *a priori* sobre o espaço de dimensão, em geral, infinita formado por todas as distribuições de probabilidade sobre um determinado espaço amostral.

Uma construção como a do Processo Dirichlet busca o equilíbrio entre o tamanho do suporte da distribuição *a priori* e a tratabilidade do processo de inferência. Apesar de suas propriedades analíticas notáveis, tais como o fechamento sob amostragem e a existência de uma expressão simples para a esperança da medida de probabilidade aleatória, foi demonstrado por Blackwell [6] que, com probabilidade um, as realizações de um Processo Dirichlet são distribuições de probabilidade discretas. Para tratar as situações em que esta descrição não é adequada, diversas modificações e extensões do Processo Dirichlet foram propostas na literatura [10].

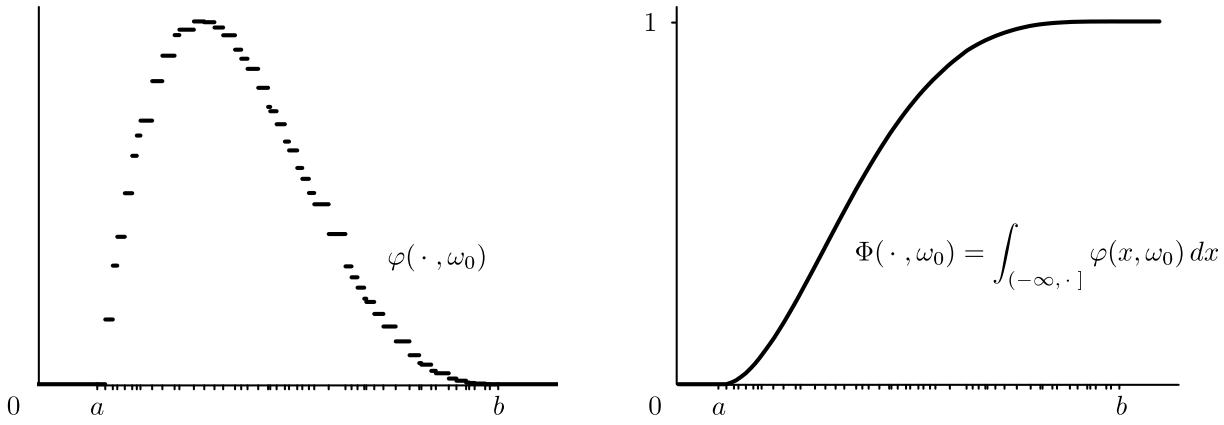
O objetivo desta tese é construir uma distribuição *a priori* sobre uma classe de densidades em relação à medida de Lebesgue e explorar as consequências inferenciais desta construção no tratamento de casos que envolvem observáveis absolutamente contínuos.

Dada uma partição de um intervalo limitado da reta real formada por subintervalos, nossa distribuição *a priori* está concentrada na classe das densidades simples em relação a esta partição, que são as funções não negativas, constantes em cada subintervalo da partição, que possuem integral unitária em relação à medida de Lebesgue. Cada densidade simples é determinada pelo vetor formado por seus valores em cada subintervalo da partição. Dizemos que as componentes deste vetor são as alturas dos degraus da densidade simples.

A ideia é definir uma *densidade aleatória simples* especificando a distribuição das alturas aleatórias dos seus degraus através da transformação e do condicionamento da distribuição normal multivariada, de maneira similar ao que foi feito nos processos estocásticos desenvolvidos por Thorburn e Lenk [15, 12].

A principal diferença entre nossa proposta e estes trabalhos é que aqui trabalhamos com um objeto aleatório de dimensão finita, o que nos leva a uma teoria muito mais simples, que prescinde de aproximações de dimensão finita para representar os objetos computacionalmente, e que não torna imperativo o uso de interpolações *ad hoc* das soluções obtidas via simulação.

Uma vez que o número de subintervalos da partição considerada determina a dimensão da família de distribuições de probabilidade, e que este número pode ser tão grande quanto se queira, nossa construção é semiparamétrica. A figura abaixo apresenta uma realização $\varphi(\,\cdot\,,\omega_0)$ de uma densidade aleatória simples e a função de distribuição $\Phi(\,\cdot\,,\omega_0)$ correspondente a esta realização.



Vemos que tais realizações são densidades de um observável absolutamente contínuo que assume, com probabilidade um, valores no intervalo limitado considerado.

A tese está organizada da seguinte maneira. No capítulo 1, definimos uma densidade aleatória simples e obtemos a expressão de uma densidade condicional, em relação a uma medida dominante adequada, do vetor aleatório formado pelas alturas dos seus degraus. No capítulo 2, utilizamos uma densidade aleatória simples para modelar uma sequência de observáveis permutáveis e provamos que a distribuição *a priori* é fechada sob amostragem. No capítulo 3, exploramos as distribuições *a priori* e *a posteriori* em alguns exemplos através de simulações estocásticas, obtendo soluções bayesianas para o problema de estimação de densidade. Os resultados destas simulações exibem o comportamento assintótico da distribuição *a posteriori* quando aumentamos o tamanho das amostras dos dados analisados. Posteriormente, consideramos o caso em que a própria partição utilizada na definição da densidade aleatória simples é desconhecida, e propomos um critério para a escolha da mesma a partir da informação contida na amostra. Apesar de a esperança de uma densidade aleatória simples ser sempre uma densidade descontínua, ao final do capítulo, sem alterar nossa distribuição *a priori*, obtemos estimativas suaves resolvendo um problema de decisão em que os estados da natureza são realizações da densidade aleatória simples e as ações são densidades suaves de uma classe adequada. Concluímos indicando algumas possíveis extensões da teoria proposta.

Dentro de cada capítulo, os exemplos, definições, proposições, lemas e teoremas são identificados pelo número do capítulo seguido por um número inteiro cuja sequência é compartilhada por todos os itens. O final das demonstrações e dos exemplos é indicado pelo símbolo ❖.

Capítulo 1

Densidades aleatórias simples

Dada uma partição de um intervalo limitado da reta real formada por subintervalos, o propósito deste capítulo é definir e caracterizar uma classe de objetos denominados densidades aleatórias simples, cujas realizações são funções simples não negativas que assumem um valor constante em cada subintervalo da partição e possuem integral unitária. Tais realizações são densidades de uma variável aleatória absolutamente contínua. Portanto, uma densidade aleatória simples determina uma distribuição *a priori* na classe das densidades para as quais o intervalo limitado especificado é o suporte da medida de probabilidade correspondente.

O capítulo está organizado da seguinte maneira. Na seção 1.1, definimos a classe das densidades simples em relação a uma partição formada por subintervalos de um intervalo limitado da reta real. Uma densidade simples é caracterizada pelos valores constantes que esta assume em cada subintervalo da partição; dizemos que estes valores são as alturas dos degraus da densidade simples. Por conseguinte, definir um objeto aleatório cujas realizações são densidades simples consiste em especificar a distribuição do vetor das alturas aleatórias, e isto é feito através do condicionamento da distribuição lognormal, definida e caracterizada no restante da seção. No início da seção 1.2, definimos informalmente uma densidade aleatória simples e observamos que esta definição envolve o condicionamento em um evento de probabilidade zero. Para tratar este caso, utilizamos a definição geral de distribuição condicional, que nos leva naturalmente ao conceito de densidade condicional. Mostramos que, em nosso caso, o cálculo de uma densidade condicional difere da situação usual em que a distribuição conjunta é dominada por uma medida produto. Isto nos leva, na seção 1.3, à construção de uma família de medidas dominantes adequada à nossa situação. De posse destes resultados, encerramos o capítulo na seção 1.4 com a definição formal de uma densidade aleatória simples.

1.1 Densidades simples

Ao longo de toda a tese, denotamos por $(\Omega, \mathscr{F}, P)$ o espaço de probabilidade subjacente, a partir do qual induzimos as distribuições de todos os objetos aleatórios considerados. Damos uma interpretação subjetivista às probabilidades calculadas com a medida P. Tais probabilidades são a expressão numérica da incerteza de um sujeito do conhecimento a respeito dos eventos

em $\mathscr{F}$. Construções axiomáticas do cálculo de probabilidades de acordo com esta interpretação podem ser encontradas em [8] e [14].

Usamos as seguintes notações. Para um inteiro $k \geq 1$, denotamos por $\mathbb{R}^k_+$ o conjunto dos vetores de $\mathbb{R}^k$ cujas componentes são todas positivas. Escrevemos $\mathscr{R}^k$ para a sigma-álgebra de Borel de $\mathbb{R}^k$. Denotamos por λ_k a medida de Lebesgue sobre $(\mathbb{R}^k, \mathscr{R}^k)$. Suprimimos os índices quando $k = 1$. As componentes de um vetor $v \in \mathbb{R}^k$ são denotadas por $v_1, \ldots, v_k$.

Daqui em diante, dado um intervalo $[a, b]$ da reta real, defina $\Delta = \{t_0, t_1, \ldots, t_k\}$, com $a = t_0 < t_1 < \cdots < t_k = b$. Seja $d_i = t_i - t_{i-1}$, para $i = 1, \ldots, k$, e defina $S_\Delta : \mathbb{R}^k \to \mathbb{R}$ por $S_\Delta(u) = \sum_{i=1}^k d_i u_i$.

Considere a partição de $[a, b]$ nos subintervalos $[a, t_1), [t_1, t_2), \ldots, [t_{k-2}, t_{k-1}), [t_{k-1}, b]$. A classe das densidades simples em relação a esta partição é formada pelas funções simples não negativas que assumem um valor constante em cada subintervalo da partição e que possuem integral unitária em relação à medida de Lebesgue. Representamos cada densidade simples f com o auxílio de um vetor $h \in \mathbb{R}^k$ tal que $h_i \geq 0$, para $i = 1, \ldots, k$, em que as h_i's são denominadas alturas dos degraus de f. De maneira explícita, uma densidade simples f admite a representação

$$f(x) = \sum_{i=1}^{k} h_i \, I_{[t_{i-1}, t_i)}(x) \, ,$$

na qual I_A é a função indicadora do conjunto A: $I_A(x) = 0$, se $x \notin A$, e $I_A(x) = 1$, se $x \in A$. A condição de unitariedade da integral de f em relação à medida de Lebesgue impõe que $S_\Delta(h) = 1$.

Daqui em diante, nos referimos apenas à classe das densidades simples, deixando subentendida a partição subjacente, que supomos ser conhecida *a priori*. Na seção 3.4, consideramos uma extensão do modelo para o caso em que a partição não é conhecida.

Para definir uma densidade aleatória cujas realizações são densidades simples, tudo o que precisamos é especificar a distribuição de um vetor aleatório que determine as alturas dos seus degraus. A ideia é usar para tanto a distribuição de um vetor aleatório lognormal U condicionada no fato de que $S_\Delta(U) = 1$. O restante desta seção estabelece algumas propriedades da distribuição lognormal. Todas as definições e resultados necessários a respeito da distribuição normal podem ser encontrados, por exemplo, em [1].

Definição 1.1 Seja $Z = (Z_1, \ldots, Z_k)$ um vetor aleatório com distribuição normal que possui vetor de médias m e matriz de covariâncias Σ. Dizemos que o vetor aleatório $U : \Omega \to \mathbb{R}^k$ definido por $U(\omega) = \left(e^{Z_1(\omega)}, \ldots, e^{Z_k(\omega)}\right)$ tem *distribuição lognormal*, com parâmetros m e Σ. Denotamos a distribuição de U por $\mu_U(A) = P\{\omega : U(\omega) \in A\}$, para $A \in \mathscr{R}^k$. Usamos a notação $U \sim \ L_k(m, \Sigma)$.

Antes de caracterizarmos a distribuição de um vetor lognormal, precisamos recordar alguns dos conceitos relacionados à definição de continuidade absoluta.

Sejam μ e ν duas medidas sobre um espaço mensurável $(S, \mathscr{S})$. A medida μ é absolutamente contínua em relação à medida ν se $\nu(A) = 0$ implica em $\mu(A) = 0$, para todo $A \in \mathscr{S}$. Ou seja, μ não tem efeitos nos conjuntos nulos de ν. Neste caso, dizemos que ν é uma medida dominante de μ, ou que μ é dominada por ν, e usamos a notação $\mu \ll \nu$.

Pelo Teorema de Radon-Nikodym [14], se $\mu \ll \nu$ e ν é sigma-finita, então existe uma função real estendida mensurável f tal que

$$\mu(A) = \int_A f(s)\, d\nu(s)\,,$$

para todo $A \in \mathscr{S}$. Dizemos que f é uma versão da densidade de μ em relação à medida ν. O termo "versão" enfatiza que, pela definição da integral de Lebesgue, qualquer função real estendida mensurável f^* tal que $f = f^*$ quase certamente $[\nu]$ também satisfaz a relação integral acima. A classe de equivalência formada por estas densidades é a derivada de Radon-Nikodym de μ em relação a ν, que denotamos por $d\mu/d\nu$. Em geral, indicamos um representante da classe de equivalência escrevendo abreviadamente $d\mu/d\nu = f$. O Teorema de Radon-Nikodym também garante que se g é um função real estendida mensurável integrável em relação a μ, então

$$\int g(s)\, d\mu(s) = \int g(s) f(s)\, d\nu(s) = \int g(s)\, \frac{d\mu}{d\nu}(s)\, d\nu(s)\,,$$

em que cometemos um ligeiro abuso de notação ao escrever o integrando da terceira integral. Este resultado é conhecido como a regra de Leibniz para as derivadas de Radon-Nikodym, uma vez que, formalmente, podemos escrever $d\mu = \dfrac{d\mu}{d\nu}\, d\nu$.

O seguinte resultado fornece a expressão analítica para uma densidade de um vetor aleatório com distribuição lognormal. O símbolo $^\top$ indica a operação de transposição de vetores e matrizes.

Proposição 1.2 *Seja $U \sim L_k(m, \Sigma)$. Suponha que Σ é não-singular. Então, $\mu_U \ll \lambda_k$, com derivada de Radon-Nikodym $d\mu_U/d\lambda_k = f_U$ dada por*

$$f_U(u) = (2\pi)^{-k/2}\, |\Sigma|^{-1/2} \left(\prod_{i=1}^{k} u_i^{-1} \right) \exp\left(-\frac{1}{2} (\log u - m)^\top \Sigma^{-1} (\log u - m) \right) I_{\mathbb{R}_+^k}(u)\,,$$

em que $|\Sigma|$ é o determinante da matriz Σ, $\log u = (\log u_1, \ldots, \log u_k)^\top$ e $m = (m_1, \ldots, m_k)^\top$.

Demonstração. Seja $f : \mathbb{R}^k \to \mathbb{R}^k_+$ a função definida por $f(z_1, \ldots, z_k) = (e^{z_1}, \ldots, e^{z_k})$. Assim, f é uma função diferenciável, com inversa diferenciável $g : \mathbb{R}^k_+ \to \mathbb{R}^k$ definida por $g(u_1, \ldots, u_k) = (\log u_1, \ldots, \log u_k)$. O valor do jacobiano no ponto $u \in \mathbb{R}^k_+$ é $J_g(u) = \prod_{i=1}^k u_i^{-1}$. Por definição, $U = f(Z)$, em que Z possui distribuição normal com vetor de médias m e matriz de covariâncias Σ. Uma vez que Σ é não-singular, sabemos que f_Z definida por

$$f_Z(z) = (2\pi)^{-k/2} \, |\Sigma|^{-1/2} \exp\left(-\frac{1}{2}(z-m)^\top \Sigma^{-1} \, (z-m)\right) I_{\mathbb{R}^k}(z) \, ,$$

com $z = (z_1, \ldots, z_k)^\top$ e $m = (m_1, \ldots, m_k)^\top$, é uma densidade da distribuição de Z em relação a λ_k, . Assim, fazendo uma mudança de variável, temos que $f_U(u) = f_Z(g(u))|J_g(u)|$, para $u \in \mathbb{R}^k_+$. Uma vez que $P\{\omega : U(\omega) \notin \mathbb{R}^k_+\} = 0$, definimos $f_U(u) = 0$, para $u \notin \mathbb{R}^k_+$. O resultado segue. ❖

Existem expressões analíticas simples para as esperanças e covariâncias de um vetor aleatório com distribuição lognormal.

Proposição 1.3 *Seja $U \sim L_k(m, \Sigma)$, com $\Sigma = (\sigma_{ij})$. Então, temos que $\mathrm{E}[U_i] = e^{m_i + \frac{1}{2}\sigma_{ii}}$ e $\mathrm{Cov}[U_i, U_j] = \mathrm{E}[U_i]\,\mathrm{E}[U_j](e^{\sigma_{ij}} - 1)$, para $i, j = 1, \ldots, k$.*

Demonstração. Seja Z um vetor aleatório com distribuição normal multivariada, com vetor de médias m e matriz de covariâncias Σ. Por definição, $U_i(\omega) = e^{Z_i(\omega)}$, para $i = 1, \ldots, k$. Para $a \in \mathbb{R}^k$, temos que

$$\mathrm{E}\left[\prod_{i=1}^k U_i^{a_i}\right] = \mathrm{E}\left[\prod_{i=1}^k e^{a_i Z_i}\right] = \mathrm{E}\left[e^{a^\top Z}\right] = e^{a^\top m + \frac{1}{2} a^\top \Sigma \, a} \, ,$$

nas quais $a = (a_1, \ldots, a_k)^\top$ e utilizamos a expressão da função geradora de momentos de Z (veja [1]). Para $i = 1, \ldots, k$, escolha $a \in \mathbb{R}^k$ de modo que sua i-ésima componente seja igual a um e as demais sejam iguais a zero. Assim, a expressão acima para $\mathrm{E}\left[\prod_{i=1}^k U_i^{a_i}\right]$ fornece que $\mathrm{E}[U_i] = e^{m_i + \frac{1}{2}\sigma_{ii}}$. Agora, para $i, j = 1, \ldots, k$, tome $a \in \mathbb{R}^k$ tal que suas i-ésima e j-ésima componentes sejam iguais a um e as demais sejam iguais a zero. Usando novamente a expressão para $\mathrm{E}\left[\prod_{i=1}^k U_i^{a_i}\right]$, temos que $\mathrm{E}[U_i \, U_j] = e^{m_i + m_j + \frac{1}{2}\sigma_{ii} + \sigma_{ij} + \frac{1}{2}\sigma_{jj}} = \mathrm{E}[U_i]\,\mathrm{E}[U_j]\, e^{\sigma_{ij}}$. Pela definição de covariância, temos que $\mathrm{Cov}[U_i, U_j] = \mathrm{E}[U_i \, U_j] - \mathrm{E}[U_i]\,\mathrm{E}[U_j] = \mathrm{E}[U_i]\,\mathrm{E}[U_j](e^{\sigma_{ij}} - 1)$. ❖

Os resultados da proposição anterior mostram como controlar o sinal das covariâncias entre as componentes de um vetor lognormal U, pois $\mathrm{Cov}[U_i, U_j] < 0$ se e somente se $\sigma_{ij} < 0$. Observe também que no caso da distribuição normal conseguimos especificar as esperanças separadamente das covariâncias, mas, examinando a expressão obtida para $\mathrm{Cov}[U_i, U_j]$, vemos que o mesmo não ocorre com a distribuição lognormal.

1.2 Distribuições condicionais e suas densidades

Vimos no início da seção anterior como uma densidade simples é caracterizada pelo vetor formado pelas alturas dos seus degraus, o que torna o nosso objetivo a construção de um vetor aleatório H tal que, com probabilidade um, suas componentes sejam não negativas e $S_\Delta(H) = 1$. Informalmente, a ideia é definir a distribuição de H com o auxílio de um vetor aleatório lognormal U através do condicionamento: $P\{H \in A\} = P\{U \in A \mid S_\Delta(U) = 1\}$, para todo $A \in \mathscr{R}^k$. Provaremos no Lema 1.11 que a distribuição de $S_\Delta(U)$ é absolutamente contínua em relação à medida de Lebesgue, o que faz que o evento $\{S_\Delta(U) = 1\}$ tenha probabilidade zero. Assim, precisamos trabalhar com uma definição de distribuição condicional que descreva este caso.

Vamos chegar à definição geral de distribuição condicional examinando e estendendo o seguinte caso particular: sejam X e Y variáveis aleatórias que assumem valores, respectivamente, em $\mathscr{X} = \{x_1, \ldots, x_m\}$ e $\mathscr{Y} = \{y_1, \ldots, y_n\}$. Defina as sigma-álgebras $\mathscr{A} = 2^{\mathscr{X}}$ e $\mathscr{B} = 2^{\mathscr{Y}}$. Suponha que $P\{Y = y_i\} > 0$, para $i = 1, \ldots, n$. Neste caso, a definição usual de probabilidade condicional diz que

$$P\{X \in A \mid Y = y_i\} = \frac{P\{X \in A, Y = y_i\}}{P\{Y = y_i\}}\,,$$

em que $A \in \mathscr{A}$ e $i = i, \ldots, n$. Para $B \in \mathscr{B}$, segue facilmente desta definição que

$$P\{X \in A, Y \in B\} = \sum_{y \in B} P\{X \in A \mid Y = y\}\, P\{Y = y\}\,. \qquad (*)$$

A ideia é usar a propriedade $(*)$ como guia para uma nova definição de $P\{X \in A \mid Y = y_i\}$ que seja válida quando $P\{Y = y_i\} = 0$.

Sejam $(\mathscr{X}, \mathscr{A})$ e $(\mathscr{Y}, \mathscr{B})$ espaços mensuráveis e considere $X : \Omega \to \mathscr{X}$ e $Y : \Omega \to \mathscr{Y}$ mensuráveis nas respectivas sigma-álgebras. Para cada $A \in \mathscr{A}$, defina a medida ν_A sobre $(\mathscr{Y}, \mathscr{B})$ por $\nu_A(B) = P\{X \in A, Y \in B\}$. Pela monotonicidade de P, temos que

$$\nu_A(B) \leq P\{Y \in B\} = \mu_Y(B)\,,$$

na qual $\mu_Y = P \circ Y^{-1}$ é a distribuição de Y. É imediato que $\nu_A \ll \mu_Y$ e, pelo Teorema de Radon-Nikodym, temos que

$$\nu_A(B) = \int_B \frac{d\nu_A}{d\mu_Y}(y)\, d\mu_Y(y)\,,$$

para cada $B \in \mathscr{B}$.

Introduzindo as notações $\mu_{X|Y}(A \mid y) = P\{X \in A \mid Y = y\} = \dfrac{d\nu_A}{d\mu_Y}(y)$ para a derivada de Radon-Nikodym, a relação integral acima pode ser interpretada intuitivamente como uma

versão generalizada da propriedade $(*)$. Informalmente, a probabilidade de observarmos que $Y = y$ é igual a $\mu_Y(dy)$ e, dado que isto ocorreu, a probabilidade de observarmos que $X \in A$ é igual a $P\{X \in A \mid Y = y\}$. Para calcularmos $P\{X \in A, Y \in B\}$, somamos (integramos) $P\{X \in A \mid Y = y\}\mu_Y(dy)$ sobre todos os y em B.

O Teorema de Radon-Nikodym garante que $\mu_{X|Y}(A \mid \cdot\ )$ é mensurável, para cada $A \in \mathscr{A}$. No entanto, não podemos dizer que, em geral, $\mu_{X|Y}(\ \cdot \mid y)$ é uma medida de probabilidade sobre $(\mathscr{X}, \mathscr{A})$, para todo $y \in \mathscr{Y}$. A seguinte definição agrega todos estes aspectos.

Definição 1.4 Seja $\mu_{X|Y} : \mathscr{A} \times \mathscr{Y} \to \mathbb{R}$ e denote o valor de $\mu_{X|Y}$ em (A, y) por $\mu_{X|Y}(A \mid y)$. Se $\mu_{X|Y}$ satisfaz

1. $\mu_{X|Y}(A \mid \cdot\ )$ é mensurável, para todo $A \in \mathscr{A}$;
2. $P\{X \in A, Y \in B\} = \displaystyle\int_B \mu_{X|Y}(A \mid y)\, d\mu_Y(y)$, para todo $A \in \mathscr{A}$ e todo $B \in \mathscr{B}$,

dizemos que $\mu_{X|Y}$ é uma *versão da distribuição condicional de X dado Y*. Ademais, se

3. $\mu_{X|Y}(\ \cdot \mid y)$ é uma medida de probabilidade sobre $(\mathscr{X}, \mathscr{A})$, para todo $y \in \mathscr{Y}$,

então $\mu_{X|Y}$ é uma versão *regular* da distribuição condicional de X dado Y.

A existência de versões regulares é garantida, sob condições bastante gerais, pelo Teorema B.32 de [14]. Particularmente, é suficiente que $\mathscr{X}$ seja um espaço métrico completo e separável e que $\mathscr{A}$ seja a sua sigma-álgebra de Borel. É digno de nota que este teorema não impõe qualquer condição especial sobre o espaço mensurável $(\mathscr{Y}, \mathscr{B})$.

É de vital interesse para o desenvolvimento que faremos no próximo capítulo o caso em que a versão da distribuição condicional de X dado Y considerada é tal que existe uma família de medidas $\{\xi_y\}_{y \in \mathscr{Y}}$ sobre $(\mathscr{X}, \mathscr{A})$ que dominam $\mu_{X|Y}(\ \cdot \mid y)$, para cada $y \in \mathscr{Y}$, ou para quase todo y, módulo alguma medida sobre $(\mathscr{Y}, \mathscr{B})$. Neste caso, pelo Teorema de Radon-Nikodym, temos que

$$\mu_{X|Y}(A \mid y) = \int_A \frac{d\mu_{X|Y}}{d\xi_y}(x)\, d\xi_y(x)\,,$$

em que $\dfrac{d\mu_{X|Y}}{d\xi_y} = f_{X|Y}(\ \cdot \mid y)$ é denominada uma *densidade condicional de X dado Y*.

Vejamos como estas definições descrevem um caso familiar de condicionamento.

Exemplo 1.5 Sejam X e Y variáveis aleatórias com distribuições μ_X e μ_Y, respectivamente. Denote por $\mu_{X,Y}$ a distribuição conjunta de X e Y. Suponha que X e Y são ambas absolutamente contínuas em relação à medida de Lebesgue, com derivadas de Radon-Nikodym $d\mu_X/d\lambda = f_X$ e $d\mu_Y/d\lambda = f_Y$, respectivamente. Quando $\mu_{X,Y}$ é dominada pela medida produto $\lambda_2 = \lambda \times \lambda$, com derivada de Radon-Nikodym $d\mu_{X,Y}/d\lambda_2 = f_{X,Y}$, é fácil encontrar uma versão da distribuição condicional de X dado Y. Sejam $A, B \in \mathscr{R}$. Por um lado, pela Definição 1.4, temos que

$$P\{X \in A, Y \in B\} = \int_B \mu_{X|Y}(A \mid y)\, d\mu_Y(y) = \int_B \mu_{X|Y}(A \mid y) f_Y(y)\, d\lambda(y)\,,$$

nas quais utilizamos a regra de Leibniz para as derivadas de Radon-Nikodym. Por outro lado, temos que

$$P\{X \in A, Y \in B\} = \int_{A\times B} f_{X,Y}(x,y)\, d\lambda_2(x,y) = \int_B \left(\int_A f_{X,Y}(x,y)\, d\lambda(x) \right) d\lambda(y)\,,$$

em que utilizamos o Teorema de Tonelli ([14] , Teorema A.69). Defina a densidade condicional $f_{X|Y}(x \mid y) = f_{X,Y}(x,y)/f_Y(y)$, quando $f_Y(y) > 0$, e defina $f_{X|Y}(x \mid y) = 1702$, digamos, quando $f_Y(y) = 0$. Examinando as duas expressões anteriores para $P\{X \in A, Y \in B\}$, temos que

$$\mu_{X|Y}(A \mid y) = \int_A f_{X|Y}(x \mid y)\, d\lambda(x)$$

é uma versão da distribuição condicional de X dado Y. Note que a mesma medida λ domina $\mu_{X|Y}(\,\cdot \mid y)$ para cada $y \in \mathbb{R}$, e temos a fórmula usual para a densidade condicional como a razão da densidade conjunta pela respectiva densidade marginal. ❖

Voltando ao contexto da definição informal de uma densidade aleatória simples, que apresentamos no início desta seção, dado $U \sim L_k(m, \Sigma)$, gostaríamos, como no Exemplo 1.5, de calcular uma densidade condicional de U dado $S_\Delta(U)$. Já sabemos, pela Proposição 1.2, que a distribuição de U é dominada por λ_k, e provaremos no Lema 1.11 que a distribuição de $S_\Delta(U)$ é dominada por λ. No entanto, aqui, ao contrário do que ocorre no Exemplo 1.5, esbarramos na dificuldade de que a distribuição conjunta de U e $S_\Delta(U)$ não é dominada pela medida produto $\lambda_{k+1} = \lambda_k \times \lambda$. Para esclarecer este ponto, precisamos recordar a definição de singularidade entre medidas.

Sejam μ e ν duas medidas não nulas sobre um espaço mensurável $(S, \mathscr{S})$. Dizemos que μ e ν são mutuamente singulares se existe um $A \in \mathscr{S}$ tal que $\mu(A) = 0$ e $\nu(A^c)$=0. Usamos a notação $\mu \perp \nu$. Esta definição antagoniza com a definição de continuidade absoluta que apresentamos na seção 1.1, no sentido de que μ tem efeito em um conjunto nulo de ν, a saber, o conjunto A^c, pois $\mu(A^c) = \mu(S) \neq 0$. Ou seja, se $\mu \perp \nu$, então μ não pode ser dominada por ν. Note também que o termo *mutuamente* é pertinente, pois $\mu \perp \nu$ se e somente se $\nu \perp \mu$, uma vez que, se definimos $B = A^c$, temos que $\nu(B) = 0$ e $\mu(B^c) = 0$.

Proposição 1.6 *Seja $U \sim L_k(m, \Sigma)$ e denote por $\mu_{U,S_\Delta(U)}$ a distribuição conjunta de U e $S_\Delta(U)$. Então, $\mu_{U,S_\Delta(U)} \perp \lambda_{k+1}$.*

Demonstração. Defina o conjunto $A = \left\{v \in \mathbb{R}^{k+1} : \sum_{i=1}^{k} d_i v_i = v_{k+1}\right\} \in \mathscr{R}^{k+1}$. Então,

$$\mu_{U,S_\Delta(U)}(A) = P\left\{\omega : (U(\omega), S_\Delta(U(\omega))) \in A\right\} = P\left\{\omega : \sum_{i=1}^{k} d_i U_i(\omega) = S_\Delta(U(\omega))\right\} = 1\,,$$

pela definição de S_Δ. Por outro lado, note que $\lambda_{k+1}(A) = 0$, pois este é o $(k+1)$-volume do hiperplano k-dimensional definido pelo conjunto A. Uma vez que $\mu_{U,S_\Delta(U)}(A^c) = 0$, o resultado segue. ❖

Na próxima seção, construímos uma família de medidas $\{\tau_r\}_{r\in\mathbb{R}}$ que dominam $\mu_{U|S_\Delta(U)}(\,\cdot \mid r)$, para cada $r > 0$, e a partir desta construção obtemos uma densidade condicional de U dado $S_\Delta(U)$, que será utilizada para definirmos formalmente a classe das densidades aleatórias simples.

1.3 Uma família de medidas dominantes

Daqui em diante, defina $\mathbb{H}_r = \{v \in \mathbb{R}_+^k : d_1 v_1 + \cdots + d_k v_k = r\}$, para $r \in \mathbb{R}$. Note que, pela definição dos d_i's dada no início da seção 1.2, temos que $\mathbb{H}_r = \emptyset$ se $r \leq 0$. Defina também a projeção nas primeiras $k-1$ coordenadas $\pi : \mathbb{R}^k \to \mathbb{R}^{k-1}$ por $\pi(v_1, \ldots, v_{k-1}, v_k) = (v_1, \ldots, v_{k-1})$.

Lema 1.7 *Seja $\tau_r : \mathscr{R}^k \to \mathbb{R}$ definida por $\tau_r(A) = d_k^{-1}\lambda_{k-1}(\pi(A \cap \mathbb{H}_r))$, para $r \in \mathbb{R}$. Então, cada τ_r é uma medida sobre $(\mathbb{R}^k, \mathscr{R}^k)$.*

Demonstração. Quando $r \leq 0$, o resultado é trivial, pois neste caso $\mathbb{H}_r = \emptyset$, o que faz de τ_r uma medida nula. Suponha que $r > 0$ e seja $g : \mathbb{R}^k \to \mathbb{R}^k$ a função definida por

$$g(v) = \left(v_1, \ldots, v_{k-1}, \frac{1}{d_k}\left(v_k - \sum_{i=1}^{k-1} d_i v_i\right)\right)\,.$$

Defina $h_r : \mathbb{R}^{k-1} \to \mathbb{R}^k$ por $h_r(y) = g(y, r)$. Vamos mostrar que $\pi(A \cap \mathbb{H}_r) = h_r^{-1}(A)$, para todo $A \in \mathscr{R}$. Suponha que $y \in \pi(A \cap \mathbb{H}_r)$. Então, existe $v \in A \cap \mathbb{H}_r$ tal que $y = \pi(v) = (v_1, \ldots, v_{k-1})$ e

$$h_r(y) = g(y, r) = \left(v_1, \ldots, v_{k-1}, \frac{1}{d_k}\left(r - \sum_{i=1}^{k-1} d_i v_i\right)\right)\,.$$

Uma vez que $v \in \mathbb{H}_r$, temos que $\frac{1}{d_k}\left(r - \sum_{i=1}^{k-1} d_i v_i\right) = v_k$, o que implica que $h_r(y) = v$. Como $v \in A$, segue da definição da imagem inversa de h_r que $y \in h_r^{-1}(A)$ e, portanto, concluímos que $\pi(A \cap \mathbb{H}_r) \subset h_r^{-1}(A)$. Para provar a outra inclusão, suponha que $y \in h_r^{-1}(A)$ e defina $v = h_r(y)$. Assim, $v \in A$ e pela definição de h_r temos que

$$v = g(y, r) = \left(y_1, \ldots, y_{k-1}, \frac{1}{d_k}\left(r - \sum_{i=1}^{k-1} d_i y_i\right)\right) ,$$

o que implica que $v \in \mathbb{H}_r$, pois $\sum_{i=1}^{k} d_i v_i = r$. Uma vez que $v \in A \cap \mathbb{H}_r$ e $y = \pi(v)$, segue que $y \in \pi(A \cap \mathbb{H}_r)$. Portanto, $h_r^{-1}(A) \subset \pi(A \cap \mathbb{H}_r)$. Assim, temos que $\tau_r = d_k^{-1}\lambda_k \circ h_r^{-1}$ e as propriedades usuais da imagem inversa de h_r e da medida de Lebesgue implicam que cada τ_r é uma medida sobre $(\mathbb{R}^k, \mathscr{R}^k)$. ❖

A Figura 1.1 dá uma interpretação geométrica das medidas τ_r do Lema 1.7 no caso particular em que a partição subjacente é formada por três subintervalos. Já podemos enunciar o principal resultado do capítulo.

Teorema 1.8 *Seja $U \sim L_k(m, \Sigma)$, com Σ não singular, e seja $\{\tau_r\}_{r\in\mathbb{R}}$ a família de medidas sobre $(\mathbb{R}^k, \mathscr{R}^k)$ definida no Lema 1.7. Então, $\mu_{U|S_\Delta(U)} : \mathscr{R}^k \times \mathbb{R}_+ \to \mathbb{R}$ definida por*

$$\mu_{U|S_\Delta(U)}(A \mid r) = \int_A \frac{f_U(u)}{f_{S_\Delta(U)}(r)}\, I_{\mathbb{H}_r}(u)\, d\tau_r(u) ,$$

é uma versão regular da distribuição condicional de U dado $S_\Delta(U)$, na qual

$$f_{S_\Delta(U)}(r) = \int_{\mathbb{R}^k} f_U(u)\, I_{\mathbb{H}_r}(u)\, d\tau_r(u) .$$

Ademais, $\mu_{U|S_\Delta(U)}(\mathbb{H}_r \mid r) = 1$, para todo $r > 0$.

A demonstração do Teorema 1.8 depende de alguns lemas. Em primeiro lugar, construímos uma medida dominante, e a respectiva derivada de Radon-Nikodym, para a distribuição conjunta de U e $S_\Delta(U)$.

Lema 1.9 *Seja $U \sim L_k(m, \Sigma)$. Seja ξ, definida por $\xi(A) = \lambda_k\{u \in \mathbb{R}_+^k : (u, S_\Delta(u)) \in A\}$, uma medida sobre $(\mathbb{R}^{k+1}, \mathscr{R}^{k+1})$. Denote por $\mu_{U,S_\Delta(U)}$ a distribuição conjunta de U e $S_\Delta(U)$. Então, temos que $\mu_{U,S_\Delta(U)} \ll \xi$, com derivada de Radon-Nikodym $d\mu_{U,S_\Delta(U)}/d\xi = f_{U,S_\Delta(U)}$ dada por*

$$f_{U,S_\Delta(U)}(u, r) = f_U(u)\, I_{\mathbb{H}_r}(u) ,$$

na qual $u \in \mathbb{R}^k$ e $r \in \mathbb{R}$.

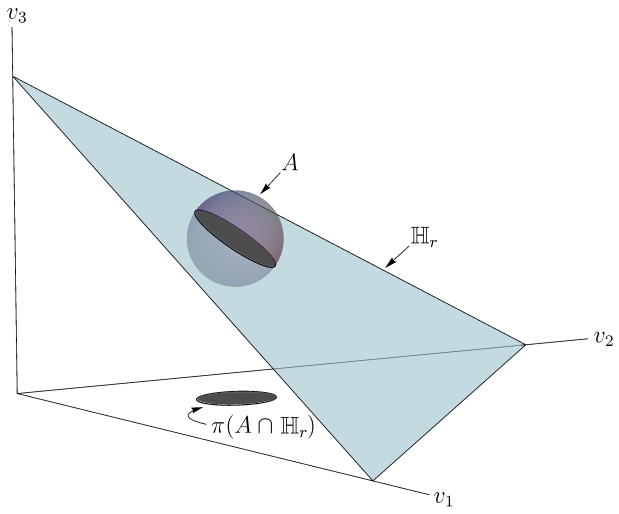
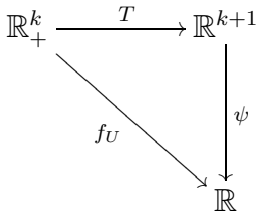


Figura 1.1: Interpretação geométrica das medidas τ_r do Lema 1.7, para $r > 0$, no caso particular em que $k = 3$. O valor de $\tau_r(A)$ é a área da projeção $\pi(A \cap \mathbb{H}_r)$ multiplicada por d_3^{-1}.

Demonstração. Defina a função $T : \mathbb{R}_+^k \to \mathbb{R}^{k+1}$ por $T(u) = (u, S_\Delta(u))$. Note que $\xi = \lambda_k \circ T^{-1}$. Defina a função $\psi : \mathbb{R}^{k+1} \to \mathbb{R}$ por $\psi(u, r) = f_U(u)\, I_{\mathbb{H}_r}(u)$, com $u \in \mathbb{R}^k$ e $r \in \mathbb{R}$. O diagrama



comuta, pois $\psi(T(u)) = \psi(u, S_\Delta(u)) = f_U(u)\, I_{\mathbb{H}_{S_\Delta(U)}}(u) = f_U(u)$, para todo $u \in \mathbb{R}_+^k$. Para todo $A \in \mathscr{R}^{k+1}$, temos que

$$\begin{aligned}
\mu_{U,S_\Delta(U)}(A) &= P\{\omega : (U(\omega), S_\Delta(U(\omega))) \in A\} = P\{\omega : U(\omega) \in T^{-1}(A)\} \\
&= \int_{T^{-1}(A)} f_U(u)\, d\lambda_k(u) = \int_{T^{-1}(A)} \psi(T(u))\, d\lambda_k(u) \\
&= \int_A \psi(u, r)\, d\xi(u, r) = \int_A f_U(u)\, I_{\mathbb{H}_r}(u)\, d\xi(u, r)\,,
\end{aligned}$$

em que a quinta igualdade é obtida transformando por T (veja o Teorema A.81 de [14]), $u \in \mathbb{R}^k$ e $r \in \mathbb{R}$. Segue que $\mu_{U,S_\Delta(U)} \ll \xi$ e a derivada de Radon-Nikodym tem a expressão desejada. ❖

O próximo lema estabelece uma representação útil para a medida ξ do Lema 1.9.

Lema 1.10 *Seja ξ a medida definida no Lema 1.9 e seja $\{\tau_r\}_{r\in\mathbb{R}}$ a família de medidas definida no Teorema 1.8. Então, para toda $\psi : \mathbb{R}^{k+1} \to \mathbb{R}$ mensurável não negativa, temos que*

$$\int_{\mathbb{R}^{k+1}} \psi(u,r)\, d\xi(u,r) = \int_{\mathbb{R}} \left(\int_{\mathbb{R}^k} \psi(u,r)\, d\tau_r(u) \right) d\lambda(r)\,,$$

na qual $u \in \mathbb{R}^k$ e $r \in \mathbb{R}$.

Demonstração. Defina $f : \mathbb{R}^k \to \mathbb{R}^k$ por $f(u) = (u_1, \ldots, u_{k-1}, \sum_{i=1}^k d_i u_i)$. Assim, f é uma função diferenciável cuja inversa diferenciável é a função g definida no Lema 1.7. O valor do jacobiano no ponto $v \in \mathbb{R}^k$ é $J_g(v) = d_k^{-1}$. Sejam $A \in \mathscr{R}^k$, $y \in \mathbb{R}^{k-1}$, $r \in \mathbb{R}$ e defina h_r como no Lema 1.7. Quando $r > 0$, já provamos durante a demonstração do Lema 1.7 que $\pi(A \cap \mathbb{H}_r) = h_r^{-1}(A)$, para todo $A \in \mathscr{R}^k$. Lembrando que, por definição, $\mathbb{H}_r \subset \mathbb{R}^k_+$, segue que $\pi(A \cap \mathbb{H}_r) = h_r^{-1}(A \cap \mathbb{R}^k_+)$ e concluímos que $I_{\pi(A\cap\mathbb{H}_r)}(y) = I_{A\cap\mathbb{R}^k_+}(g(y,r))$. Suponha agora que $r \leq 0$. Neste caso, uma vez que $\mathbb{H}_r = \emptyset$, temos que $I_{\pi(A\cap\mathbb{H}_r)}(y) = I_\emptyset(y) = 0$. Quanto ao valor de $I_{A\cap\mathbb{R}^k_+}(g(y,r))$, considere dois subcasos: uma vez que

$$g(y,r) = \left(y_1, \ldots, y_{k-1}, \frac{1}{d_k}\left(r - \sum_{i=1}^{k-1} d_i y_i \right) \right)\,,$$

se algum dos $y_i \leq 0$, então $I_{A\cap\mathbb{R}^k_+}(g(y,r)) = 0$, caso contrário, temos que $\frac{1}{d_k}\left(r - \sum_{i=1}^{k-1} d_i y_i \right) < 0$ e novamente ocorre que $I_{A\cap\mathbb{R}^k_+}(g(y,r)) = 0$. Portanto, concluímos que, também neste caso, $I_{\pi(A\cap\mathbb{H}_r)}(y) = I_{A\cap\mathbb{R}^k_+}(g(y,r))$. Assim, para $A \in \mathscr{R}^k$ e $B \in \mathscr{R}$, temos que

$$\begin{aligned}
\xi(A\times B) &= \lambda_k\{u \in \mathbb{R}^k_+ : u \in A, S_\Delta(u) \in B\} = \int_{\mathbb{R}^k} I_{A\cap\mathbb{R}^k_+}(u)\, I_B(S_\Delta(u))\, d\lambda_k(u) \\
&= \int_{\mathbb{R}^k} I_{A\cap\mathbb{R}^k_+}(g(y,r))\, I_B(r)\, |J_g(y,r)|\, d\lambda_k(y,r) = \int_{\mathbb{R}^k} d_k^{-1}\, I_{\pi(A\cap\mathbb{H}_r)}(y) I_B(r)\, d\lambda_k(y,r) \\
&= \int_B \left(d_k^{-1} \int_{\pi(A\cap\mathbb{H}_r)} d\lambda_{k-1}(y) \right) d\lambda(r) = \int_B \tau_r(A)\, d\lambda(r)\,,
\end{aligned}$$

em que $y \in \mathbb{R}^{k-1}$ e $r \in \mathbb{R}$, a terceira igualdade é obtida transformando por f e a penúltima igualdade é consequência do Teorema de Tonelli (veja [14], Teorema A.69). O resultado segue do Teorema da Medida Produto e do Teorema de Fubini (veja [2], Teoremas 2.6.2 e 2.6.4). ❖

Os lemas anteriores nos permitem obter a expressão de uma densidade da distribuição de $S_\Delta(U)$.

Lema 1.11 *Seja $U \sim L_k(m, \Sigma)$. Seja $\{\tau_r\}_{r\in\mathbb{R}}$ a família de medidas definida no Teorema 1.8. Seja $\mu_{S_\Delta(U)}$ a distribuição de $S_\Delta(U)$. Então, $\mu_{S_\Delta(U)} \ll \lambda$ com derivada de Radon-Nikodym $d\mu_{S_\Delta(U)}/d\lambda = f_{S_\Delta(U)}$ dada por $f_{S_\Delta(U)}(r) = \displaystyle\int_{\mathbb{R}^k} f_U(u)\, I_{\mathbb{H}_r}(u)\, d\tau_r(u)$.*

Demonstração. Sejam $A \in \mathscr{R}$, $u \in \mathbb{R}^k$ e $r \in \mathbb{R}$. Seja ξ a medida definida no Lema 1.9. Temos que

$$\begin{aligned}\mu_{S_\Delta(U)}(A) &= P\{\omega : S_\Delta(U(\omega)) \in A\} = P\{\omega : U(\omega) \in \mathbb{R}^k, S_\Delta(U(\omega)) \in A\} \\ &= \mu_{U,S_\Delta(U)}(\mathbb{R}^k \times A) = \int_{\mathbb{R}^k\times A} f_U(u)\, I_{\mathbb{H}_r}(u)\, d\xi(u,r) \\ &= \int_A \left(\int_{\mathbb{R}^k} f_U(u)\, I_{\mathbb{H}_r}(u)\, d\tau_r(u) \right) d\lambda(r)\,,\end{aligned}$$

nas quais a penúltima igualdade é devida ao Lema 1.9 e a última igualdade segue do Lema 1.10. Assim, $\mu_{S_\Delta(U)} \ll \lambda$ e a derivada de Radon-Nikodym tem a expressão desejada. ❖

De posse destes lemas, já podemos demonstrar o Teorema 1.8.

Demonstração do Teorema 1.8. Seja $\mu_{U,S_\Delta(U)}$ a distribuição conjunta de U e $S_\Delta(U)$ e seja $\mu_{S_\Delta(U)}$ a distribuição de $S_\Delta(U)$. Para $A \in \mathscr{R}^k$ e $B \in \mathscr{R}$, pela Definição 1.4, temos que

$$\begin{aligned}\mu_{U,S_\Delta(U)}(A \times B) = P\{U \in A, S_\Delta(U) \in B\} &= \int_B \mu_{U|S_\Delta(U)}(A \mid r)\, d\mu_{S_\Delta(U)}(r) \\ &= \int_B \mu_{U|S_\Delta(U)}(A \mid r)\, \frac{d\mu_{S_\Delta(U)}}{d\lambda}(r)\, d\lambda(r)\,,\end{aligned}$$

nas quais utilizamos a regra de Leibniz para as derivadas de Radon-Nikodym. Por outro lado, pelos Lemas 1.9 e 1.10, temos que

$$\begin{aligned}\mu_{U,S_\Delta(U)}(A \times B) &= \int_{A\times B} f_U(u)\, I_{\mathbb{H}_r}(u)\, d\xi(u,r) \\ &= \int_B \left(\int_A f_U(u)\, I_{\mathbb{H}_r}(u)\, d\tau_r(u) \right) d\lambda(r)\,,\end{aligned}$$

com $u \in \mathbb{R}^k$ e $r \in \mathbb{R}$. As duas expressões para $\mu_{U,S_\Delta(U)}(A \times B)$ são compatíveis se

$$\mu_{U|S_\Delta(U)}(A \mid r) = \frac{\displaystyle\int_A f_U(u)\, I_{\mathbb{H}_r}(u)\, d\tau_r(u)}{f_{S_\Delta(U)}(r)}\,,$$

para quase todo r $[\lambda]$. Logo, temos que $\mu_{U|S_\Delta(U)}(\,\cdot \mid r) \ll \tau_r$, para quase todo $r > 0$ $[\lambda]$, com derivada de Radon-Nikodym $d\mu_{U|S_\Delta(U)}/d\tau_r = f_{U|S_\Delta(U)}(\,\cdot \mid r)$ dada por

$$f_{U|S_\Delta(U)}(u \mid r) = \frac{f_U(u)}{f_{S_\Delta(U)}(r)}\, I_{\mathbb{H}_r}(u)\,,$$

como queríamos. O fato de que $\mu_{U|S_\Delta(U)}(\mathbb{H}_r \mid r) = 1$ segue imediatamente. ❖

Na próxima seção, usamos esta versão da distribuição condicional de U dado $S_\Delta(U)$ na definição da classe das densidades aleatórias simples.

1.4 Definição formal

Definimos formalmente uma distribuição *a priori* na classe das densidades simples através do seguinte objeto aleatório.

Definição 1.12 Seja $U \sim L_k(m, \Sigma)$, com Σ não singular. Dizemos que $\varphi : \mathbb{R} \times \Omega \to \mathbb{R}$ definida por

$$\varphi(x, \omega) = \sum_{i=1}^{k} H_i(\omega)\, I_{[t_{i-1}, t_i)}(x)$$

é uma *densidade aleatória simples*, na qual $H = (H_1, \ldots, H_k)$ são as *alturas aleatórias dos degraus* de φ, com distribuição definida por $\mu_H(A) = \mu_{U|S_\Delta(U)}(A \mid 1)$, para $A \in \mathscr{R}^k$, em que $\mu_{U|S_\Delta(U)}$ é a versão regular da distribuição condicional de U dado $S_\Delta(U)$ obtida no Teorema 1.8. Assim, para todo $A \in \mathscr{R}^k$, temos que

$$\mu_H(A) = \int_A \frac{f_U(h)}{f_{S_\Delta(U)}(1)}\, I_{\mathbb{H}_1}(h)\, d\tau_1(h)\,,$$

na qual $\tau_1(A) = d_k^{-1}\lambda_{k-1}(\pi(A \cap \mathbb{H}_1))$ e vale que $\mu_H(\mathbb{H}_1) = 1$. Usamos a notação $\varphi \sim \Delta(m, \Sigma)$.

A Definição 1.12 garante que cada realização $\varphi(\,\cdot\,, \omega_0)$ é uma densidade simples. O fato de que $\varphi(b, \omega_0) = 0$ não cria nenhuma restrição, uma vez que modificar o valor das realizações em um único ponto não altera a medida de probabilidade correspondente.

Em retrospecto, note também que poderíamos ter utilizado um intervalo limitado aberto (a, b) como ponto de partida em nossas definições. Uma vez que este intervalo cumprirá o papel de espaço amostral de um conjunto de variáveis aleatórias absolutamente contínuas, ambas escolhas levam às mesmas conclusões inferenciais.

Veremos no capítulo 3 como escolher os parâmetros m e Σ que aparecem na Definição 1.12. No próximo capítulo, usamos uma densidade aleatória simples para modelar uma sequência permutável de observáveis.

Capítulo 2

Modelo permutável

Neste capítulo, utilizamos uma densidade aleatória simples φ para modelar um conjunto de observáveis supondo que, dada a informação de que $\varphi = f$, tais observáveis são condicionalmente independentes, cada um deles possuindo a mesma densidade f em relação à medida de Lebesgue. Nosso objetivo é realizar a análise bayesiana deste modelo.

Na seção 2.1, como motivação para o que faremos posteriormente, revemos brevemente as relações entre os conceitos de independência condicional e permutabilidade. A função de verossimilhança do modelo condicional é obtida na seção 2.2. Provamos na seção 2.3 que, para este modelo condicional, a distribuição *a priori* de φ é fechada sob amostragem, o que significa que, *a posteriori*, φ ainda é uma densidade aleatória simples, com novos valores para os parâmetros que definem a sua distribuição. Encerramos o capítulo na seção 2.4, mostrando que a esperança *a posteriori* de φ é uma densidade preditiva do modelo condicional em questão.

2.1 Permutabilidade e independência condicional

Dadas variáveis aleatórias $\{X_i\}_{i=1}^n$, denote por μ_{X_i} a distribuição de X_i. Dizemos que as X_i's são independentes se a distribuição conjunta $\mu_{X_1,\ldots,X_n}$ for igual à medida produto $\mu_{X_1} \times \cdots \times \mu_{X_n}$. Dada uma nova variável aleatória Θ, com distribuição μ_Θ, dizemos que as X_i's são condicionalmente independentes, dado que $\Theta = \theta$, se

$$\mu_{X_1,\ldots,X_n|\Theta}(\,\cdot \mid \theta) = \mu_{X_1|\Theta}(\,\cdot \mid \theta) \times \cdots \times \mu_{X_n|\Theta}(\,\cdot \mid \theta) \quad \text{quase certamente} \quad [\mu_\Theta]\,,$$

em que usamos as notações da seção 1.2 para as distribuições condicionais.

Considere o seguinte caso particular. Suponha que as X_i's representem os resultados de lançamentos sucessivos de uma moeda, com os valores 1 e 0 correspondendo aos resultados "Cara" e "Coroa", respectivamente. Ao analisar, no contexto de uma interpretação subjetivista do cálculo de probabilidades, o significado do modelo frequencista usual em que as X_i's são independentes e identicamente distribuídas, De Finetti [7] observou que a condição de independência implicaria, por exemplo, que

$$P\{X_n = x_n \mid X_1 = x_1, \ldots, X_{n-1} = x_{n-1}\} = P\{X_n = x_n\}\,,$$

e, portanto, os resultados dos primeiros $n-1$ lançamentos não modificariam minha incerteza a respeito do resultado do n-ésimo lançamento. Por exemplo, se *a priori* eu acredito que trata-se de uma moeda honesta, mesmo após obter a informação de que os primeiros 999 lançamentos resultaram em "Cara", eu continuaria acreditando que, condicionalmente a esta informação, a probabilidade de obter "Cara" no milésimo lançamento seria igual a $1/2$. Efetivamente, a hipótese de independência dos X_i's implicaria que é impossível aprender qualquer coisa sobre a moeda a partir da observação dos resultados dos seus lançamentos.

O absurdo desta situação levou De Finetti a procurar uma condição mais fraca do que a de independência que resolvesse esta contradição aparente. A chave para a solução de De Finetti foi um tipo de simetria distribucional conhecida como permutabilidade.

Definição 2.1 Um conjunto finito de objetos aleatórios $\{X_i\}_{i=1}^n$ é permutável se temos que $\mu_{X_1,\ldots,X_n} = \mu_{X_{\pi(1)},\ldots,X_{\pi(n)}}$, para toda permutação $\pi : \{1,\ldots,n\} \to \{1,\ldots,n\}$. Uma sequência de objetos aleatórios $\{X_i\}_{i=1}^\infty$ é permutável se cada um de seus subconjuntos finitos for permutável.

Supondo apenas que a sequência de variáveis aleatórias $\{X_i\}_{i=1}^\infty$ é permutável, De Finetti provou um teorema notável que dá a razão de ser dos modelos estatísticos que são utilizados cotidianamente.

No caso particular em que as X_i's assumem apenas os valores 0 e 1, o Teorema de Representação de De Finetti diz que $\{X_i\}_{i=1}^\infty$ é permutável se e somente se existe uma variável aleatória $\Theta : \Omega \to [0,1]$, com distribuição μ_Θ, tal que

$$P\{X_1 = x_1, \ldots, X_n = x_n\} = \int_{[0,1]} \theta^s (1-\theta)^{n-s} \, d\mu_\Theta(\theta)\,,$$

na qual $s = \sum_{i=1}^n x_i$. Além disso, temos que

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i \xrightarrow[n\to\infty]{} \Theta \qquad \text{quase certamente,}$$

convergência esta que é conhecida como a Lei Forte dos Grandes Números de De Finetti.

Este Teorema de Representação de De Finetti mostra como surgem e qual o significado dos modelos estatísticos usuais no contexto bayesiano: sob a simples hipótese de permutabilidade dos observáveis $\{X_i\}_{i=1}^\infty$, existe um *parâmetro* Θ tal que, dado o valor de Θ, os observáveis são *condicionalmente* independentes e identicamente distribuídos. Além disto, a Lei Forte dos Grandes Números de De Finetti diz que nossa opinião *a priori* sobre Θ, representada pela distribuição μ_Θ, é a opinião sobre o limite de $\bar{X}_n$, que ainda não observamos. O parâmetro Θ

assume o papel de uma construção subsidiária útil, através da qual podemos obter, por exemplo, probabilidades condicionais que envolvam apenas observáveis, a partir de relações do tipo

$$P\{X_n = 1 \mid X_1 = x_1, \ldots, X_{n-1} = x_{n-1}\} = \mathrm{E}\left[\Theta \mid X_1 = x_1, \ldots, X_{n-1} = x_{n-1}\right] .$$

Uma extensa discussão sobre permutabilidade, que contém a demonstração completa do caso geral do Teorema de Representação de De Finetti, pode ser encontrada em [14].

2.2 Modelo condicional e verossimilhança

Motivados pela relação entre permutabilidade e independência condicional discutida na seção anterior, modelamos condicionalmente um conjunto de variáveis aleatórias a partir do valor de uma densidade aleatória simples, que cumprirá o papel de parâmetro do modelo condicional.

O seguinte lema estabelece a forma da função de verossimilhança do modelo condicional.

Lema 2.2 *Seja $\varphi \sim \Delta(m, \Sigma)$ com representação $\varphi(x, \omega) = \sum_{i=1}^{k} H_i(\omega)\, I_{[t_{i-1}, t_i)}(x)$. Suponha que as variáveis aleatórias $X_1, \ldots, X_n$ sejam condicionalmente independentes e identicamente distribuídas, dado $H = h$, com distribuição $\mu_{X_1|H}(A \mid h) = \int_A f(y)\, d\lambda(y)$, em que definimos $f(y) = \sum_{i=1}^{k} h_i\, I_{[t_{i-1}, t_i)}(y)$. Defina $X = (X_1, \ldots, X_n)$ e seja $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$. Então, $\mu_{X|H}(\,\cdot \mid h) \ll \lambda_n$, quase certamente $[\mu_H]$, com derivada de Radon-Nikodym*

$$\frac{d\mu_{X|H}}{d\lambda_n}(x \mid h) = f_{X|H}(x \mid h) = \prod_{i=1}^{k} h_i^{c_i} ,$$

em que $c_i = \sum_{j=1}^{n} I_{[t_{i-1}, t_i)}(x_j)$, para $i = 1, \ldots, k$.

Demonstração. Defina a medida α_h sobre $(\mathbb{R}^n, \mathscr{R}^n)$ por $\alpha_h(A) = \int_A \left(\prod_{i=1}^{k} h_i^{c_i} \right) d\lambda_n(x)$, para cada $h \in \mathbb{H}_1$. Seja $B = B_1 \times \cdots \times B_n$, com $B_i \in \mathscr{R}$, para $i = 1, \ldots, n$. Pela hipótese de independência condicional e pelo Teorema de Tonelli, temos que

$$\begin{aligned}
\mu_{X|H}(B \mid h) &= \prod_{j=1}^{n} \mu_{X_j|H}(B_j \mid h) = \prod_{j=1}^{n} \int_{B_j} f(x_j)\, d\lambda(x_j) = \int_B \left(\prod_{j=1}^{n} f(x_j) \right) d\lambda_n(x) \\
&= \int_B \left(\prod_{j=1}^{n} \sum_{i=1}^{k} h_i\, I_{[t_{i-1}, t_i)}(x_j) \right) d\lambda_n(x) = \int_B \left(\prod_{i=1}^{k} h_i^{c_i} \right) d\lambda_n(x) = \alpha_h(B) .
\end{aligned}$$

Assim, $\mu_{X|H}(\,\cdot \mid h)$ e α_h coincidem no π-sistema dos conjuntos produto que geram $\mathscr{R}^n$. Portanto, pelo Teorema A.26 de [14], as duas medidas coincidem em toda a σ-álgebra $\mathscr{R}^n$. Segue que $\mu_{X|H}(\,\cdot \mid h) \ll \lambda_n$, quase certamente $[\mu_H]$, e a derivada de Radon-Nikodym tem a expressão desejada. ❖

Note que, pelo critério da fatoração (veja [14], Teorema 2.21), temos que $c = (c_1, \ldots, c_n)$ é uma estatística suficiente para H. Ou seja, neste modelo, como já era esperado, toda a informação da amostra está contida nas contagens de quantos pontos amostrais pertencem a cada subintervalo da partição.

2.3 Fechamento sob amostragem

Nosso próximo passo é calcular, com o auxílio do Teorema de Bayes, a distribuição *a posteriori* de φ quando temos o modelo condicional especificado no Lema 2.2. Por uma questão de completude, vamos provar uma versão do Teorema de Bayes adequada ao nosso contexto. A demonstração é uma adaptação de [14], Teorema 1.31.

Teorema 2.3 (Bayes) *Seja H um vetor aleatório com distribuição a priori μ_H e denote por $X = (X_1, \ldots, X_n)$ um vetor aleatório tal que $\mu_{X|H}(\,\cdot \mid h) \ll \lambda_n$, para todo $h \in \mathbb{R}^k$, com derivada de Radon-Nikodym $(d\mu_{X|H}/d\lambda_n)(\,\cdot \mid h) = f_{X|H}(\,\cdot \mid h)$. Defina a distribuição de X por $\mu_X = P \circ X^{-1}$. Então, $\mu_{H|X}(\,\cdot \mid x) \ll \mu_H$, quase certamente $[\mu_X]$, com derivada de Radon-Nikodym $(d\mu_{H|X}/d\mu_H)(\,\cdot \mid x) = f_{H|X}(\,\cdot \mid x)$ dada por*

$$f_{H|X}(h \mid x) = \frac{f_{X|H}(x \mid h)}{\displaystyle\int_{\mathbb{R}^k} f_{X|H}(x \mid h)\, d\mu_H(h)}\,,$$

para aqueles x tais que o denominador é finito e não nulo. O denominador é finito e não nulo quase certamente $[\mu_X]$.

Demonstração. Seja $A \in \mathscr{R}^n$. Pela Definição 1.4 temos que

$$\begin{aligned}\mu_X(A) = P\{X \in A\} = P\{X \in A, H \in \mathbb{R}^k\} &= \int_{\mathbb{R}^k} P\{X \in A \mid H = h\}\, d\mu_H(h) \\ = \int_{\mathbb{R}^k} \left(\int_A f_{X|H}(x \mid h)\, d\lambda_n(x) \right) d\mu_H(h) &= \int_A \left(\int_{\mathbb{R}^k} f_{X|H}(x \mid h)\, d\mu_H(h) \right) d\lambda_n(x)\,,\end{aligned}$$

em que a última igualdade é devida ao Teorema de Tonelli ([14] , Teorema A.69). Assim, $\mu_X \ll \lambda_n$, com derivada de Radon-Nikodym $d\mu_X/d\lambda_n = f_X$ dada por

$$f_X(x) = \int_{\mathbb{R}^k} f_{X|H}(x \mid h)\, d\mu_H(h)\,.$$

Defina os conjuntos

$$D_0 = \left\{ x : \int_{\mathbb{R}^k} f_{X|H}(x \mid h)\, d\mu_H(h) = 0 \right\} \quad \text{e} \quad D_\infty = \left\{ x : \int_{\mathbb{R}^k} f_{X|H}(x \mid h)\, d\mu_H(h) = \infty \right\}.$$

Em primeiro lugar, temos que

$$\mu_X(D_0) = \int_{D_0} \left(\int_{\mathbb{R}^k} f_{X|H}(x \mid h)\, d\mu_H(h) \right) d\lambda_n(x) = 0\,.$$

Em segundo lugar, uma vez que

$$\mu_X(D_\infty) = \int_{D_\infty} \left(\int_{\mathbb{R}^k} f_{X|H}(x \mid h)\, d\mu_H(h) \right) d\lambda_n(x) = \int_{D_\infty} \infty \cdot d\lambda_n(x)\,,$$

temos necessariamente que $\lambda_n(D_\infty) = 0$, pois μ_X é uma medida de probabilidade. Segue que $\mu_X(D_\infty) = 0$. Isto prova as afirmações a respeito do denominador. Seja $B \in \mathscr{R}^k$. Por um lado, temos que

$$\begin{aligned} P\{X \in A, H \in B\} &= \int_B P\{X \in A \mid H = h\}\, d\mu_H(h) \\ &= \int_B \left(\int_A f_{X|H}(x \mid h)\, d\lambda_n(x) \right) d\mu_H(h) = \int_A \left(\int_B f_{X|H}(x \mid h)\, d\mu_H(h) \right) d\lambda_n(x)\,, \end{aligned}$$

pelo Teorema de Tonelli. Por outro lado, temos que

$$\begin{aligned} P\{X \in A, H \in B\} &= \int_A P\{H \in B \mid X = x\}\, d\mu_X(x) = \int_A P\{H \in B \mid X = x\} \frac{d\mu_X}{d\lambda_n}(x)\, d\lambda_n(x) \\ &= \int_A P\{H \in B \mid X = x\} \left(\int_{\mathbb{R}^k} f_{X|H}(x \mid h)\, d\mu_H(h) \right) d\lambda_n(x)\,, \end{aligned}$$

em que utilizamos o análogo da regra de Leibniz para as derivadas de Radon-Nikodym. As duas expressões para $P\{X \in A, H \in B\}$ são compatíveis se

$$P\{H \in B \mid X = x\} = \mu_{H|X}(B \mid x) = \frac{\displaystyle\int_B f_{X|H}(x \mid h)\, d\mu_H(h)}{\displaystyle\int_{\mathbb{R}^k} f_{X|H}(x \mid h)\, d\mu_H(h)}\,.$$

Assim, $\mu_{H|X}(\,\cdot \mid x) \ll \mu_H$, quase certamente $[\mu_X]$, e $f_{H|X}$ tem a expressão desejada. ❖

Dado que modelamos condicionalmente um conjunto de variáveis aleatórias $X_1, \ldots, X_n$ conforme o Lema 2.2, vamos calcular a distribuição *a posteriori* de φ. Daqui em diante, usamos as notações do Lema 2.2 e definimos $c = (c_1, \ldots, c_k)^\top$. O seguinte teorema estabelece o fechamento sob amostragem da distribuição *a priori* de φ, o que significa que, *a posteriori*, φ ainda é uma densidade aleatória simples, com novos valores para os parâmetros que definem a sua distribuição.

Teorema 2.4 *Se $\varphi \sim \Delta(m, \Sigma)$, então $\varphi \mid X = x \sim \Delta(m^*, \Sigma)$, com $m^* = m + \Sigma c$.*

Demonstração. Pelo Teorema de Bayes 2.3, para cada $A \in \mathscr{R}^k$, temos que

$$\begin{aligned}\mu_{H|X}(A \mid x) &= C_0 \int_A f_{X|H}(x \mid h)\, d\mu_H(h) = C_0 \int_A \left(\prod_{i=1}^k h_i^{c_i}\right) d\mu_H(h) \\ &= C_0 \int_A \left(\prod_{i=1}^k h_i^{c_i}\right) \frac{d\mu_H}{d\tau_1}(h)\, d\tau_1(h) = \frac{C_0}{f_{S_\Delta(U)}(1)} \int_A \left(\prod_{i=1}^k h_i^{c_i}\right) f_U(h)\, I_{\mathbb{H}_1}(h)\, d\tau_1(h)\,,\end{aligned}$$

nas quais utilizamos a expressão da função de verossimilhança obtida no Lema 2.2, a regra de Leibniz para as derivadas de Radon-Nikodym, a expressão de $d\mu_H/d\tau_1$ da Definição 1.12 e a constante C_0 é tal que $\mu_{H|X}(\mathbb{H}_1 \mid x) = 1$. O restante da demonstração depende de alguma álgebra matricial. Seja I a matriz identidade. Uma vez que, por definição, Σ é simétrica, temos que $I = I^\top = (\Sigma\Sigma^{-1})^\top = (\Sigma^{-1})^\top\Sigma^\top = (\Sigma^{-1})^\top\Sigma$. Portanto, temos que $(\Sigma^{-1})^\top = \Sigma^{-1}$. Escreva $l = \log h$. Uma vez que o escalar $l^\top\Sigma^{-1}m$ é igual ao seu transposto $(l^\top\Sigma^{-1}m)^\top = m^\top\Sigma^{-1}l$, temos que $(l-m)^\top\Sigma^{-1}(l-m) = l^\top\Sigma^{-1}l - 2m^\top\Sigma^{-1}l + m^\top\Sigma^{-1}m$. Definindo $d = \Sigma\, c$, temos que

$$\begin{aligned}&\left(\prod_{i=1}^k h_i^{c_i}\right) \exp\left(-\frac{1}{2}(l-m^*)^\top\Sigma^{-1}(l-m^*)\right) \\ &\quad = \exp\left(-\frac{1}{2}\left(-2d^\top\Sigma^{-1}l + l^\top\Sigma^{-1}l - 2m^\top\Sigma^{-1}l + m^\top\Sigma^{-1}m\right)\right) \\ &\quad = C_1 \exp\left(-\frac{1}{2}\left(-2d^\top\Sigma^{-1}l + l^\top\Sigma^{-1}l - 2m^\top\Sigma^{-1}l + m^\top\Sigma^{-1}m\right) + 2m^\top\Sigma^{-1}d + d^\top\Sigma^{-1}d\right)\,,\end{aligned}$$

com $C_1 = \exp\left(-(1/2)\left(-2m^\top\Sigma^{-1}d - d^\top\Sigma^{-1}d\right)\right)$. Defina $m^* = m + d$. Uma vez que o escalar $d^\top\Sigma^{-1}m = (d^\top\Sigma^{-1}m)^\top = m^\top\Sigma^{-1}d$, temos que $(m^*)^\top\Sigma^{-1}m^* = m^\top\Sigma^{-1}m + 2m^\top\Sigma^{-1}d + d^\top\Sigma^{-1}d$. Assim, obtemos que

$$\begin{aligned}&\left(\prod_{i=1}^k h_i^{c_i}\right) \exp\left(-\frac{1}{2}(l-m^*)^\top\Sigma^{-1}(l-m^*)\right) \\ &\qquad = C_1 \exp\left(-\frac{1}{2}\left(l^\top\Sigma^{-1}l - 2(m^*)^\top\Sigma^{-1}l + (m^*)^\top\Sigma^{-1}m^*\right)\right) \\ &\qquad = C_1 \exp\left(-\frac{1}{2}(l-m^*)^\top\Sigma^{-1}(l-m^*)\right)\,.\end{aligned}$$

Usando este resultado na expressão de $\mu_{H|X}$ juntamente com a expressão de f_U obtida na Proposição 1.2, chegamos a

$$\mu_{H|X}(A \mid x) = C_2 \int_A f_{U^*}(h)\, I_{\mathbb{H}_1}(h)\, d\tau_1(h)\,,$$

na qual $C_2 = (C_0\, C_1)/f_{S_\Delta(U)}(1)$ e f_{U^*} é uma densidade de um vetor aleatório $U^* \sim L_k(m^*, \Sigma)$. Concluímos que, dado que $X = x$, o vetor H tem a distribuição das alturas dos degraus de uma densidade aleatória simples $\varphi^* \sim \Delta(m^*, \Sigma)$, que é o resultado desejado. ❖

2.4 Densidades preditivas como estimativas de Bayes

No modelo condicional do Lema 2.2 a esperança *a priori* de φ é uma versão da densidade preditiva *a priori*. Este resultado é o ponto de partida para a escolha de m e Σ. De maneira análoga, a esperança *a posteriori* de φ determina uma versão da densidade preditiva plena, o que torna esta estimativa de Bayes o sumário de maior interesse ao bayesiano.

Proposição 2.5 *Suponha que as variáveis aleatórias $X_1, \ldots, X_{n+1}$ sejam modeladas condicionalmente conforme o Lema 2.2. Denote por μ_{X_i} a distribuição de X_i, para $i = 1, \ldots, n+1$. Por conveniência, utilize as notações $X^{(n)} = (X_1, \ldots, X_n)$ e $x^{(n)} = (x_1, \ldots, x_n) \in \mathbb{R}^k$. Então, para todo $A \in \mathscr{R}$, temos que*

(a) $\mu_{X_i}(A) = \displaystyle\int_A \mathrm{E}[\varphi(y)]\, d\lambda(y)$, *para* $i = 1, \ldots, n+1$;

(b) $\mu_{X_{n+1}|X^{(n)}}(A \mid x^{(n)}) = \displaystyle\int_A \mathrm{E}[\varphi(y) \mid X^{(n)} = x^{(n)}]\, d\lambda(y)$, *quase certamente* $[\mu_{X^{(n)}}]$.

Demonstração. Pela Definição 1.12, temos que

$$\mathrm{E}[\varphi(y)] = \mathrm{E}\left[\sum_{i=1}^{k} H_i\, I_{[t_{i-1},t_i)}(y)\right] = \int_{\mathbb{R}^k} f(y)\, d\mu_H(h)\,,$$

na qual $h \in \mathbb{R}^k$ e $f(y) = \sum_{i=1}^k h_i\, I_{[t_{i-1},t_i)}(y)$, para $y \in \mathbb{R}$. De maneira análoga, temos que

$$\mathrm{E}[\varphi(y) \mid X^{(n)} = x^{(n)}] = \int_{\mathbb{R}^k} f(y)\, d\mu_{H|X^{(n)}}(h \mid x^{(n)})\,.$$

Para o item (a), note que

$$\begin{aligned}\mu_{X_i}(A) &= P\{X_i \in A, H \in \mathbb{R}^k\} = \int_{\mathbb{R}^k} \mu_{X_i|H}(A \mid h)\, d\mu_H(h) \\ &= \int_{\mathbb{R}^k} \left(\int_A f(y)\, d\lambda(y)\right) d\mu_H(h) = \int_A \left(\int_{\mathbb{R}^k} f(y)\, d\mu_H(h)\right) d\lambda(y) = \int_A \mathrm{E}[\varphi(y)]\, d\lambda(y)\,,\end{aligned}$$

nas quais a quarta igualdade é devida ao Teorema de Tonelli. Para o item (b), para cada $B \in \mathscr{R}^n$, temos que

$$P\{X_{n+1} \in A, X^{(n)} \in B\} = \int_B \mu_{X_{n+1}|X^{(n)}}(A \mid x^{(n)})\, d\mu_{X^{(n)}}(x^{(n)})\,.$$

Por outro lado, temos que

$$\begin{aligned}
P\{X_{n+1} \in A, X^{(n)} \in B\} &= P\{X_{n+1} \in A, X^{(n)} \in B, H \in \mathbb{R}^k\} \\
&= \int_{B\times\mathbb{R}^k} \mu_{X_{n+1}|X^{(n)},H}(A \mid x^{(n)}, h)\, d\mu_{X^{(n)},H}(x^{(n)}, h) \\
&= \int_{B\times\mathbb{R}^k} \mu_{X_{n+1}|H}(A \mid h)\, d\mu_{X^{(n)},H}(x^{(n)}, h) \\
&= \int_B \left(\int_{\mathbb{R}^k} \mu_{X_{n+1}|H}(A \mid h)\, d\mu_{H|X^{(n)}}(h \mid x^{(n)}) \right) d\mu_{X^{(n)}}(x^{(n)}) \\
&= \int_B \left(\int_{\mathbb{R}^k} \left(\int_A f(y)\, d\lambda(y) \right) d\mu_{H|X^{(n)}}(h \mid x^{(n)}) \right) d\mu_{X^{(n)}}(x^{(n)}) \\
&= \int_B \left(\int_A \left(\int_{\mathbb{R}^k} f(y)\, d\mu_{H|X^{(n)}}(h \mid x^{(n)}) \right) d\lambda(y) \right) d\mu_{X^{(n)}}(x^{(n)}) \\
&= \int_B \left(\int_A \mathrm{E}[\varphi(y) \mid X^{(n)} = x^{(n)}]\, d\lambda(y) \right) d\mu_{X^{(n)}}(x^{(n)}) \,,
\end{aligned}$$

nas quais a terceira igualdade segue da hipótese de independência condicional e do Teorema B.61 de [14], a quarta igualdade é consequência do Teorema 2.6.4 de [2] e a sexta igualdade é devida ao Teorema de Tonelli. Comparando as duas expressões anteriores para $P\{X_{n+1} \in A, X^{(n)} \in B\}$, obtemos o resultado desejado. ❖

No próximo capítulo exploramos através de simulações estocásticas as distribuições *a priori* e *a posteriori* de uma densidade aleatória simples φ com o auxílio do Teorema 2.4.

Capítulo 3

Simulações estocásticas

De posse da teoria das densidades aleatórias simples e do modelo condicional desenvolvidos nos dois capítulos anteriores, exploramos neste capítulo as distribuições *a priori* e *a posteriori* de uma densidade aleatória simples através de simulações estocásticas. Como resultado desta análise, obtemos soluções bayesianas para o problema de estimação de densidade.

O capítulo está organizado da seguinte maneira. Uma vez que as explorações das distribuições *a priori* e *a posteriori* são feitas através de simulações estocásticas baseadas no algoritmo de Metropolis-Hastings, descrevemos este algoritmo brevemente na seção 3.1 e mostramos como as realizações de uma densidade aleatória simples geradas via simulação nos permitem obter uma estimativa e um conjunto crível com nível de credibilidade predeterminado. O problema geral da especificação da matriz de covariâncias existente na distribuição *a priori* de uma densidade aleatória simples é tratado na seção 3.2. Ainda nesta seção, discutimos algumas questões de análise numérica pertinentes às implementações das simulações. Apresentamos na seção 3.3 resultados das simulações utilizando dados gerados a partir de diferentes distribuições. Ao crescermos progressivamente o tamanho das amostras dos dados simulados, registramos o comportamento assintótico, no sentido bayesiano, da distribuição *a posteriori*. Na seção 3.4, estendemos o modelo considerando os casos em que a partição não é conhecida *a priori* e propomos um critério de escolha da mesma a partir da informação contida na amostra dos dados analisados. Finalizamos o capítulo na seção 3.5 mostrando como obter, com o auxílio da Teoria da Decisão, estimativas que são densidades suaves de uma classe adequada.

3.1 Algoritmo de Metropolis-Hastings

Precisamos de métodos que possibilitem a exploração das distribuições *a priori* e *a posteriori* de uma densidade aleatória simples. Isto será feito a partir de uma amostra de realizações da densidade aleatória simples obtida via simulação. Para este fim, utilizamos o algoritmo de Metropolis-Hastings, cujo propósito é definir uma cadeia de Markov de tal sorte que sua distribuição de equilíbrio possua uma densidade predeterminada em relação a uma certa medida dominante. Temos a seguir uma breve descrição do algoritmo já adaptado às nossas necessidades, utilizando as notações introduzidas na Definição 1.12.

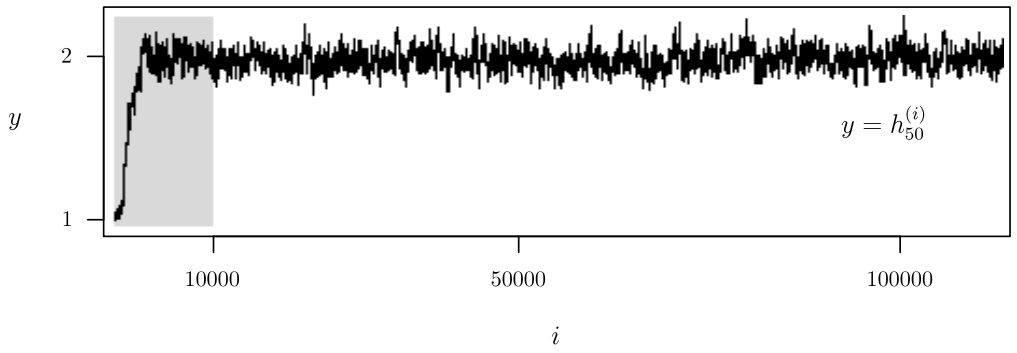


Figura 3.1: Exemplo de monitoramento de uma das coordenadas da cadeia de Markov. A região cinza corresponde ao período de aquecimento.

Para uma sequência de vetores aleatórios $\{H^{(i)}\}_{i\geq 0}$ em $\mathbb{H}_1$, o algoritmo de Metropolis-Hastings inicia a sequência em um ponto $x_0 \in \mathbb{H}_1$. Para cada $i \geq 0$, dado que $H^{(i)} = x$, geramos um ponto $y \in \mathbb{H}_1$ a partir de uma distribuição cuja densidade condicional em relação a τ_1 é $q(\,\cdot \mid x)$. Com probabilidade

$$\alpha(x, y) = \min\left\{\frac{f_H(y)}{f_H(x)}\frac{q(x \mid y)}{q(y \mid x)},\ 1\right\},$$

aceitamos a proposta y, fazendo $H^{(i+1)} = y$, ou, com probabilidade $1-\alpha(x, y)$, rejeitamos o ponto y proposto e fazemos $H^{(i+1)} = x$. Esta expressão incrivelmente simples para a probabilidade de aceitação das propostas implica que $\{H^{(i)}\}_{i\geq 0}$ é uma cadeia de Markov cuja distribuição de equilíbrio possui densidade f_H em relação à medida dominante τ_1. Note que a expressão de $\alpha(x, y)$ depende apenas da razão $f_H(y)/f_H(x)$, de modo que não temos que conhecer o valor da constante de normalização de f_H presente na Definição 1.12. Nossa implementação é baseada em um caso particular do algoritmo de Metropolis-Hastings, conhecido como o algoritmo de Metropolis com Passeio Aleatório, que consiste em considerar propostas obtidas a partir de um pequeno deslocamento aleatório do valor atual da cadeia, de maneira que a densidade condicional possua a simetria $q(x \mid y) = q(y \mid x)$. Uma discussão da teoria necessária para justificar o algoritmo de Metropolis-Hastings e suas variantes pode ser encontrada em [13].

O algoritmo de Metropolis-Hastings é implementado gerando-se valores da cadeia de Markov por um período inicial suficientemente longo, denominado *período de aquecimento*, a partir do qual assumimos que a cadeia está em equilíbrio e passamos a obter realizações dependentes com a distribuição desejada. Na Figura 3.1 temos um exemplo do monitoramento em uma das simulações dos valores de uma das coordenadas do vetor das alturas aleatórias de uma densidade aleatória simples. Note a mudança qualitativa da evolução após o período de aquecimento, indicado pela cor cinza.

Vejamos como as realizações de uma densidade aleatória simples obtidas através do algoritmo de Metropolis-Hastings nos permitem obter sumários da sua distribuição. Em particular, queremos obter uma estimativa e um conjunto crível com nível de credibilidade predeterminado.

Técnicas mais tradicionais de simulação, tais quais Aceitação-Rejeição ou Amostragem por Importância [13], produzem realizações independentes do objeto aleatório considerado. A metodologia de simulação através de cadeias de Markov difere destas técnicas por haver dependência entre as realizações obtidas. Apesar desta dependência entre os valores sucessivos da cadeia de Markov, o Teorema Ergódico [13] dá a chave para calcularmos as esperanças em que estamos interessados.

Em primeiro lugar, obtemos a esperança do vetor das alturas dos degraus de uma densidade aleatória simples a partir da convergência quase certa da média ergódica

$$\frac{1}{N}\sum_{i=0}^{N}\left(H_1^{(i)},\ldots,H_k^{(i)}\right) \xrightarrow[N\to\infty]{} \left(\mathrm{E}[H_1],\ldots,\mathrm{E}[H_k]\right).$$

Vale lembrar que já provamos na Proposição 2.5 que esta esperança é um sumário particularmente importante, pois determina uma densidade preditiva do nosso modelo condicional.

Isso posto, necessitamos de um sumário que indique a concentração da distribuição da densidade aleatória simples em torno desta estimativa.

Dados dois pontos $x, y \in \mathbb{H}_1$, defina a distância $d(x,y) = \max_{1\leq i\leq k}|x_i - y_i|$. A partir da estimativa $\hat{h} = (\mathrm{E}[H_1],\ldots,\mathrm{E}[H_k])$, defina

$$B(\hat{h},\epsilon) = \left\{h \in \mathbb{H}_1 : d(\hat{h},h) < \epsilon\right\},$$

para $\epsilon > 0$.

Definimos nosso conjunto crível, com nível de credibilidade $\gamma \in (0,1)$, tomando o menor $\epsilon > 0$ tal que $P\{\omega : H(\omega) \in B(\hat{h},\epsilon)\} = \gamma$.

Esta probabilidade é obtida via simulação valendo-se da convergência quase certa da média ergódica

$$\frac{1}{N}\sum_{i=0}^{N} I_{B(\hat{h},\epsilon)}\left(H^{(i)}\right) \xrightarrow[N\to\infty]{} \mathrm{E}\left[I_{B(\hat{h},\epsilon)}(H)\right] = P\left\{\omega : H(\omega) \in B(\hat{h},\epsilon)\right\}.$$

Estes conceitos são suficientes para nossas implementações das simulações. Na próxima seção, discutimos como escolher a matriz de covariâncias Σ que cumpre o papel de um dos parâmetros da Definição 1.12.

3.2 Estrutura de covariâncias

Utilizando a notação introduzida no início da seção 1.1, dado um intervalo $[a, b]$ e sua partição nos subintervalos determinados por $\Delta = \{a = t_0, t_1, \ldots, t_k = b\}$, para especificar a distribuição *a priori* de uma densidade aleatória simples precisamos escolher o vetor m e a matriz de covariâncias Σ que ocorrem na Definição 1.12. Nesta seção, discutimos os aspectos relacionados à escolha de Σ.

Em princípio, a única restrição existente na escolha de Σ é que esta matriz precisa ser simétrica e definida positiva, ou seja, que $\Sigma = \Sigma^\top$ e $x^\top \Sigma\, x > 0$, para todo $x \in \mathbb{R}^k$ não nulo. A ideia é construir Σ a partir de uma função de covariância $C : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ que reflita as características desejadas *a priori* da densidade aleatória simples, levando em conta a estrutura da partição determinada por Δ.

A função de covariância C precisa ser simétrica em seus argumentos e definida positiva, no sentido de que, para quaisquer $x_1, \ldots, x_k \in \mathbb{R}$, a matriz $\Sigma = (C(x_i, x_j))$ seja definida positiva. Uma condição necessária e suficiente para que C da forma $C(x, y) = \Psi(x - y)$ seja definida positiva é dada pelo Teorema de Bochner [3], que afirma que C é uma função real definida positiva se e somente se Ψ é a função característica de uma variável aleatória com densidade simétrica cujo suporte é a reta real.

Definição 3.1 Dada uma função de covariância C, definimos a matriz de covariâncias Σ calculando as covariâncias entre os pontos médios dos subintervalos que formam a partição determinada por Δ. Explicitamente, escrevendo $\Sigma = (\sigma_{ij})$, definimos

$$\sigma_{ij} = C\left(\frac{t_{i-1} + t_i}{2}, \frac{t_{j-1} + t_j}{2}\right),$$

para $i, j = 1, \ldots, k$. Dizemos que Σ é a *matriz de covariâncias induzida por* C.

Nos exemplos desta tese utilizamos matrizes de covariâncias induzidas pela família de funções de covariância gaussianas definida por $C_{\rho,\theta}(x, y) = \rho\, e^{-\theta\,(x-y)^2}$, com $\rho > 0$ e $\theta > 0$.

O tratamento computacional das matrizes de covariâncias induzidas por esta família de funções de covariância gaussianas demanda uma breve discussão sobre dois conceitos de análise numérica, o que faremos a seguir. Uma discussão mais detalhada pode ser encontrada em [11].

Para um número real x, escreva $\mathrm{fl}(x)$ para a representação em ponto flutuante de x. O *épsilon de máquina* é definido como o maior número ϵ_m tal que $\mathrm{fl}(1 + \epsilon_m) = 1$. Dada uma

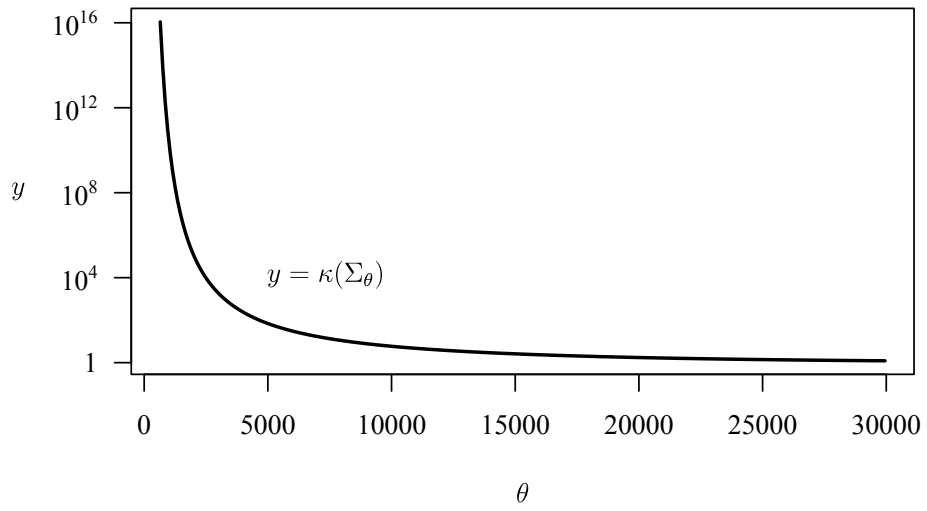


Figura 3.2: Números de condição $\kappa(\Sigma_\theta)$ em função de θ para as duas partições utilizadas nos exemplos da seção 3.3.

matriz Σ não singular, o *número de condição* $\kappa(\Sigma)$ é uma medida da estabilidade numérica de operações que envolvem Σ, tal qual o cálculo da sua inversa Σ^{-1}. Operações numéricas com matrizes para as quais o número de condição é muito pequeno, da ordem de $1/\epsilon_m$, produzem resultados incorretos. É ideal, do ponto de vista do cálculo numérico, que $\kappa(\Sigma)$ fique próximo de 1. Em nossa arquitetura, ϵ_m é da ordem de 10^{-16}. Diversas bibliotecas de cálculo numérico contém rotinas que permitem estimar $\kappa(\Sigma)$.

Por conveniência, até o final deste capítulo utilizamos o ponto como separador decimal.

Nos exemplos da próxima seção consideramos dois espaços amostrais. Inicialmente, nossas observações assumem valores no intervalo [0,1], que dividimos em subintervalos de comprimento 0.01, o que corresponde a $\Delta = \{0, 0.01, \ldots, 0.99, 1\}$. Posteriormente, analisamos amostras de dados com valores no intervalo $[-0.5, 0.5]$, que também é dividido em subintervalos de comprimento 0.01, correspondendo a $\Delta = \{-0.5, -0.49, \ldots, 0.49, 0.5\}$.

Para as duas partições, utilizamos as matrizes de covariâncias Σ_θ induzidas a partir da família de funções de covariância gaussianas, tomando $\rho = 1$, e determinamos numericamente os valores de $\kappa(\Sigma_\theta)$ para diversos valores de θ. O gráfico da Figura 3.2 apresenta os resultados desta análise, que são os mesmos para as duas partições consideradas e não dependem do valor particular de ρ que escolhemos.

Devido a estes resultados, nos exemplos da seção 3.3 consideramos apenas valores de θ maiores do que o valor crítico $\theta_c = 20\,000$.

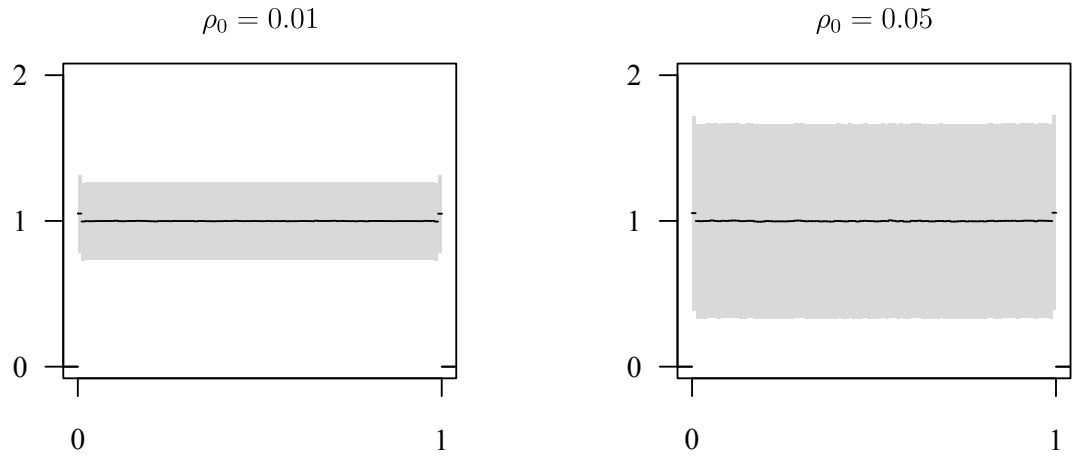


Figura 3.3: Efeito do valor de ρ_0 na concentração da distribuição *a priori*. As curvas em preto são as esperanças *a priori* e as regiões em cinza são conjuntos críveis *a priori* com nível de credibilidade 95%.

3.3 Assintótica bayesiana

O objetivo desta seção é observar o comportamento assintótico, no sentido bayesiano, da distribuição *a posteriori* de uma densidade aleatória simples. Ou seja, no caso de dados simulados, em que conhecemos a densidade a partir da qual estes foram gerados, queremos observar a concentração da distribuição *a posteriori* em torno da densidade simples que aproxima a densidade de origem conforme crescemos o tamanho das amostras.

Note que, em virtude da propriedade de fechamento enunciada no Teorema 2.4, as simulações das distribuições *a priori* e *a posteriori* feitas nesta seção dependem, essencialmente, de uma mesma implementação computacional. A diferença reside na utilização de m ou m^* na expressão da densidade de H especificada na Definição 1.12.

Almejando um pouco mais de generalidade na escolha dos parâmetros que definem a distribuição *a priori*, o que pode ser útil em eventuais aplicações, consideramos matrizes de covariância induzidas pela família de funções de covariância gaussianas com ρ fixado em um certo valor ρ_0, mas com Θ aleatório. Ao fazermos isto, as estimativas são obtidas com o auxílio do Teorema da Probabilidade Total (veja [14], Teorema B.70). Por exemplo, temos que $\hat{h}_i = \mathrm{E}[H_i] = \mathrm{E}[\mathrm{E}[H_i \mid \Theta]]$, para $i = 1, \ldots, k$.

Usamos as seguintes notações. Se Y é uma variável aleatória com distribuição Gama, com parâmetros α e β, escrevemos $Y \sim Ga(\alpha, \beta)$. Para Y com distribuição Beta, com parâmetros α e β, escrevemos $Y \sim Be(\alpha, \beta)$. Se Y tem distribuição normal com média μ e variância σ^2,

escrevemos $Y \sim N(\mu, \sigma^2)$. Denotamos misturas por $Y \sim \sum_{i=1}^{m} p_i \, Be(\alpha_i, \beta_i)$, por exemplo, em que $p_i \geq 0$, para $i = 1, \ldots, m$, e $\sum_{i=1}^{m} p_i = 1$.

Antes de apresentarmos os resultados dos exemplos, vejamos qual o efeito da escolha do valor de ρ_0 na distribuição *a priori*. Escolhemos $\alpha = 2$, $\beta = 0.001$, $\theta_c = 20\,000$ e definimos $\Theta = Y + \theta_c$, com $Y \sim Ga(\alpha, \beta)$. Para o espaço amostral $[0, 1]$ e $\Delta = \{0, 0.01, 0.02, \ldots, 0.98, 0.99, 1\}$, com todos os m_i's da Definição 1.12 iguais a 1, temos na Figura 3.3 os sumários *a priori* da distribuição de φ para $\rho_0 = 0.01$ e $\rho_0 = 0.05$, supondo que, dado que $\Theta = \theta$, temos que $\varphi \sim \Delta(m, \Sigma_\theta)$, em que Σ_θ é induzida pela família de funções de covariância gaussianas com $\rho = \rho_0$.

Foram simulados 30 valores de Θ. Para cada um destes valores a cadeia foi gerada com um período de aquecimento de 10 000 iterações, a partir do qual utilizamos os valores das 100 000 iterações seguintes. Note que, conforme aumentamos o valor de ρ_0, a distribuição *a priori* fica menos concentrada em torno da esperança *a priori*. Note também a uniformidade da esperança *a priori*.

Exemplo 3.2 (Distribuição Triangular) Sejam Y_1 e Y_2 variáveis aleatórias independentes absolutamente contínuas com distribuição uniforme no intervalo $[0, 1]$. Dizemos que $Z = (Y_1 + Y_2)/2$ tem distribuição triangular. Escolha, como antes, $\Delta = \{0, 0.01, 0.02, \ldots, 0.98, 0.99, 1\}$, fazendo todos os m_i's da Definição 1.12 iguais a 1. Escolha $\alpha = 2$, $\beta = 0.001$, $\theta_c = 20\,000$ e defina $\Theta = Y + \theta_c$, com $Y \sim Ga(\alpha, \beta)$. Escolha $\rho_0 = 0.05$ e suponha que, dado que $\Theta = \theta$, temos que $\varphi \sim \Delta(m, \Sigma_\theta)$, em que Σ_θ é induzida pela família de funções de covariância gaussianas com $\rho = \rho_0$. Com estas escolhas, os sumários *a priori* da distribuição de φ são aqueles apresentados no segundo gráfico da Figura 3.3. Supondo que os observáveis são modelados condicionalmente conforme descrito no Lema 2.2, a Figura 3.4 apresenta os sumários das distribuições *a posteriori* de φ para diversos tamanhos de amostra, quando os dados analisados são gerados a partir de uma distribuição triangular. Note a concentração da distribuição *a posteriori* em torno da densidade simples que aproxima a densidade de origem dos dados conforme crescemos o tamanho das amostras. ❖

Exemplo 3.3 (Mistura de Betas) Considere o caso em que os dados são gerados a partir da mistura

$$\frac{1}{3} \cdot Be(1, 10) + \frac{1}{3} \cdot Be(10, 10) + \frac{1}{3} \cdot Be(30, 5)\,,$$

e suponha que modelamos Θ e φ *a priori* como no Exemplo 3.2. Temos na Figura 3.5 os sumários das distribuições *a posteriori* de φ para diversos tamanhos de amostra. Novamente, o estreitamento da banda definida pelo conjunto crível indica a concentração da distribuição *a posteriori* em torno da densidade simples que aproxima a densidade de origem dos dados. ❖

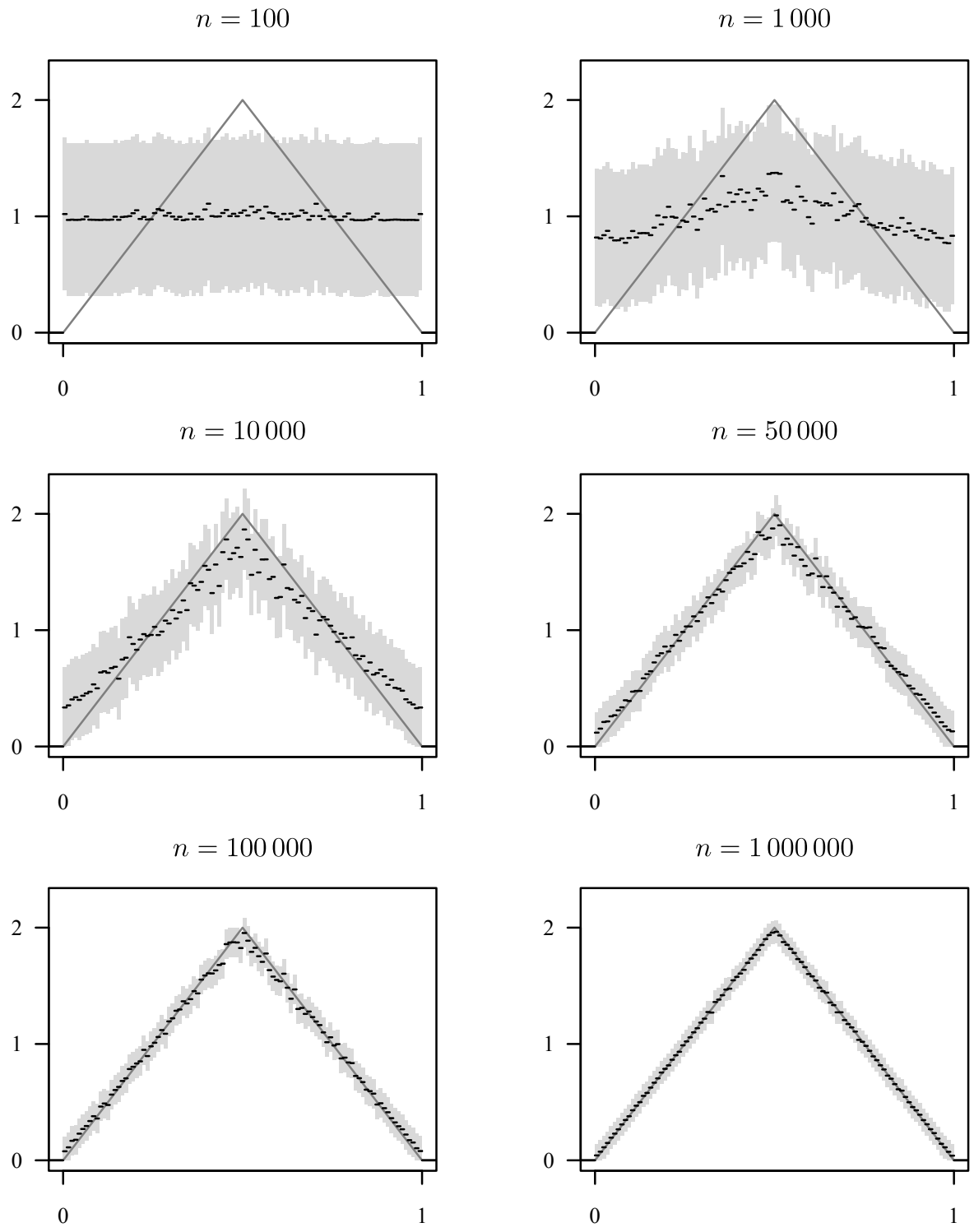


Figura 3.4: Sumários *a posteriori* para o Exemplo 3.2. Em cada gráfico, a densidade simples em preto é a estimativa $\hat{\varphi}$, a região cinza é um conjunto crível com nível de credibilidade de 95% e a curva em cinza escuro é a densidade a partir da qual os dados das amostras foram simulados.

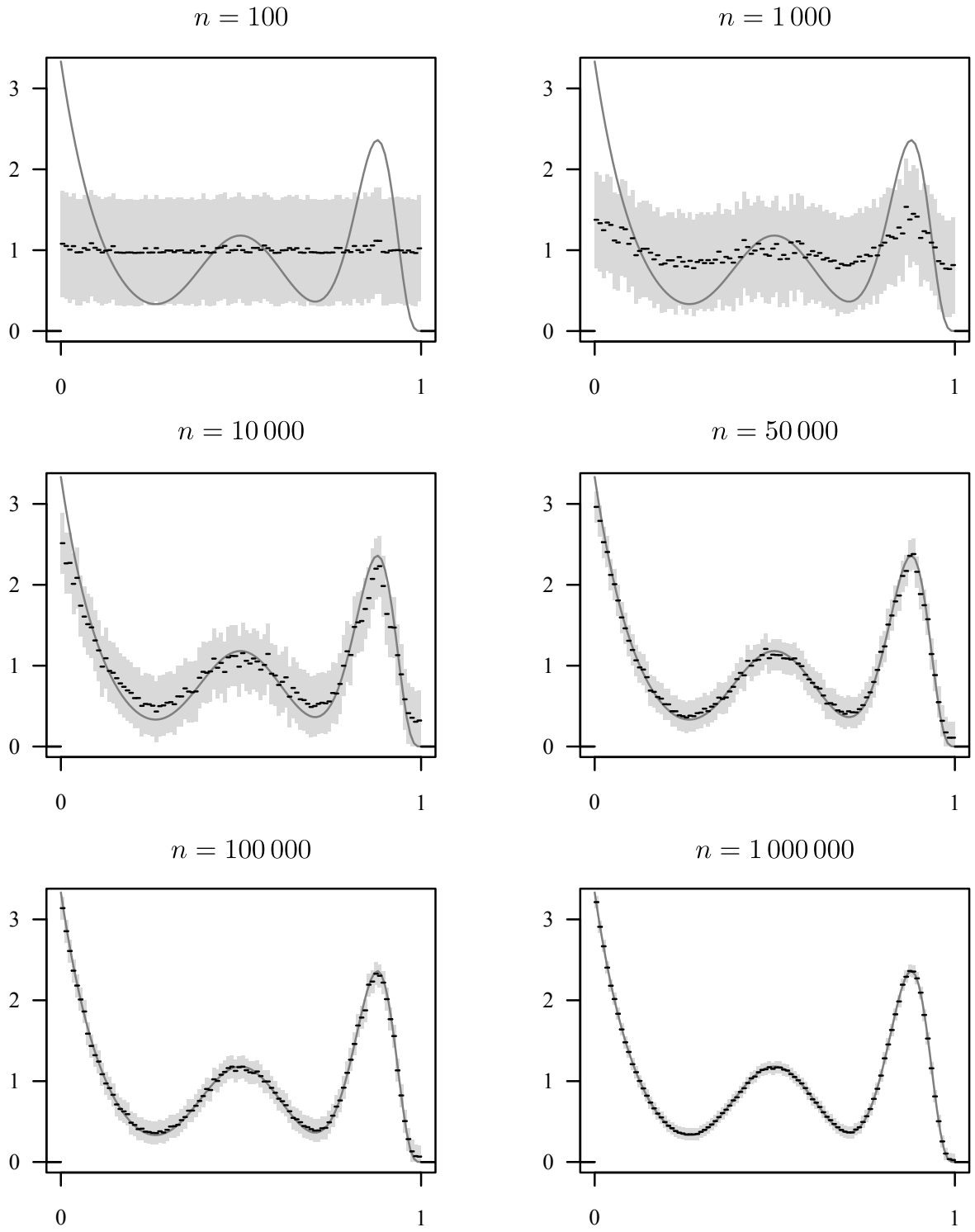


Figura 3.5: Sumários *a posteriori* para o Exemplo 3.3. Em cada gráfico, a densidade simples em preto é a estimativa $\hat{\varphi}$, a região cinza é um conjunto crível com nível de credibilidade de 95% e a curva em cinza escuro é a densidade a partir da qual os dados das amostras foram simulados.

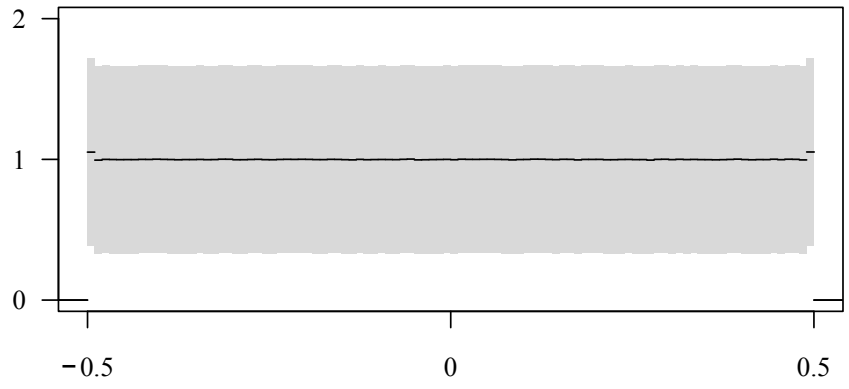


Figura 3.6: Sumários *a priori* para o Exemplo 3.4. A curva em preto é a esperança *a priori*. A região cinza é um conjunto crível com nível de credibilidade de 95%.

O próximo exemplo permite avaliar o comportamento do modelo quando o suporte não é limitado. A distribuição *a priori* é especificada supondo-se um truncamento do suporte das observações.

Exemplo 3.4 (Mistura de Normais) Suponha que o espaço amostral é o intervalo $[-0.5, 0.5]$ e que $\Delta = \{-0.5, -0.49, \ldots, 0.49, 0.5\}$. Escolha a mesma distribuição *a priori* para Θ utilizada no Exemplo 3.2. Escolha todos os m_i's iguais a 1 e faça $\rho_0 = 0.05$. Novamente, suponha que, dado que $\Theta = \theta$, temos que $\varphi \sim \Delta(m, \Sigma_\theta)$, em que Σ_θ é induzida pela família de funções de covariância gaussianas com $\rho = \rho_0$. Os sumários *a priori* de φ são apresentados na Figura 3.6. Gerando amostras de dados a partir da mistura

$$\frac{2}{3} \cdot N(-1/7, 1/12) + \frac{1}{3} \cdot N(1/7, 1/12)\,,$$

obtivemos os resultados da Figura 3.7, que indicam novamente o comportamento assintótico da distribuição *a posteriori*. ❖

Até agora, consideramos apenas situações em que a partição utilizada na definição da distribuição *a priori* da densidade aleatória simples é fixada de antemão. Na próxima seção estendemos esta análise para o caso em que Δ é aleatório e propomos um critério que permite escolher uma partição de uma determinada classe a partir da informação contida na amostra dos dados analisados.

3.4 Um critério para a escolha da partição

Nosso próximo passo é estender a definição da distribuição das densidades aleatórias simples para os casos em que a partição não é conhecida previamente. Tal extensão pode ser motivada

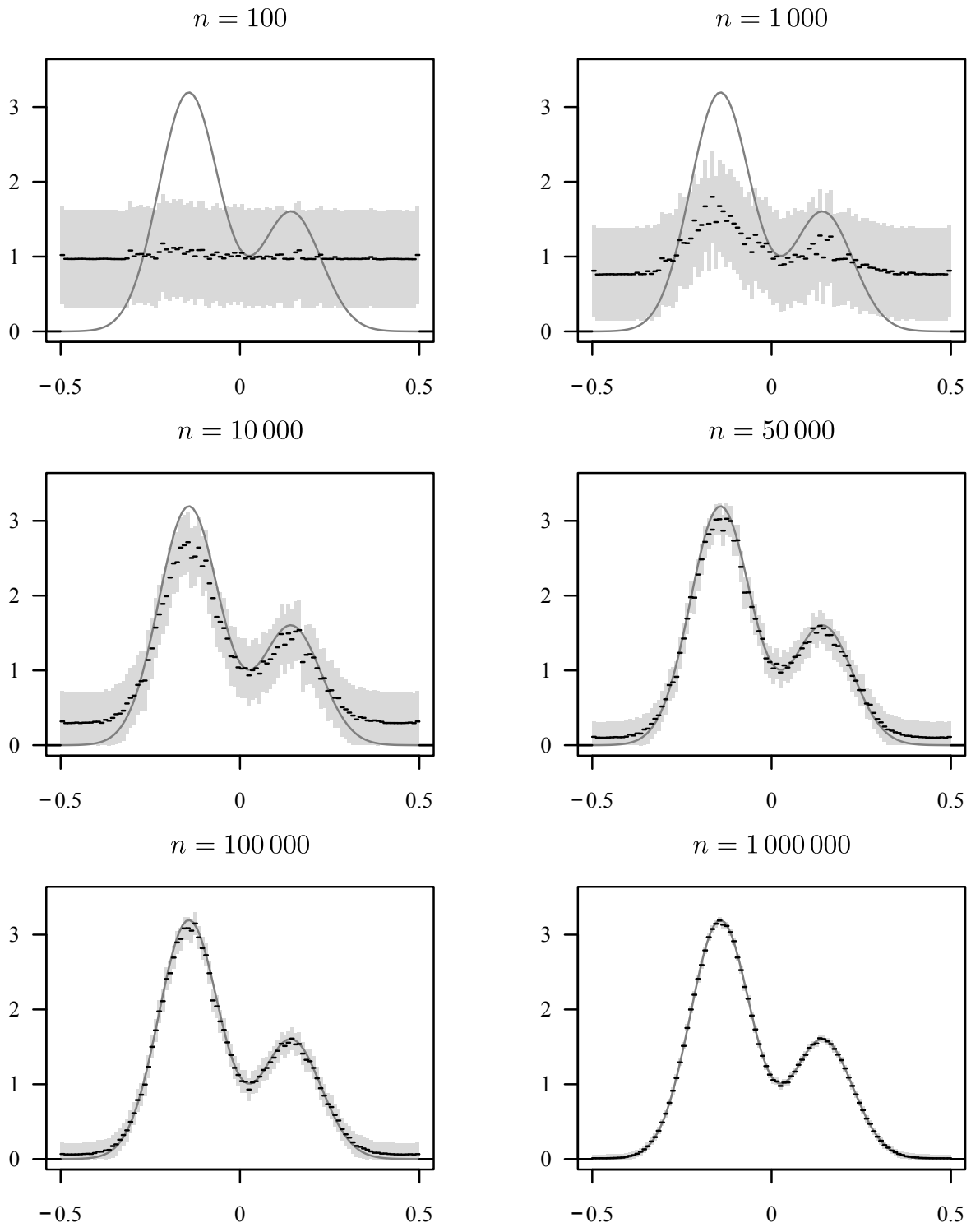


Figura 3.7: Sumários *a posteriori* para o Exemplo 3.4. Em cada gráfico, a densidade simples em preto é a estimativa $\hat{\varphi}$, a região cinza é um conjunto crível com nível de credibilidade de 95% e a curva em cinza escuro é a densidade a partir da qual os dados das amostras foram simulados.

intuitivamente da seguinte maneira.

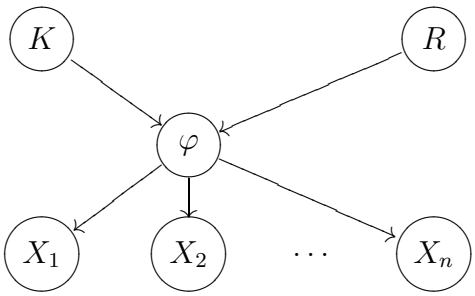
Suponha que escolhemos uma "partição" extremamente grosseira do intervalo $[a, b]$ contendo apenas um subintervalo (o próprio intervalo $[a, b]$). É claro que, neste caso, sejam quais forem os dados observados, as esperanças *a priori* e *a posteriori* serão ambas densidades uniformes. No outro extremo, considere o caso em que escolhemos uma partição ultra refinada, com um número enorme de subintervalos. Como podemos observar analisando os resultados das simulações da seção anterior, neste caso, nossa opinião *a priori* só seria alterada substancialmente se tivéssemos em mãos uma amostra gigantesca de dados. Por estes motivos, gostaríamos de construir uma extensão do modelo que, de certo modo, escolhesse automaticamente a partição que fosse mais adequada para os dados que queremos analisar.

Vamos considerar o caso em que temos uma família de partições uniformes do intervalo $[a, b]$, ou seja, os subintervalos que formam cada partição terão todos o mesmo comprimento. Cada partição desta família será descrita pela variável aleatória K, que determina o número de subintervalos da partição. Uma vez que o parâmetro ρ da família de funções de covariância utilizada para induzir a matriz de covariâncias, necessária para definirmos a distribuição da densidade aleatória simples, pode ter significados distintos para partições diferentes, também trataremos este parâmetro como uma variável aleatória positiva R.

Suponha que temos θ fixado. Explicitamente, estamos considerando o seguinte modelo hierárquico: K e R são independentes *a priori*. Dado que $K = k$ e $R = \rho$, escolhemos a partição uniforme do intervalo $[a, b]$ determinada por

$$\Delta = \left\{ a, a + \frac{b-a}{k}, a + \frac{2(b-a)}{k}, \ldots, a + \frac{(k-1)(b-a)}{k}, b \right\},$$

induzimos a matriz de covariâncias $\Sigma_{\rho,\theta}$ a partir da família de funções de covariâncias gaussianas, e fazemos $\varphi \sim \Delta(m, \Sigma_{\rho,\theta})$. Finalmente, os observáveis são modelados como no Lema 2.2. Esta hierarquia é descrita graficamente pela seguinte rede bayesiana.



Aqui não faremos a análise bayesiana completa deste modelo hierárquico, o que demandaria a especificação das distribuições *a priori* de K e R. Seguiremos uma via empírica (veja [14], seção

8.4), calculando a função de verossimilhança de K e R e escolhendo valores destes parâmetros que a maximizem.

Vamos usar as seguintes notações: seja $\mu_K = P \circ K^{-1}$ sobre $(\mathbb{N}, 2^{\mathbb{N}})$ a distribuição de K e seja $\mu_R = P \circ R^{-1}$ sobre $(\mathbb{R}, \mathscr{R})$ a distribuição de R. Denote por $\mu_{K,R}$ a distribuição conjunta de K e R, que pela independência de K e R é igual à medida produto $\mu_K \times \mu_R$, e seja $\mu_{K,R,H}$ a distribuição conjunta de K, R e H.

Proposição 3.5 *No modelo hierárquico descrito acima, temos que* $\mu_{X|K,R}(\,\cdot \mid k, \rho) \ll \lambda_n$, *quase certamente* $[\mu_{K,R}]$, *com derivada de Radon-Nikodym*

$$\frac{d\mu_{X|K,R}}{d\lambda_n}(x \mid k, \rho) = f_{X|K,R}(x \mid k, \rho) = \int_{\mathbb{R}^k} f_{X|H}(x \mid h)\, d\mu_{H|K,R}(h \mid k, \rho)\,,$$

para a $f_{X|H}$ *definida no Lema 2.2.*

Demonstração. Sejam $A \in \mathscr{R}^n$ e $B \in 2^{\mathbb{N}} \otimes \mathscr{R}$. Pela definição de distribuição condicional 1.4, temos que

$$P\{X \in A, (K,R) \in B\} = \int_B \mu_{X|K,R}(A \mid k, \rho)\, d\mu_{K,R}(k, \rho)\,.$$

Por outro lado, por argumentos similares aos usados na demonstração do item (b) da Proposição 2.5, temos que

$$\begin{aligned}
P\{X \in A, (K,R) \in B\} &= P\{X \in A, (K,R) \in B, H \in \mathbb{R}^k\} \\
&= \int_{B \times \mathbb{R}^k} \mu_{X|K,R,H}(A \mid k, \rho, h)\, d\mu_{K,R,H}(k, \rho, h) \\
&= \int_{B \times \mathbb{R}^k} \mu_{X|H}(A \mid h)\, d\mu_{K,R,H}(k, \rho, h) \\
&= \int_B \left(\int_{\mathbb{R}^k} \mu_{X|H}(A \mid h)\, d\mu_{H|K,R}(h \mid k, \rho) \right) d\mu_{K,R}(k, \rho) \\
&= \int_B \left(\int_{\mathbb{R}^k} \left(\int_A f_{X|H}(x \mid h)\, d\lambda_n(x) \right) d\mu_{H|K,R}(h \mid k, \rho) \right) d\mu_{K,R}(k, \rho) \\
&= \int_B \left(\int_A \left(\int_{\mathbb{R}^k} f_{X|H}(x \mid h)\, d\mu_{H|K,R}(h \mid k, \rho) \right) d\lambda_n(x) \right) d\mu_{K,R}(k, \rho)\,.
\end{aligned}$$

Comparando as duas expressões para $P\{X \in A, (K,R) \in B\}$, temos que

$$\mu_{X|K,R}(A \mid k, \rho) = \int_A \left(\int_{\mathbb{R}^k} f_{X|H}(x \mid h)\, d\mu_{H|K,R}(h \mid k, \rho) \right) d\lambda_n(x)\,,$$

quase certamente $[\mu_{K,R}]$. O resultado segue. ❖

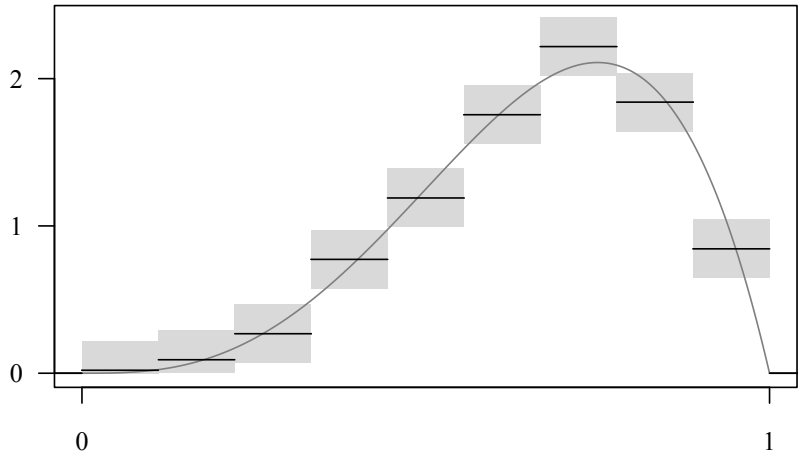


Figura 3.8: Sumários *a posteriori* para o Exemplo 3.6. A densidade simples em preto é a estimativa $\hat{\varphi}$, a região cinza é um conjunto crível com nível de credibilidade de 95% e a curva em cinza escuro é a densidade a partir da qual os dados das amostras foram simulados.

Definindo a verossimilhança de K e R por $L_x(k, \rho) = f_{X|K,R}(x \mid k, \rho)$, obtemos as estimativas $(\hat{k}, \hat{\rho}) = \arg\max_{k,\rho} L_x(k, \rho)$ e utilizamos estes valores na definição da distribuição *a priori* da densidade aleatória simples, determinando os sumários *a posteriori* como fizemos na seção anterior.

Exemplo 3.6 Com uma amostra de 2 000 dados simulados a partir de uma distribuição $Be(4, 2)$, encontramos o máximo da verossimilhança de K e R no ponto $(\hat{k}, \hat{\rho}) = (9, 1.43)$. Na Figura 3.8 temos os sumários *a posteriori* obtidos utilizando estes valores para a especificação da distribuição *a priori* da densidade aleatória simples. Além disto, temos no primeiro gráfico da Figura 3.9 a função de distribuição $\hat{F}$ correspondente à densidade estimada *a posteriori*. A título de comparação, alguns dos quantis desta função de distribuição estimada $\hat{F}$ foram graficados contra os da distribuição geradora F_0 no segundo gráfico da Figura 3.9. ❖

3.5 Estimativas suaves

A Definição 1.12 implica que a distribuição *a priori* de uma densidade aleatória simples está concentrada em uma classe de densidades descontínuas. Além disso, a esperança de uma densidade aleatória simples é sempre uma densidade descontínua, conforme ilustrado por todas as estimativas obtidas nos exemplos das seções 3.3 e 3.4.

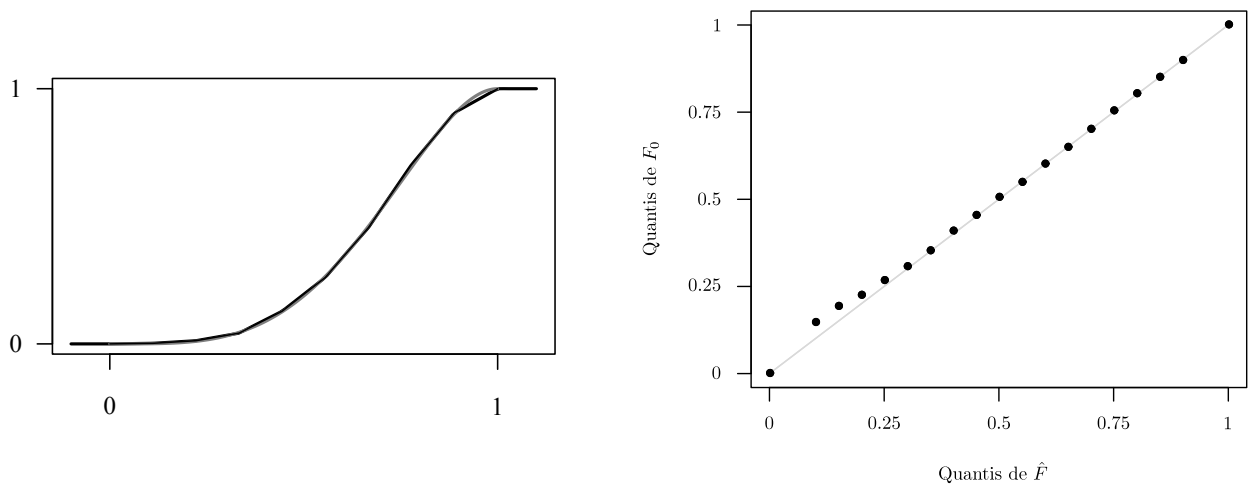


Figura 3.9: Exemplo 3.6. No gráfico da esquerda a curva em preto é a função de distribuição estimada $\hat{F}$ e a curva em cinza é a função de distribuição F_0 a partir da qual os dados foram gerados. No gráfico da direita temos a comparação de alguns dos quantis de $\hat{F}$ e F_0.

Nesta seção, queremos obter estimativas que sejam densidades diferenciáveis de uma classe predeterminada, sem alterar a definição da nossa distribuição *a priori*. No entanto, não desejamos que tais estimativas suaves sejam obtidas por uma mera interpolação *ad hoc* da esperança da densidade aleatória simples. Nosso objetivo é obter estimativas suaves a partir de primeiros princípios.

O conflito aparente entre uma distribuição *a priori* concentrada em uma classe de densidades descontínuas e a busca de uma estimativa que seja uma densidade diferenciável é dissipado se nos lembrarmos dos dois aspectos proeminentes da Inferência bayesiana: a representação de nossas incertezas através de probabilidades e a escolha de nossas ações através dos critérios da Teoria da Decisão.

Nossa proposta é obter estimativas suaves como soluções de um problema de decisão em que os estados da natureza são realizações de uma densidade aleatória simples e as ações são densidades suaves de uma classe que iremos determinar.

A proposição seguinte formula o problema de decisão, especificando a função de perda e calculando a estimativa de Bayes. Em virtude do fechamento sob amostragem que demonstramos no Teorema 2.4, é suficiente resolver o problema de decisão sem dados. Como antes, nosso espaço amostral é o intervalo $[a, b]$, que particionamos de acordo com algum Δ. Para uma densidade f em relação à medida de Lebesgue, denotamos sua norma L_2 por $\|f\|_2 = \left(\int f^2 d\lambda\right)^{1/2}$.

Proposição 3.7 *Para $N \geq 1$, sejam $g_1, \ldots, g_N$ densidades em relação à medida de Lebesgue, com suporte no intervalo $[a, b]$, tais que $\|g_i\|_2 < \infty$, e seja $\mathscr{D}$ a classe das densidades da forma $\sum_{i=1}^N \alpha_i\, g_i$, com $\alpha_i \geq 0$, para $i = 1, \ldots, N$, e $\sum_{i=1}^N \alpha_i = 1$. Seja $\varphi \sim \Delta(m, \Sigma)$ e defina $\mathscr{S}$ como a classe das densidades simples que são realizações de φ. Defina a função de perda $L : \mathscr{S} \times \mathscr{D} \to \mathbb{R}$ por*

$$L(s, f) = \|s - f\|_2^2 = \int_a^b \left(s(x) - f(x)\right)^2 \, d\lambda(x)\,.$$

Então, a decisão de Bayes é $\hat{\varphi} = \sum_{i=1}^N \hat{\alpha}_i\, g_i$, na qual os $\hat{\alpha}_i$ minimizam globalmente a forma quadrática

$$Q = \sum_{i,j=1}^N \alpha_i \alpha_j\, M_{ij} - \sum_{i=1}^N \alpha_i\, J_i\,,$$

sujeita às restrições $\alpha_i \geq 0$, para $i = 1, \ldots, N$, e $\sum_{i=1}^N \alpha_i = 1$, com as definições

$$M_{ij} = \int_a^b g_i(x) g_j(x)\, d\lambda(x) \qquad e \qquad J_i = 2 \int_a^b g_i(x) \mathrm{E}[\varphi(x)]\, d\lambda(x)\,.$$

Demonstração. Pelo Teorema de Tonelli, a perda esperada é

$$\mathrm{E}[L(\varphi, f)] = \int_a^b f^2(x)\, d\lambda(x) - 2 \int_a^b f(x) \mathrm{E}[\varphi(x)]\, d\lambda(x) + C_0\,,$$

na qual definimos a constante positiva $C_0 = \int_a^b \mathrm{E}[\varphi^2(x)]\, d\lambda(x)$. Por hipótese, cada f é da forma $f(x) = \sum_{i=1}^N \alpha_i\, g_i(x)$, o que nos leva a

$$\mathrm{E}[L(\varphi, f)] = \sum_{i,j=1}^N \left(\alpha_i \alpha_j \int_a^b g_i(x) g_j(x)\, d\lambda(x) \right) - 2 \sum_{i=1}^N \left(\alpha_i \int_a^b g_i(x) \mathrm{E}[\varphi(x)]\, d\lambda(x) \right) + C_0\,,$$

em que utilizamos a linearidade da integral. Portanto, minimizar a perda esperada equivale a resolver o problema de minimização restrita da forma quadrática Q do enunciado. Para a matriz $M = (M_{ij})$, note que, para todo $y = (y_1, \ldots, y_N)^\top \in \mathbb{R}^N$ não nulo, temos que

$$\begin{aligned} y^\top M y &= \sum_{i,j=1}^N y_i y_j M_{ij} = \sum_{i,j=1}^N \left(y_i y_j \int_a^b g_i(x) g_j(x)\, d\lambda(x) \right) \\ &= \int_a^b \sum_{i,j=1}^N \left(y_i\, g_i(x)\, y_j\, g_j(x) \right)\, d\lambda(x) = \int_a^b \left(\sum_{i=1}^N y_i\, g_i(x) \right)^2 \, d\lambda(x) > 0\,, \end{aligned}$$

nas quais utilizamos a linearidade da integral. Logo, a matriz M é definida positiva. Isso implica [4] que a forma quadrática Q é convexa e que o problema de minimização restrita de Q possui uma solução global $(\hat{\alpha_1}, \ldots, \hat{\alpha_N})$. Uma vez que a decisão de Bayes é a f que minimiza a perda esperada, o resultado segue. ❖

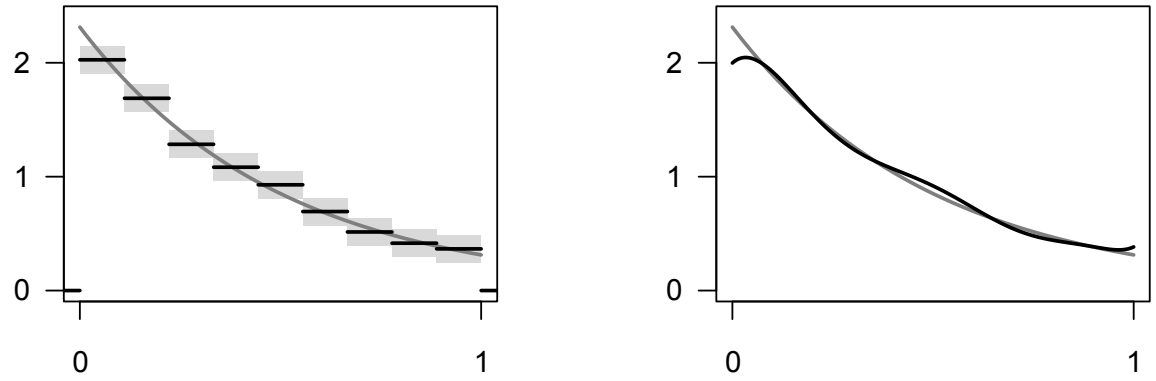


Figura 3.10: Exemplo 3.8. No gráfico da direita, a densidade simples em preto é a estimativa $\hat{\varphi}$, a região cinza é um conjunto crível com nível de credibilidade de 95% e a curva em cinza escuro é a densidade a partir da qual os dados das amostras foram simulados. No gráfico da esquerda, a densidade suave em preto é a decisão de Bayes da Proposição 3.7.

Vamos utilizar o resultado da Proposição 3.7 escolhendo as g_i's dentro de uma classe de densidades diferenciáveis que sirvam aproximadamente como uma base para representar qualquer densidade contínua com suporte no intervalo considerado.

Para o exemplo que veremos a seguir, suponha que o suporte das densidades é o intervalo $[0, 1]$. O Teorema de Bernstein [5] diz que o polinômio

$$B_N(x) = \sum_{i=0}^{N} f\left(\frac{i}{N}\right) \binom{N}{i} x^i (1-x)^{N-i}$$

aproxima uniformemente qualquer função contínua f definida no intervalo $[0, 1]$, quando $N \to \infty$. Suponha que f é uma densidade. Se definimos

$$\alpha_i = f\left(\frac{i}{N}\right) \binom{N}{i} \frac{\Gamma(i+1)\Gamma(N-i+1)}{\Gamma(N+2)},$$

para $i = 0, \ldots, N$, podemos reescrever o polinômio aproximante como $B_N(x) = \sum_{i=0}^{N} \alpha_i \, g_i(x)$, na qual g_i é a densidade de uma variável aleatória com distribuição $Be(i+1, N-i+1)$.

Assim, se tomarmos N suficientemente grande, esperamos que qualquer densidade contínua com suporte no intervalo $[0, 1]$ seja aproximada razoavelmente por uma mistura destas g_i. No próximo exemplo, a classe $\mathscr{D}$ da Proposição 3.7 foi construída a partir destas g_i's, com $i = 0, 1, \ldots, 10$.

Exemplo 3.8 Suponha que temos uma amostra de 5 000 dados simulados a partir de uma distribuição Exponencial, com taxa igual a 2, truncada ao intervalo $[0, 1]$, cuja densidade em relação à medida de Lebesgue é

$$f_0(x) = \frac{2\, e^{-2(x-1)}}{e^2 - 1}\; I_{[0,1]}(x)\,.$$

Repetindo a análise feita no Exemplo 3.8, encontramos o máximo da verossimilhança de K e R no ponto $(\hat{k}, \hat{\rho}) = (9, 0.86)$. O primeiro gráfico da Figura 3.10 apresenta os sumários *a posteriori*. Após isto, resolvemos numericamente o problema de minimização restrita da Proposição 3.7 e chegamos ao resultado do segundo gráfico da Figura 3.10. ❖

O exemplo anterior pode ser generalizado para o caso em que o espaço amostral é o intervalo $[a, b]$ construindo a classe $\mathscr{D}$ a partir de translações e mudanças de escala da mesma família de densidades beta.

Conclusões

Encerramos com breves comentários sobre possíveis modificações e extensões dos modelos considerados na tese.

No Exemplo 3.4, em que simulamos os dados da amostra a partir de uma mistura de normais, tratamos simplificadamente a situação em que os observáveis tem suporte ilimitado, através de um truncamento do espaço amostral. Uma outra possibilidade para tratar estes casos é utilizar uma transformação do espaço amostral. Por exemplo, se tivermos observáveis que assumem valores na reta real, podemos escolher alguma transformação suave com imagem no intervalo $[0, 1]$, digamos, e proceder a análise neste novo espaço amostral utilizando uma densidade aleatória simples. Ao final da análise, utilizamos a transformação inversa para obter respostas no espaço amostral original.

O critério de escolha da partição proposto na seção 3.4, em que estimamos os parâmetros no nível mais alto do modelo hierárquico, pode ser sofisticado se considerarmos partições formadas por subintervalos que possuem comprimentos distintos. Intuitivamente, gostaríamos que a partição escolhida se adaptasse ao fato de que porções do espaço amostral em que observamos mais dados deveriam ser particionadas de maneira mais refinada.

Ainda em relação à técnica utilizada na seção 3.4, temos a alternativa de realizar uma análise bayesiana completa do modelo hierárquico, especificando distribuições *a priori* para K e R. Ao fazermos isto, criamos uma situação em que a própria dimensão do espaço paramétrico não é conhecida. Algoritmos de salto reversível [13] podem ser úteis nas simulações deste caso.

Finalmente, podemos modificar a definição das densidades aleatórias simples para que possamos tratar o caso multivariado, em que cada ponto amostral é um vetor de $\mathbb{R}^d$. Por exemplo, no caso bivariado, podemos particionar uma região limitada do plano e definir uma densidade aleatória simples cujas realizações assumem um valor constante em cada subconjunto da partição, de maneira análoga ao que foi feito no caso univariado estudado na tese.

Referências

[1] R.B. Ash, *Basic Probability Theory*, Dover, New York (2008).

[2] R.B. Ash e C. Doléans-Dade, *Probability and Measure Theory*, segunda edição, Harcourt/Academic Press (2000).

[3] R.B. Ash e M.F. Gardner, *Topics in Stochastic Processes*, Academic Press (1975).

[4] M.S. Bazaraa e C.M. Shetty, *Nonlinear Programming: Theory and Algorithms*, John Wiley & Sons (1979).

[5] P. Billingsley, *Probability and Measure*, terceira edição, Wiley-Interscience (1995).

[6] D. Blackwell, *The Annals of Statistics*, **1**, 356 (1973).

[7] B. De Finetti, *Theory of Probability*, volumes 1 e 2, John Wiley & Sons (1992).

[8] M.H. De Groot, *Optimal Statistical Decisions*, Wiley-Interscience (2004).

[9] T. Ferguson, *The Annals of Statistics*, **1**, 209 (1973).

[10] J.K. Gosh e R.V. Ramamoorthi, *Bayesian Nonparametrics*, Springer (2002).

[11] N.J. Higham, *Accuracy and Stability of Numerical Algorithms*, segunda edição, SIAM (2002).

[12] P.J. Lenk, *Journal of the American Statistical Association*, **83**, 509 (1988).

[13] C.P. Robert e G. Casella, *Monte Carlo Statistical Methods*, segunda edição, Springer (2004).

[14] M.J. Schervish, *Theory of Statistics*, Springer-Verlag (1997).

[15] D. Thorburn, *Biometrika*, **73**, 65 (1986).