
CAMPOS ALEATÓRIOS DE MARKOV E DISTRIBUIÇÕES CONDICIONALMENTE ESPECIFICADAS

Renato Martins Assunção, Erica Castilho Rodrigues
Departamento de Estatística e Laboratório de Estatística Espacial
Universidade Federal de Minas Gerais - UFMG

Elias Teixeira Krainski
Departamento de Estatística e Laboratório de Estatística e Geoinformação
Universidade Federal do Paraná - UFPR

XIX SINAPE, Simpósio Nacional de Probabilidade e Estatística
25 a 30 de Julho de 2010, São Pedro, SP, Brasil.

Palavras, palavras, palavras...

Uma imagem vale mil palavras, diz a convenção.

Agora me diga isso sem usar palavras.

Millor Fernandes, Revista Pif Paf.

Prefácio

Este texto foi escrito para um minicurso apresentado pelos autores no XIX SINAPE, Simpósio Brasileiro de Probabilidade e Estatística, realizado de 26 a 30 de julho de 2010, no Hotel Fazenda da Colina Verde, São Pedro, SP.

O processo de Markov e, em particular, a cadeia de Markov, é um dos tipos de modelos mais populares para representar dados dependentes no tempo. A estrutura uni-dimensional do tempo simplifica tremendamente os cálculos e propriedades desses modelos. Os campos aleatórios de Markov são uma generalização das cadeias de Markov substituindo o espaço-índice unidimensional do tempo por um espaço-índice mais genérico, tal como o espaço geográfico ou uma posição num grafo de vizinhança.

O estudo dos campos aleatórios de Markov levou a um problema teórico em probabilidade. A propriedade de Markov está associada com distribuições condicionais. Sabemos que a partir da distribuição conjunta de n variáveis aleatórias, podemos deduzir as distribuições marginais de cada uma das variáveis. Podemos também obter as distribuições condicionais de cada variável dados os valores das demais variáveis. Algumas vezes, é possível obter o resultado reverso. Isto é, podemos especificar a distribuição condicional de cada variável dados os valores das demais variáveis e, a partir disso, obter a distribuição conjunta. Não é simples determinar quando este resultado é válido, nem saber como obter a conjunta a partir das distribuições condicionais.

Este problema teórico perturbou os estatísticos por vários anos, pois ele era importante para a modelagem de alguns fenômenos aleatórios. Por exemplo, em estatística espacial, uma abordagem natural é a especificação de qual é a distribuição condicional de uma área dado todo o restante do mapa. A propriedade de Markov implica que esta distribuição depende apenas dos va-

lores de suas áreas vizinhas, e não dos valores de áreas mais distantes. Outras aplicações recentes envolvem a modelagem de dados aleatórios em grafos tais como o tráfego na internet, onde cada página da *web* é vista como um nó conectado a outras páginas por meio dos *links*, que fazem o papel de arestas.

O problema de determinar se existe uma única distribuição conjunta associada com as distribuições condicionais e qual é esta distribuição foi resolvido no início da década de 70 por John Hammersley e Peter Clifford. Eles descobriram uma ligação fundamental entre o problema teórico da especificação de uma distribuição via suas condicionais e os campos aleatórios de Markov.

Faz parte da história da estatística o fato de que, incomodados com uma hipótese necessária em sua demonstração do teorema, Hammersley e Clifford nunca publicaram a prova de seu resultado. Em 1974, num dos artigos mais citados da estatística e que deu origem a uma imensa quantidade de pesquisa teórica e aplicada, Julian Besag apresentou uma prova do teorema de Hammersley-Clifford que não exigia conhecimentos avançados de probabilidade e matemática. Esta demonstração faz parte dos clássicos da estatística e esse artigo é um dos aparecem na coleção *Breakthroughs in Statistics*.

Peter Clifford conta um pouco desta história em Clifford (1990): *Much of physics is concerned with providing an understanding of the spatial organisation of matter and it is not surprising that many of the ideas which have become central in the theory of spatial statistics should have their origins in physical theory. The introduction of MRFs into the theory of statistics is yet another example of the continuing transfer of knowledge from the world of theoretical physics. John Hammersley whose interests include both domains of study, was ideally placed to facilitate the process of cross-fertilisation. Others who were involved in this instance include Neyman and Besag. Neyman was responsible for bringing Hammersley and a number of other visitors, including myself, to the University of California, Berkeley in the summer of 1971. Hammersley gave an advanced course of lectures on probabilistic problems in physics, which included among other things a discussion of Spitzer's (1971) characterisation of two-state MRFs on a square lattice. This characterisation had been obtained independently by Averintsev (1970). Hammersley and I were able to generalise the results to arbitrary graphs and lattices, and to identify the central importance of the clique functions, as terms in the potential of a generalised Gibbs distribution. Hammersley returned to Oxford and sent a*

copy of the Berkeley paper to Besag who had already obtained partial results for rectangular lattices (Besag 1972). Besag then wrote to Hammersley with a much simpler, analytical proof of the general result, which appeared later in his very influential paper on spatial statistics (Besag 1974). Three other authors published proofs of the main theorem at about this time (Grimmett 1973, Preston 1973, Sherman 1973).

Este texto faz uma revisão da pesquisa desenvolvida nesta área. Vamos começar apresentando os conceitos básicos envolvidos no problema da especificação de uma distribuição conjunta a partir das distribuições condicionais completas. No Capítulo 2, vamos estudar o caso bivariado, um caso muito simples, para que o leitor fique familiarizado com o problema. No Capítulo 3, vamos introduzir os campos de Markov e as distribuições de Gibbs. No capítulo seguinte, o caso especial de campos markovianos gaussianos será abordado. O Capítulo 5 vai apresentar alguns dos modelos mais comuns de campos gaussianos espaciais. Vamos focar no caso particular de distribuições normais e em modelos de análise geográfica. O modelo mais usado como distribuição a priori na análise de dados espaciais é um modelo autoregressivo condicional (CAR). Vamos estudar este modelo em detalhes, apresentando suas principais propriedades.

O Capítulo 6 é o principal deste livro e ele apresenta o enunciado e a demonstração do teorema de Hammersley-Clifford. No Capítulo seguinte introduzimos os auto-modelos, uma classe de modelos da família exponencial onde a especificação da distribuição conjunta é feita através das distribuições condicionais. Nos dois capítulos finais tratamos da simulação de auto-modelos e da estimação desses modelos.

Nosso objetivo ao escrever este texto é fornecer uma introdução didática para um aluno de mestrado em estatística de alguns dos resultados mais bonitos de inferência estatística. Se ao final da leitura o leitor resolver estudar um pouco mais estatística espacial, nós sentiremos que nosso trabalho terá valido a pena.

Renato M. Assunção, Erica Castilho Rodrigues, Elias Teixeira Krainski
Belo Horizonte, 2010.

Sumário

1	Introdução	1
1.1	Conjunta e marginais	1
1.2	Conjunta e condicionais	4
1.2.1	Encontrando condicionais rapidamente	6
1.2.2	Conjunta a partir de $f_{X Y}(x y)$ e $f_{Y X}(y x)$	9
1.2.3	Conjunta a partir de $a(x, y)$ e $b(x, y)$	9
1.3	Unicidade	13
1.4	Resumindo	13
1.5	Modelos de regressão	14
1.6	Modelos para séries temporais	15
1.7	Modelos para estatística espacial	18
2	Caso Bivariado	25
2.1	Condicionais determinam conjunta	25
2.2	Positividade no caso bivariado	28
2.3	Expansão de Brook: caso bivariado	29
2.4	Existência de conjunta bivariada compatível	30
2.5	Exemplos no caso bivariado	32
2.6	Unicidade	36
2.7	Considerações finais	36
3	Campos de Markov	37
3.1	Grafos	37
3.2	Vizinhança e cliques	42
3.3	Campos aleatórios de Markov	45

3.3.1	Markov no tempo	47
3.3.2	Caso Markov em duas dimensões	48
3.4	Energia e densidades	50
3.5	Distribuições de Gibbs	53
3.6	Distribuições de Gibbs e campos de Markov	55
4	Campos de Marvok Gaussianos	57
4.1	Definição	57
4.2	Independência Condicional	60
4.3	Condicionais obtidas a partir da conjunta	65
4.4	Conjunta obtida a partir das condicionais	68
5	Exemplos de Campos de Markov Gaussianos	73
5.1	Modelo CAR	73
5.2	Modelo ICAR	77
5.3	Modelo de Leroux	78
6	Teorema de Hammersley-Clifford	81
6.1	Expansão de Brook	81
6.2	Algumas definições importantes	84
6.3	O Teorema de Hammersley-Clifford	87
6.4	Extensões	90
6.5	Voltando com energia	91
7	Exemplos de Auto-modelos	95
7.1	Introdução	95
7.2	Modelos autonormal	99
7.3	Modelos autologístico	102
7.4	Modelo autobinomial	103
7.5	Modelo autoPoisson	103
8	Simulação de Auto-modelos	107
8.1	Modelo CAR	107
8.2	Os outros auto-modelos	110
8.3	Modelo autologístico e de Ising	112
8.4	Simulando dos modelos ICAR e autoPoisson	117

SUMÁRIO

ix

9 Estimação de Auto-modelos	123
9.1 Introdução	123
9.2 Modelo CAR	125
9.3 Modelo autologístico	128
9.3.1 Usando pseudo-verossimilhança	128
9.3.2 Estimação via sistema de coding	129
9.3.3 Estimação da variância via bootstrap	131
9.3.4 Pequena avaliação no procedimento de bootstrap	134
Bibliografia	137

Capítulo 1

Introdução

Um dos objetivos principais da análise estatística é conhecer o comportamento probabilístico de várias variáveis aleatórias simultaneamente. A maneira matemática de descrever este comportamento simultâneo é estabelecer a distribuição conjunta para as variáveis. Esta distribuição conjunta é tudo o que é necessário para responder a qualquer pergunta sobre o comportamento estocástico das variáveis.

Vamos adotar a notação de variáveis aleatórias contínuas neste texto, identificando a distribuição conjunta de um vetor (X, Y) com a densidade conjunta $f_{XY}(x, y)$. No caso de variáveis aleatórias discretas, basta substituir a densidade conjunta pela função de probabilidade conjunta, as integrais por somas, etc.

No caso de duas variáveis aleatórias X e Y , através da distribuição conjunta podemos calcular a probabilidade de que X pertença a certo conjunto A e, ao mesmo tempo, Y pertença a um outro conjunto B . Ou calcular a esperança de X , a correlação entre X e Y , ou qualquer outro aspecto que seja de interesse sobre as variáveis.

1.1 Conjunta e marginais

Denote por \mathcal{S}_x , \mathcal{S}_y e \mathcal{S}_{xy} os conjuntos suporte de X , Y e do vetor (X, Y) , respectivamente. Isto é, \mathcal{S}_x é o conjunto de valores em que a densidade marginal $f_X(x) > 0$ e analogamente para \mathcal{S}_y . O conjunto \mathcal{S}_{xy} é formado pelos

pares de pontos (x, y) onde a densidade conjunta é positiva: $f_{XY}(x, y) > 0$.

Através da distribuição conjunta $f_{XY}(x, y)$ obtemos as distribuições marginais simplesmente integrando a conjunta sobre a outra variável. Por exemplo,

$$f_X(x) = \int f_{XY}(x, y) dy.$$

Desta forma, a distribuição conjunta determina as distribuições marginais.

É bem conhecido que o reverso não é verdadeiro: as distribuições marginais não determinam a distribuição conjunta. A menos que as variáveis aleatórias sejam independentes, em geral, conhecer as distribuições marginais $f_X(x)$ e $f_Y(y)$ não implica em conhecer a distribuição conjunta $f_{XY}(x, y)$. Imagine, por exemplo, que você saiba que X e Y possuem ambas distribuição normal com média 0 e variância 1. Qual a distribuição conjunta de X e Y ? É impossível saber a partir apenas dessa informação sobre as distribuições marginais.

Se a distribuição conjunta for uma normal bivariada, precisamos conhecer também o coeficiente de correlação linear $\rho \in [-1, 1]$. Com as distribuições marginais normais e o coeficiente de correlação linear ρ obtemos a distribuição conjunta. Valores muito diferentes de ρ implicam em distribuições conjuntas muito diferentes. Um gráfico de dispersão de uma amostra aleatória de pontos $(x_1, y_1), \dots, (x_n, y_n)$ dá uma boa idéia do que pode ser esta distribuição conjunta. Na Figura 1.1, mostramos 4 gráficos de dispersão de uma amostra aleatória de tamanho $n = 100$ do vetor bivariado (X, Y) variando desde a independência, quando $\rho = 0$, até uma situação de forte dependência positiva, quando $\rho = 0.99$. Valores negativos de ρ geram gráficos similares, refletidos em torno do eixo vertical.

Dessa forma, basta um número adicional, o coeficiente de correlação linear ρ , para determinar a distribuição conjunta neste caso. No entanto, isto depende de sabermos que a distribuição conjunta é uma normal bivariada. Ter as distribuições marginais como normais não implica que a distribuição conjunta seja normal bivariada. Por exemplo, considere $X \sim N(0, 1)$ e obtenha Y da seguinte forma: jogue uma moeda honesta para cima; caso saia cara, faça $Y = X$; caso saia coroa, faça $Y = -X$. A distribuição marginal de Y é também $N(0, 1)$ mas a distribuição conjunta não é normal bivariada. O gráfico de dispersão para uma amostra desta distribuição para (X, Y) está à esquerda na Figura 1.2. No lado direito dessa figura, temos um exemplo menos extremo

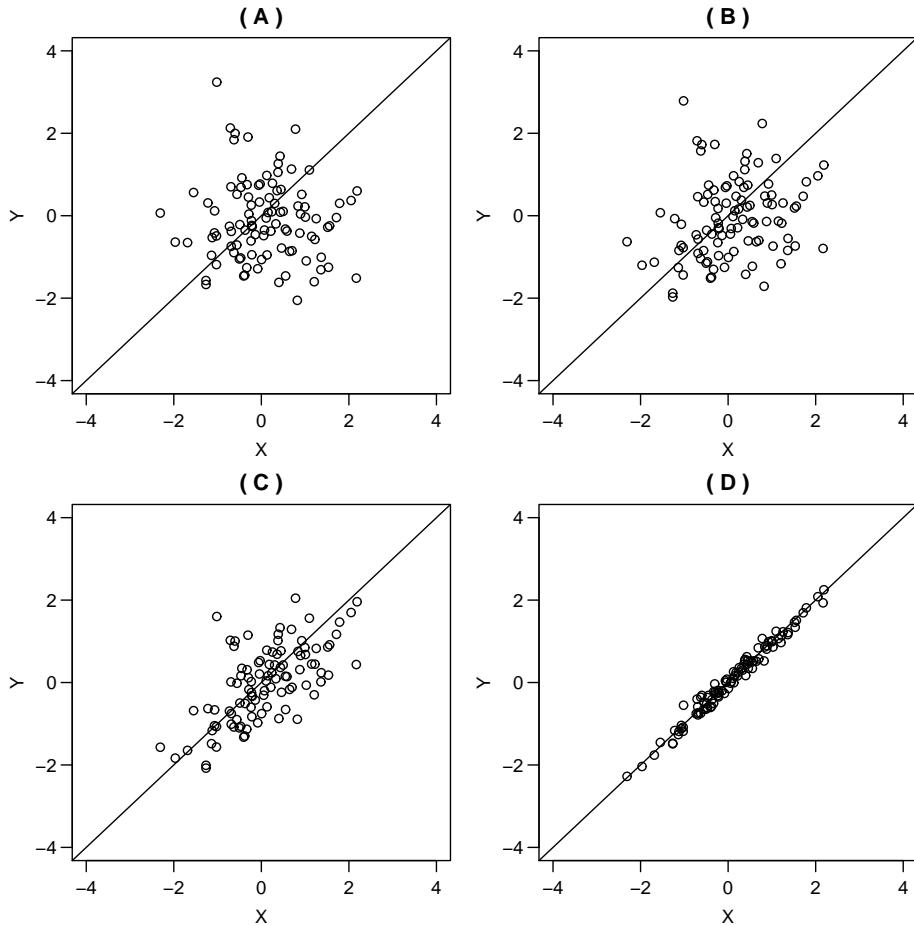


Figura 1.1: Diagrama de dispersão de quatro amostras aleatórias de tamanho 100 da distribuição Normal bivariada com $\rho = 0$ (gráfico A), $\rho = 0.3$ (gráfico B), $\rho = 0.7$ (gráfico C) e $\rho = 0.99$ (gráfico D) simuladas com a mesma semente.

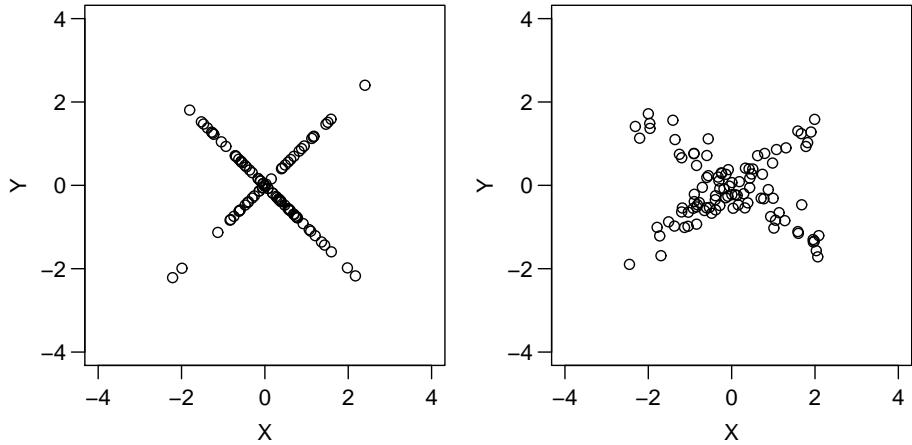


Figura 1.2: Diagramas de dispersão de $X \sim N(0, 1)$ com $Y = X$ ou $Y = -X$ (esquerda) e $Y = \beta X + \epsilon$ ou $Y = -\beta X + \epsilon$ (direita).

mas similar em que $Y = \beta X + \epsilon$ ou $Y = -\beta X + \epsilon$ com probabilidade $1/2$, $\beta = 1/\sqrt{2}$ e ϵ tendo uma distribuição $N(0, 1/2)$ independente de X .

1.2 Conjunta e condicionais

A distribuição conjunta $f_{XY}(x, y)$ determina as distribuições condicionais. Por exemplo, a densidade condicional de Y dado que $X = x$ é definida como

$$f_{Y|X}(y|x) = \frac{f_{XY}(x, y)}{f_X(x)} \quad (1.2.1)$$

desde que $f_X(x) > 0$ (a definição é arbitrária se $f_X(x) = 0$). Revertendo a direção do condicionamento, se $f_Y(y) > 0$, temos

$$f_{X|Y}(x|y) = \frac{f_{XY}(x, y)}{f_Y(y)}. \quad (1.2.2)$$

Como a distribuição conjunta determina as distribuições marginais $f_X(x)$ e $f_Y(y)$, as densidades (1.2.1) e (1.2.2) podem ser calculadas a partir do conhecimento de $f_{XY}(x, y)$.

Para ter uma idéia qualitativa do que é uma densidade condicional para X com $Y = y^*$, basta olhar a distribuição conjunta como função de x deixando y como uma constante com valor y^* . Este é o caso porque, como (1.2.2) mostra, a densidade condicional, como uma função de x , é proporcional à densidade conjunta. Isto é,

$$f_{X|Y}(x|y^*) = \frac{f_{XY}(x, y^*)}{f_Y(y^*)} \propto f_{XY}(x, y^*). \quad (1.2.3)$$

Exercício 1.2.1. Considere duas variáveis aleatórias X e Y com densidade conjunta representada no gráfico tri-dimensional da Figura 1.3(a). Sem fazer nenhum cálculo identifique dentre os demais gráficos apresentados na Figura 1.3 (b, c e d) qual deles melhor representa a distribuição condicional de $X|Y = y$.

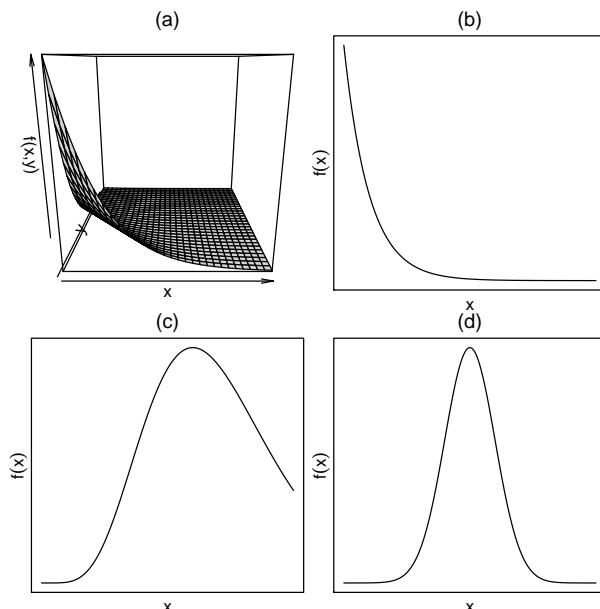


Figura 1.3: Densidade conjunta de duas variáveis aleatórias independentes com distribuição exponencial e possíveis formas da distribuição conjunta $X|Y = y$.

Resposta: Opção (b). ♠

1.2.1 Encontrando condicionais rapidamente

Uma maneira rápida de encontrar a densidade condicional de X dado que $Y = y$ a partir da densidade conjunta é ignorar todos os fatores multiplicativos que não envolvam x e procurar identificar no que sobrar o núcleo de uma densidade de probabilidade. Se reconhecermos no que sobrar uma densidade de probabilidade a menos da constante de integração então teremos encontrado a densidade condicional desejada.

Que podemos ignorar fatores multiplicativos que não envolvam x é fácil de ver. Suponha que $f_{XY}(x, y) = ch(x, y)g(y)$ onde c é uma constante. Então

$$\begin{aligned} f_{X|Y}(x|y) &= \frac{f_{XY}(x, y)}{f_Y(y)} \\ &= \frac{f_{XY}(x, y)}{\int f_{XY}(u, y) du} \\ &= \frac{ch(x, y)g(y)}{\int ch(u, y)g(y) du} \\ &= \frac{ch(x, y)g(y)}{cg(y) \int h(u, y) du} \\ &= \frac{h(x, y)}{\int h(u, y) du} \end{aligned}$$

mostrando que nem c nem $g(y)$ tem qualquer papel na determinação da densidade condicional. Isto é, todos os fatores multiplicativos que não envolvem x podem ser ignorados.

Para ilustrar como podemos também ignorar o denominador da distribuição condicional, vamos usar um exemplo simples. Suponha que (X, Y) é um vetor de variáveis aleatórias com distribuição normal bivariada centrada em zero e com a seguinte matriz de covariância

$$\Sigma = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$$

ou seja, a correlação entre as variáveis é igual a ρ e cada uma delas tem variância unitária. Sabemos então que a densidade conjunta é dada por

$$f_{X,Y}(x,y) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left\{\frac{-1}{2(1-\rho^2)} [x^2 + y^2 - 2\rho xy]\right\}.$$

Para encontrarmos a densidade condicional de $X|Y$, podemos proceder de duas maneiras. A primeira delas é encontrar inicialmente a distribuição marginal de Y e depois utilizar a definição de distribuição condicional:

$$f_X(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}.$$

A segunda maneira seria proceder como apresentado logo acima, ou seja, fazer

$$f_X(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} \propto f_{X,Y}(x,y).$$

Após ignorar o denominador, vamos desconsiderar todos os fatores multiplicativos que não dependem de x na distribuição conjunta e tentamos identificar o núcleo de uma distribuição conhecida.

Iremos fazer aqui das duas maneiras para que fique claro para o leitor como funciona esse último procedimento e qual seu significado. Para o primeiro modo precisamos encontrar a distribuição marginal de Y , que é dada por

$$f_Y(y) = \frac{1}{2\pi\sqrt{1-\rho^2}} \int_{-\infty}^{\infty} \exp\left(\frac{1}{2(1-\rho^2)} [x^2 + y^2 - 2\rho xy]\right) dx.$$

Completando quadrados dentro do parênteses para encontrarmos o núcleo de uma normal temos que

$$\begin{aligned} f_Y(y) &= \frac{1}{2\pi\sqrt{1-\rho^2}} \int_{-\infty}^{\infty} \exp\left(\frac{1}{2(1-\rho^2)} [x^2 + y^2 - 2\rho xy + \rho^2y - \rho^2y]\right) dx \\ &= \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left(\frac{-y^2}{2(1-\rho^2)} + \frac{\rho^2y^2}{2(1-\rho^2)}\right) \int_{-\infty}^{\infty} \exp\left(\frac{1}{2(1-\rho^2)} [x - \rho y]^2\right) dx \end{aligned}$$

Dentro da integral temos então um núcleo de uma distribuição normal com média ρy e variância $(1-\rho^2)$ e, portanto, essa integral é igual a $\sqrt{2\pi(1-\rho^2)}$. Substituindo esse valor e cancelando alguns termos chegamos a

$$f_Y(y) = \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-y^2}{2}\right)$$

ou seja, uma normal padrão. Percebe-se que essa função não depende de x , ela depende apenas de y . Ela entra no cálculo da distribuição condicional e modifica seu formato, pois para valores diferentes de y teremos distribuições condicionais diferentes, como será visto a seguir. Utilizando então a definição de distribuição condicional temos que

$$\begin{aligned} f_X(x|y) &= \frac{f(x,y)}{f(y)} = \frac{\frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left\{\frac{-1}{2(1-\rho^2)} [x^2 + y^2 - 2\rho xy]\right\}}{\frac{1}{\sqrt{2\pi}} \exp\left(\frac{-y^2}{2}\right)} \\ &= \frac{1}{\sqrt{2\pi}\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)} [x^2 + y^2 - 2\rho xy - y^2(1-\rho^2)]\right\} \\ &= \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)} (x - \rho y)^2\right\}. \end{aligned}$$

Isso significa que $X|Y$ é uma distribuição normal com média ρy e variância $\sqrt{1-\rho^2}$. A distribuição condicional de X dado $Y = y$ é uma função desse valor y : quando y varia, a média dessa distribuição também varia.

Vamos agora encontrar a distribuição de $X|Y = y$ do segundo modo mencionado. Como observado anteriormente, o valor de y modifica o formato da distribuição condicional, mas não modifica o núcleo da distribuição. Dessa maneira tudo que depende apenas de y pode ser desconsiderado quando vamos identificar o núcleo. Observe que desconsideramos esses termos apenas para essa identificação: no formato final da distribuição condicional eles terão influência. Descartando da distribuição conjunta os fatores multiplicativos que não dependem de x temos

$$f_{X|Y}(x) \propto f_{X,Y}(x,y) \propto \exp\left\{-\frac{1}{2(1-\rho^2)} [x^2 - 2\rho xy]\right\}.$$

Esse é apenas o núcleo da distribuição, não é ainda sua forma fechada. Para encontrarmos sua forma final precisamos ainda da constante de integração. Porém, muitas vezes, não precisamos encontrar essa constante integrando diretamente, pois o núcleo é de uma distribuição conhecida. Este é exatamente o caso deste exemplo. Vamos completar o quadrado, para tanto

precisamos somar e subtrair o termo ρy , note que como esse termo também não depende de x , ele fará parte da constante de integração e portanto não é necessário escrever o $-\rho y$ aqui.

$$f_{X|Y}(x) \propto \exp \left\{ -\frac{1}{2(1-\rho^2)} [x^2 - 2\rho xy + \rho^2 y^2] \right\} = \exp \left\{ -\frac{1}{2(1-\rho^2)} [x - \rho y]^2 \right\}.$$

Podemos ver então o núcleo de uma distribuição normal com média ρy e variância $(1 - \rho^2)$, assim como foi encontrado procedendo do outro modo.

1.2.2 Conjunta a partir de $f_{X|Y}(x|y)$ e $f_{Y|X}(y|x)$

Dado que a conjunta determina as distribuições condicionais, uma pergunta natural é: as distribuições condicionais determinam a distribuição conjunta? Aqui nós precisamos ser cuidadosos. Devemos quebrar esta pergunta em duas perguntas distintas.

Vamos imaginar inicialmente que sejam dadas as densidades condicionais $f_{X|Y}(x|y)$ e $f_{Y|X}(y|x)$ para todo par de valores $(x, y) \in \mathbb{R}^2$. Estas densidades foram obtidas a partir de uma distribuição conjunta $f_{XY}(x, y)$ que sabemos existir. De posse destas densidades condicionais, é possível recuperar a densidade conjunta? Neste problema não existe nenhuma dúvida sobre a existência das densidade conjunta ou a validade das condicionais. A densidade conjunta existe e foi a partir dela que as condicionais forma obtidas. O que queremos saber é se é possível recuperar esta densidade conjunta tendo apenas as duas condicionais $f_{X|Y}(x|y)$ e $f_{Y|X}(y|x)$.

Diferentemente do caso das densidades marginais, conhecer as densidades condicionais *implica* que conhecemos a densidade conjunta. Veremos este resultado, como a conjunta pode ser recuperada a partir das condicionais, no caso bivariado no próximo capítulo. O caso geral será discutido no Capítulo 6.

1.2.3 Conjunta a partir de $a(x, y)$ e $b(x, y)$

No entanto, o problema acima é artificial. Ele não se parece com o que os estatísticos fazem na prática. O problema que realmente interessa é quando temos duas funções, $a(x, y)$ e $b(x, y)$, que são propostas como distribuições

condicionais. Cada uma delas representa uma família de distribuições de probabilidade. Para cada valor de $y \in \mathcal{S}_y$ fixo, temos uma distribuição de probabilidade $a(x, y)$ para $x \in \mathcal{S}_x$. Do mesmo modo, para cada valor de $x \in \mathcal{S}_x$ fixo, temos uma distribuição de probabilidade $b(x, y)$ para $y \in \mathcal{S}_y$. O problema é que nós não sabemos se $a(x, y)$ e $b(x, y)$ vieram de uma conjunta. Nós não sabemos se esta conjunta existe.

Considere as duas famílias $a(x, y)$ e $b(x, y)$ de densidades condicionais. A questão que nos importa é a seguinte: existe uma densidade conjunta $f_{XY}(x, y)$ tal que

$$f_{X|Y}(x|y) = a(x, y) \text{ e } f_{Y|X}(y|x) = b(x, y).$$

Pode ser que não exista uma distribuição conjunta $f_{XY}(x, y)$ tal que suas densidades condicionais sejam $a(x, y)$ e $b(x, y)$, as duas especificadas pelo analista. Ao especificar um modelo estatístico usando duas famílias de distribuições condicionais $a(x, y)$ e $b(x, y)$, podemos estar cometendo um erro: pode não haver uma densidade conjunta $f_{XY}(x, y)$ tal que, ao calcularmos as condicionais usando as definições (1.2.1) e (1.2.2), encontremos as densidades $a(x, y)$ e $b(x, y)$ especificadas. Isto implica que o modelo baseado nas especificações condicionais não faz sentido, não existe conjunta que gere estas condicionais.

Este problema não aflige a relação entre distribuições marginais e conjunta. De fato, dadas quaisquer duas distribuições de probabilidade $a(x)$ e $b(y)$ existe pelo menos uma densidade conjunta cujas marginais são estas duas especificadas. Basta tomar a conjunta de duas variáveis independentes. Isto é, tome $f_{XY}(x, y) = a(x)b(y)$. Então, integrando e usando o fato de que $a(x)$ e $b(y)$ são densidades de probabilidade, temos $f_X(x) = a(x)$ e $f_Y(y) = b(y)$. Desta forma, pelo menos uma densidade conjunta existe tal que suas marginais serão as duas especificadas, $a(x)$ e $b(y)$.

Esta situação simples não acontece no caso da especificação via distribuições condicionais. Um exemplo muito simples deve ser suficiente para ilustrar a dificuldade para as densidades condicionais determinarem a distribuição conjunta.

Exemplo 1.2.1: Suponha que o vetor aleatório assuma valores no quadrado unitário $[0, 1]^2$ e que as propostas de distribuições condicionais, para $x \in [0, 1]$ e $y \in [0, 1]$, sejam as seguintes:

$$[Y|X = x] \sim U(0, x) \quad (1.2.4)$$

e

$$[X|Y = y] \sim U(0, y). \quad (1.2.5)$$

A Figura 1.4 mostra um gráfico no qual estão destacados os locais em que cada uma dessas variáveis está definida. A região em cinza mais escuro é o local que a variável aleatória X coloca massa de probabilidade quando definimos a condicional $X|Y$ e o cinza mais claro, quando definimos $Y|X$.

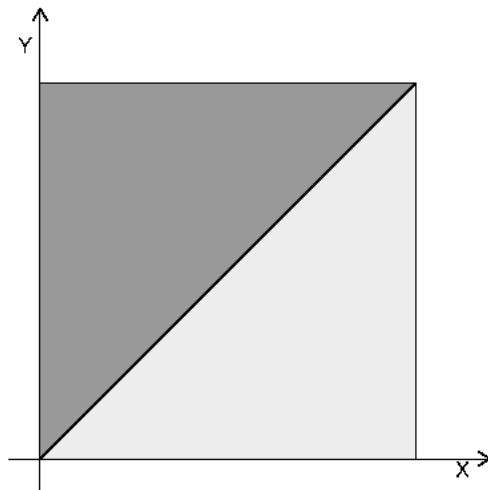


Figura 1.4: Espaço de Probabilidade de duas variáveis para as quais as condicionais não determinam a conjunta

A distribuição em (1.2.4) implica que, para qualquer $x \in (0, 1)$ fixado, os valores que Y pode assumir são aqueles entre 0 e x . Portanto, $Y < X$ com probabilidade 1. De maneira um pouco mais formal,

$$\mathbb{P}(Y < X) = \int \mathbb{P}(Y < X|Y = y) f_Y(y) dy = \int 1 f_Y(y) dy = 1.$$

De maneira análoga, a distribuição em (1.2.5) implica que $X < Y$ com probabilidade 1. Como estas duas condições não podem ocorrer simultaneamente, as distribuições condicionais em (1.2.4) e (1.2.5) não são compatíveis com nenhuma distribuição conjunta. Queremos dizer com isto que não existe uma

distribuição conjunta $f_{XY}(x, y)$ tal que, ao calcular as suas densidades condicionais, encontremos aquelas especificadas em (1.2.4) e (1.2.5). ♠

Exemplo 1.2.2: Assuma que

$$(X|Y = y) \sim N(\mu_1(y), \sigma_1^2)$$

$$(Y|X = x) \sim N(\mu_2(x), \sigma_2^2).$$

Isto é, cada condicional é uma normal. Vamos verificar quando pode existir uma distribuição normal bivariada compatível com estas condicionais¹. Como veremos no Capítulo 2, se existe uma conjunta seguindo uma distribuição normal bivariada, é necessário que

$$\mu_1(y) = a_1 + b_1 y \text{ e que } \mu_2(x) = a_2 + b_2 x$$

e também que

$$b_1 \sigma_2^2 = b_2 \sigma_1^2$$

com $b_1 b_2 < 1$.

Se especificarmos $\mu_1(y) = y^2$ e $\mu_2(x) = x$, não vai existir uma distribuição conjunta normal bivariada compatível com estas distribuições condicionais. ♠

Exemplo 1.2.3: Assuma que

$$(X|Y = y) \sim N(\beta y, \sigma_1^2)$$

$$(Y|X = x) \sim N(\gamma x, \sigma_2^2).$$

Então, só existe uma conjunta se $\beta \gamma < 1$. ♠

O caso bivariado será tratado em detalhes no Capítulo 2. O caso geral será tratado no restante do livro.

¹A distribuição conjunta poderia não ser normal bivariada, como no Exemplo 3 a seguir.

1.3 Unicidade

Outro problema um pouco mais delicado é que talvez exista uma distribuição conjunta compatível com as densidades condicionais especificadas. Mas ela pode não ser única. Podem existir duas ou mais densidades conjuntas tais que suas condicionais sejam aquelas especificadas pelo analista.

Isto ocorre também na relação entre distribuições marginais e conjunta. Por exemplo, se $X \sim N(0, 1)$ e $Y \sim N(0, 1)$, existem infinitas distribuições conjuntas compatíveis com estas distribuições marginais (por exemplo, as normais bivariadas que se obtém variando ρ).

1.4 Resumindo

O problema descrito acima é que as distribuições condicionais especificadas por um analista podem não determinar a distribuição conjunta. Nós usamos a palavra “podem” porque, em certos casos, elas podem sim determinar a distribuição conjunta. A maior parte deste texto é voltada para esclarecer quando isto ocorre. Além disso, vimos que, se existir uma distribuição conjunta associada com as distribuições condicionais, ela pode não ser única.

Muito bem, este é um problema de probabilidade. Qual a relevância disso para a análise de dados? Talvez esse problema não passe de uma questão teórica com pouca importância para a estatística como é praticada. Vale a pena um estatístico estudar este assunto, dentre tantos outros competindo por sua atenção? As próximas seções vão procurar dar a motivação para que este estudo seja feito.

Como veremos, a maior parte dos modelos estatísticos usados na análise de dados é baseada em modelos para distribuições condicionais. A idéia fundamental é que podemos prever o comportamento estatístico de uma variável aleatória Y a partir do conhecimento de uma ou mais variáveis. Em geral, queremos saber qual a distribuição de Y (seu valor esperado, sua variância, etc) sabendo qual é o valor de uma outra variável X . A variável X pode ser fácil de ser obtida ou pode ser uma variável observada antes da realização de Y . Ou ainda X pode ser uma variável passível de manipulação ou intervenção. Todas estas são razões para usar um modelo condicional.

Todos os modelos de regressão são deste tipo, modelos condicionais. Os

modelos de séries temporais e cadeias de Markov, que prevêem Y_t a partir das observações passadas, também são modelo condicionais. Modelos de estatística espacial se encaixam também nesta categoria. São todos modelos definidos a partir de distribuições condicionais. No entanto, veremos que o problema com o qual vamos lidar neste texto aflige principalmente a área de estatística espacial. Para entender isto, vamos começar revendo os modelos de regressão.

1.5 Modelos de regressão

Considere o modelo mais simples e uma dos mais conhecidos, o modelo de regressão linear simples. Existe uma variável X , chamada de covariável ou variável independente e outra variável Y , chamada de resposta ou variável dependente. Pode haver interesse em estudar a variável X por si mesma mas, no modelo de regressão linear, o interesse é saber como Y varia quando X possui certo valor arbitrário. Isto é, dado que $X = x$, qual a distribuição condicional de Y ? A suposição clássica é que $[Y|X = x]$ possui distribuição normal com média $\mu(x) = \beta_0 + \beta_1 x$ e variância σ^2 . Assim, apenas a média muda quando X assume valores diferentes. Somando-se a suposição de independência estocástica entre observações sucessivas, temos o modelo de regressão linear simples.

No problema de regressão linear simples e no de regressão em geral (modelo linear generalizado, modelo aditivo generalizado, etc) o interesse não está na distribuição conjunta das variáveis (X, Y) mas somente na distribuição condicional $[Y|X = x]$. Neste caso, especificamos apenas esta distribuição (para todo valor de x) sem nos preocupar em especificar a distribuição conjunta de X e Y . Por esta razão, não especificamos a distribuição condicionada reversa $[X|Y = y]$ nem a distribuição marginal de X . A razão, como dissemos, é que a variação conjunta de Y e X não é de interesse. Em geral, queremos conhecer apenas a variação de Y sabendo-se que X possui algum valor fixo.

Por exemplo, queremos saber qual a renda Y de um indivíduo dado seu nível de escolaridade X , ou qual o status sócio-econômico Y da ocupação de um indivíduo aos 40 anos sabendo a status X da ocupação de seu pai. Neste caso, é óbvio a relação causal: mudar X implica numa mudança nos valores típicos que podemos esperar para Y .

Dessa forma, embora os modelos de regressão sejam baseados em dis-

distribuições condicionais de Y dado o valor de $X = x$, estes modelos não justificam uma preocupação com os problemas mencionados na seção 1.2. A razão é que não é de interesse, na maioria dos casos, considerar a distribuição condicional reversa, a distribuição de X dado o valor $Y = y$. Isto é, não estamos interessados em partir de modelos onde ambas as distribuições condicionais, $f_{X|Y}(x|y)$ e $f_{Y|X}(y|x)$, são especificadas de alguma forma e esperamos que exista uma única conjunta compatível com elas. Não é assim que fazemos análise de dados com modelos de regressão. A motivação para o estudo dos problemas da seção 1.2 começa a ser desvelada de fato na próxima seção.

1.6 Modelos para séries temporais

Um pouco diferente são os modelos condicionais para séries temporais. Temos n variáveis aleatórias Y_1, Y_2, \dots, Y_n observadas seqüencialmente no tempo. Um dos principais interesses é ser capaz de fazer previsões sobre Y_t conhecendo todo os valores anteriores da série Y_1, Y_2, \dots, Y_{t-1} .

Isto só faz sentido quando acreditamos que o passado pode dizer algo sobre o futuro. Se as variáveis formassem uma seqüência i.i.d. poderíamos ignorar todo o passado já que o que vai acontecer no futuro depende apenas da distribuição marginal de Y_t e não dos valores específicos acontecidos no passado. Com a independência, saber que hoje ocorreu um evento extremo não traz nenhuma informação para a previsão do que vai acontecer amanhã.

É claro que, na maioria das situações práticas, o passado de uma série é relevante para prever o seu futuro. Assim, modelos para séries temporais introduzem dependência entre observações sucessivas. Existe uma infinidade de modelos estatísticos para analisar este tipo de dados temporais. O que queremos é um modelo para a distribuição condicional do valor futuro Y_{t+1} conhecendo-se todo o passado da série. Precisamos especificar as densidades condicionais

$$f_{Y_{t+1}|Y_t, Y_{t-1}, \dots, Y_1}(y|y_t, y_{t-1}, \dots, y_1) \quad (1.6.6)$$

para todos os valores possíveis do vetor $(y_1, \dots, y_{t-1}, y_t, y) \in \mathbb{R}^{t+1}$ e para todos os valores de $t = 1, 2, \dots$

É preciso especificar $n - 1$ densidades condicionais, uma para cada tempo $t = 2, \dots, n$. Para o primeiro momento $t = 1$, supomos que é possível determinar a distribuição marginal $f_{Y_1}(y)$ (ou fazemos inferência condicionalmente no

valor observado de Y_1). Estas n distribuições determinam a distribuição conjunta sem maiores problemas pois a conjunta pode sempre ser escrita como um produto destas distribuições condicionais. Usando a definição de densidade condicional para a variável Y_n e o vetor (Y_1, \dots, Y_{n-1}) , podemos fatorar a densidade conjunta do seguinte modo:

$$f_{Y_1, \dots, Y_n}(y_1, \dots, y_n) = f_{Y_n|Y_1, \dots, Y_{n-1}}(y_n|y_1, \dots, y_{n-1}) \\ f_{Y_1, \dots, Y_{n-1}}(y_1, \dots, y_{n-1}).$$

Aplicando a mesma idéia ao segundo termo do lado direito com a variável Y_{n-1} e o vetor (Y_1, \dots, Y_{n-2}) , temos

$$f_{Y_1, \dots, Y_n}(y_1, \dots, y_n) = f_{Y_n|Y_1, \dots, Y_{n-1}}(y_n|y_1, \dots, y_{n-1}) \\ f_{Y_{n-1}|Y_1, \dots, Y_{n-2}}(y_{n-1}|y_1, \dots, y_{n-2}) \\ f_{Y_1, \dots, Y_{n-2}}(y_1, \dots, y_{n-2}).$$

Prosseguindo dessa forma, obtemos a distribuição conjunta como o produto

$$f_{Y_1, \dots, Y_n}(y_1, \dots, y_n) = \prod_{t=2}^n f_{Y_t|Y_1, \dots, Y_{t-1}}(y_t|y_1, \dots, y_{t-1}) \times f_{Y_1}(y_1). \quad (1.6.7)$$

Assim, a obtenção da conjunta a partir das distribuições condicionais (1.6.6) não apresenta maiores problemas. A questão de compatibilidade não existe e a densidade conjunta é unicamente determinada.

No entanto, especificar as n distribuições condicionais de (1.6.6) pode ser uma tarefa quase impossível. Pense, por exemplo, em especificar um modelo de probabilidade para a temperatura média de hoje baseada em todos os valores diários de temperatura que aconteceram nos últimos dois anos. No caso mais extremo, para cada seqüência de 730 possíveis valores para as temperaturas diárias, precisamos fornecer uma distribuição de probabilidade distinta para os valores da temperatura de hoje.

É preciso simplificar de algum modo esta tarefa e este é o principal papel dos modelos estocásticos: fazer um retrato distorcido e grosseiro da realidade, sem todos os infinitos detalhes que a compõem, mas que capture os traços mais essenciais da situação. Esta é quase a definição de caricatura. De fato, uma caricatura enfatiza e exagera as características do indivíduo retratado, sendo

comum o uso de poucos traços no desenho. Esta é também a característica de um bom modelo estatístico: capturar a essência da situação real em poucas características das distribuições de probabilidade envolvidas.

O modelo mais simples

No modelo mais simples para uma série temporal, assumimos que podemos ignorar todo o passado mais distante (isto é, podemos ignorar $Y_{t-1}, Y_{t-2}, \dots, Y_1$) desde que o presente Y_t seja conhecido. Em outras palavras, a distribuição condicional de Y_{t+1} , dados todos os valores anteriores da série, depende apenas do seu valor mais recente Y_t :

$$f_{Y_{t+1}|Y_t, Y_{t-1}, \dots, Y_1}(y|y_t, y_{t-1}, \dots, y_1) = f_{Y_{t+1}|Y_t}(y|y_t). \quad (1.6.8)$$

Este é um modelo de Markov e a propriedade (1.6.8) é chamada de propriedade markoviana. Ela diz que, dado o valor de Y_t , o futuro Y_{t+1} e o passado $Y_{t-1}, Y_{t-2}, \dots, Y_1$ são independentes.

Pensando em fenômenos reais, parece improvável que este modelo de Markov possa ser útil. Na maioria das vezes, o que vai acontecer amanhã depende não apenas do que acontece hoje mas também do que aconteceu ontem, anteontem, etc. No entanto, este modelo é uma boa aproximação para várias situações práticas. Um dos casos mais impressionantes de sucesso desse modelo é a propagação de genes através de uma linha hereditária. Para prever o DNA de um indivíduo não é necessário a informação do DNA de seu pai e do seu avô paterno, e do seu bisavô paterno, e assim por diante. Basta o DNA do pai para tanto. Todo o resto é dispensável. Nada no DNA do avô ou bisavô paterno pode passar ao indivíduo se não for através de seu pai.

Esse exemplo serve também para mostrar um erro comum ao conhecer o modelo de Markov pela primeira vez. A propriedade markoviana (1.6.8) não implica que Y_{t+1} e Y_{t-1} sejam independentes. Ela diz que Y_{t+1} e Y_{t-1} são *condicionalmente* independentes, dado que Y_t seja conhecido. No contexto do exemplo genético, o DNA do seu avô paterno é relevante para prever seu DNA caso o DNA de seu pai não esteja à mão. Eles não são independentes. No entanto, caso o DNA de seu pai seja conhecido, podemos ignorar todo o DNA de seu avô para efeito de predição do seu DNA. Ele não acrescenta nada em termos da previsão do que pode ser o seu próprio DNA.

No caso em que a propriedade markoviana (1.6.8) é válida, a densidade conjunta (1.6.7) fica reduzida a

$$f_{Y_1, \dots, Y_n}(y_1, \dots, y_n) = \prod_{t=2}^n f_{Y_t|Y_{t-1}}(y_t|y_{t-1}) \times f_{Y_1}(y_1). \quad (1.6.9)$$

1.7 Modelos para estatística espacial

Como vimos na seção anterior, a obtenção da conjunta a partir das distribuições condicionais numa série temporal é simples. Basta usar a igualdade (1.6.7) para estabelecer a distribuição conjunta. A razão principal pela qual o modelo de séries temporais é simples neste aspecto é o fato de que o tempo tem dimensão um, pode ser representado numa reta. Em uma dimensão não existe a necessidade de definir ambas, tanto a distribuição de y_{t+1} em função de y_t quanto a distribuição de y_t em função de y_{t+1} . Como vimos, basta definir a distribuição de cada y_t em função de seu passado. Isto é natural no contexto prático, quando queremos predizer valores futuros usando o que foi observado até o presente.

Na situação espacial (ou em mais dimensões) a situação é um pouco mais complicada. Muitos modelos especificam a distribuição das variáveis envolvidas com certo grau de circularidade. A distribuição do que podemos ver em certo local de um mapa é definida em termos do que é visto em seus locais vizinhos. Mas estes vizinhos, por sua vez, têm a sua distribuição especificada em termos do que o primeiro sítio exibe. Como sair desta circularidade que lembra o ovo e a galinha? São nestes casos que aparece o problema da existência de uma distribuição conjunta a partir da especificação das distribuições condicionais.

Vamos ser mais específicos apresentando um exemplo. Imagine uma imagem quadrada subdividida em quadrados menores. Em cada subquadrado observa-se o valor de uma variável aleatória binária com valores -1 ou 1. A Figura 1.5 mostra 4 imagens deste tipo com os -1s representados por quadrados brancos e os 1s por quadrados pretos. Em cada uma das imagens temos uma grade regular (ou reticulado) que pode ser representada como uma matriz 128×128 com elementos $Y(i, j)$, com $i, j = 1, \dots, 64$. Os subquadrados são chamados de nós, sítios ou píxeis (plural de píxel).

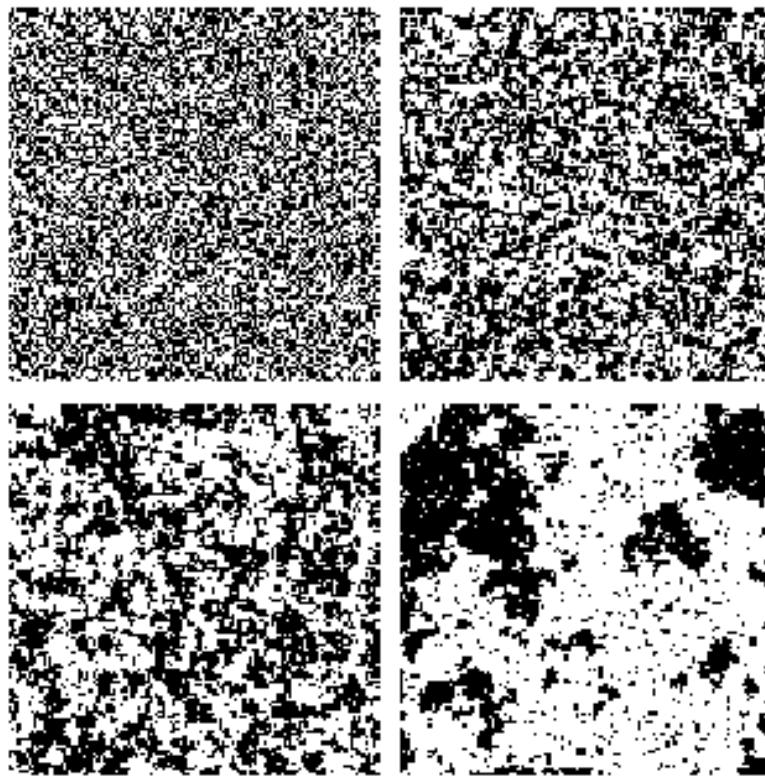


Figura 1.5: Imagens binárias de tamanho 128×128 geradas pelo modelo de Ising.

Várias aplicações geram dados desta forma. Por exemplo, podemos pensar numa plantação onde cada subquadrado representa uma árvore e as variáveis binárias indicam a presença ou ausência de certo atributo na árvore tal como uma doença.

Qual modelo estatístico pode ser usado para representar o processo de geração destas imagens? Como especificar uma distribuição conjunta para a matriz aleatória composta pelos 128^2 elementos $Y(i, j)$ da matriz-imagem? Como cada variável possui apenas dois valores possíveis, a distribuição de probabilidade deve ser discreta. Numa grade regular com $n \times m$ nós, existem

2^{nm} configurações ou imagens possíveis. Este número é extremamente grande mesmo para n e m pequenos. Numa grade 8×8 com 64 sítios, teríamos 2^{64} , ou da ordem de 10^{20} , configurações possíveis.

Ao invés de especificar a distribuição conjunta $f(y(1,1), \dots, y(n,m))$ a abordagem preferida em estatística espacial é especificar a distribuição condicional do valor $y(i,j)$ de um pixel dados os valores de todos os outros pixels na imagem. Um dos modelos mais famosos é o modelo de Ising proposto por físicos, que descreveremos abaixo.

Para cada pixel (i,j) no interior da imagem, considere como sítios vizinhos aqueles obtidos pelos possíveis movimentos do peão num tabuleiro de xadrez. Isto é, os sítios vizinhos de (i,j) formam o conjunto

$$\mathcal{N}_{ij} = \{(i-1,j), (i+1,j), (i,j-1), (i,j+1)\} .$$

No caso de sítios nas bordas, o conjunto \mathcal{N}_{ij} fica menor pois alguns dos vizinhos definidos acima não existem. Cada um dos sítios nos quatro cantos extremos terá apenas dois vizinhos enquanto os demais sítios ao longo das bordas da imagem terão três vizinhos.

Seja p_{ij} a probabilidade condicional de $Y(i,j)$ ser preto dado o restante da imagem (ou todos os outros sítios):

$$p_{ij} = \mathbb{P}(Y(i,j) = 1 | y(i^*, j^*), i^* \neq i, j^* \neq j) .$$

Sendo binária a variável $Y(i,j)$, basta esta probabilidade para especificar completamente a distribuição condicional. Como no modelo de Markov para séries temporais, o modelo de Ising simplifica a dependência supondo que p_{ij} depende apenas dos valores nos sítios \mathcal{N}_{ij} vizinhos de (i,j) , dispensando o restante da imagem. Isto é, nós assumimos que

$$\begin{aligned} p_{ij} &= \mathbb{P}(Y(i,j) = 1 | y(i^*, j^*), i^* \neq i, j^* \neq j) \\ &= \mathbb{P}(Y(i,j) = 1 | y(i^*, j^*), \text{ com } (i^*, j^*) \in \mathcal{N}_{ij}) . \end{aligned}$$

Como veremos mais tarde, este é um modelo de Markov em duas dimensões.

Ernst Ising (10 de maio, 1900, Colônia, Alemanha - 11 de maio, 1998, Peoria, Illinois, EUA) foi um físico alemão conhecido pelo desenvolvimento do modelo Ising. Ising obteve seu doutorado em física na Universidade de Hamburgo, Alemanha, em 1924. Sua tese abordou um problema sugerido e introduzido em 1920 por seu orientador, Wilhelm Lenz. Ele investigou o caso especial de uma cadeia linear de partículas que são capazes de assumir apenas dois estados, -1 e 1 , e que interagem com os vizinhos mais próximos. O modelo era uma aproximação para materiais ferromagnéticos em que cada átomo pode estar polarizado de uma entre duas possíveis formas. Ele investigou o fenômeno de transição de fase nestes materiais com seu modelo, que se tornou o famoso modelo de Ising. Usando o modelo de Ising, os estudos de Rudolf Peierls, Hendrik Kramers, Gregory Wannier e Lars Onsager tiveram sucesso em explicar as transições de fase entre estados ferromagnéticos e paramagnéticos.

Ising desistiu de continuar fazendo pesquisa científica após seu doutorado por acreditar que seu modelo não tinha nenhuma utilidade para o estudo de fenômenos físicos. Somente depois da segunda guerra mundial, após vários incidentes, ele fica ciente de que era um nome famoso por causa do trabalho de outros cientistas com seu modelo (ver Brush, 1967, para mais detalhes).

Após seu doutorado, ele foi trabalhar na General Electric Company. Insatisfeito neste emprego, em 1930 ele começou a trabalhar como professor de segundo grau de uma escola pública em Berlim e casou-se no mesmo ano. Quando Hitler chegou ao poder em 1933 as coisas ficaram mais complicadas pois Ernst era judeu. Ele foi demitido de seu emprego de funcionário público. Em 1934, ele tornou-se professor e mais tarde também diretor de uma escola-internato onde estudavam apenas os alunos judeus que haviam sido excluídos das escolas públicas na Alemanha. Durante os ataques violentos (pogrom) contra a população judia alemã em 1938, a escola foi destruída e Ernest foi interrogado pela Gestapo.

Ising e sua esposa fugiram para Luxemburgo em 1939 planejando emigrar para os EUA assim que possível. Antes disso, Hitler invadiu Luxemburgo. A maioria dos judeus era deportada para os campos de concentração na Alemanha mas como a esposa de Ernst não era judia ele pode permanecer em Luxemburgo sobrevivendo com trabalhos braçais. Os Isings sobreviveram à guerra e chegaram aos EUA em 1946 com um filho pequeno. Ele ensinou física no State Teachers College em Minot, North Dakota e em 1948 foi para a Bradley University em Peoria, Illinois, onde ele ficou até sua aposentadoria em 1976. Em 1998, ele faleceu em sua residência aos 98 anos.

Atualmente, mais de 800 artigos por ano são publicados usando modelo de Ising. Estes artigos versam sobre temas tão diversos quanto redes neurais, estrutura tri-dimensional das proteínas, membranas biológicas e interação social. Existem aproximadamente 40 mil referências ao modelo de Ising na web. A versão em inglês de sua tese de doutorado pode ser encontrada em http://www.hs-augsburg.de/~harsch/anglica/Chronology/20thC/Ising/isi_intr.html.



Além da estrutura markoviana, o modelo de Ising assume também que a chance de um sítio (i, j) ser preto é uma função apenas do número de sítios pretos dentre os seus vizinhos e de um parâmetro β :

$$\begin{aligned} p_{ij} &= \mathbb{P}(Y(i, j) = 1 | y(i^*, j^*), \text{ com } (i^*, j^*) \in \mathcal{N}_{ij}) \\ &\propto \exp(\beta s_{ij}) \end{aligned}$$

onde

$$s_{ij} = \sum_{(i^*, j^*) \in \mathcal{N}_{ij}} y_{i^*, j^*} \quad \text{e} \quad \beta \text{ é uma constante pertencente a } \mathbb{R}.$$

Como cada valor é igual a -1 ou 1, isto quer dizer que s_{ij} é igual ao número de vizinhos que possuem valor igual a 1 (preto) menos o número de vizinhos que possuem valor igual a -1 (branco). Como o número de vizinhos é fixo, no final s_{ij} depende apenas de quantos vizinhos são pretos.

Como a probabilidade p_{ij} está dada a amenos de uma constante de integração, precisamos também saber o que é a probabilidade de um sítio ser branco dados seus vizinhos. O modelo de Ising assume que:

$$\begin{aligned} 1 - p_{ij} &= \mathbb{P}(Y(i, j) = -1 | y(i^*, j^*), \text{ com } (i^*, j^*) \in \mathcal{N}_{ij}) \\ &\propto \exp(-\beta s_{ij}). \end{aligned}$$

Como devemos ter

$$1 = p_{ij} + (1 - p_{ij}) = c (\exp(\beta s_{ij}) + \exp(-\beta s_{ij}))$$

temos então

$$p_{ij} = \frac{\exp(\beta s_{ij})}{\exp(\beta s_{ij}) + \exp(-\beta s_{ij})} = \frac{1}{1 + \exp(\beta s_{ij})}. \quad (1.7.10)$$

Para um sítio que não esteja na borda, o número total de vizinhos é igual a 4 e portanto o número de vizinhos pretos é igual a 4 menos o número de vizinhos brancos. Escrevendo de forma mais explícita, temos

$$p_{ij} \propto \begin{cases} (\exp(\beta))^4, & \text{se todos vizinhos são pretos} \\ (\exp(\beta))^2, & \text{se três vizinhos são pretos} \\ 1, & \text{se dois vizinhos são pretos} \\ (\exp(\beta))^{-2}, & \text{se apenas um vizinho é preto} \\ (\exp(\beta))^{-4}, & \text{se nenhum vizinho é preto} \end{cases}$$

Assim, se $\beta > 0$, quanto mais vizinhos pretos tem um sítio, maior é a probabilidade de que ele mesmo seja também preto. Se os vizinhos são brancos, sua probabilidade de ser preto será pequena levando o sítio a tornar-se branco, como seus vizinhos. E quanto maior o valor de β , mais forte será esta interação.

A partir de (1.7.10) podemos deduzir que

$$\log \left(\frac{p_{ij}}{1 - p_{ij}} \right) = 2\beta s_{ij} .$$

Esta última expressão mostra que temos um modelo logístico para a probabilidade condicional de um sítio ser preto com a diferença entre o número de vizinhos pretos e brancos como covariável. O parâmetro $\beta \in \mathbb{R}$ mostra o grau de dependência de $Y(i, j)$ com seus vizinhos. Se $\beta = 0$, os sítios escolhem suas cores independentemente e com probabilidade 1/2. Se $\beta > 0$, a chance de um sítio ser preto aumenta a medida que cresce o número de seus vizinhos pretos. A Figura 1.5 apresenta configurações geradas com $\beta = 0, 0.25, 0.35, 0.44$.

Um valor negativo para β indica uma espécie de repulsão entre os valores: a probabilidade de um sítio ser negro diminui a medida que aumentam seus vizinhos pretos. No caso extremo de β muito negativo teríamos configurações semelhantes a um tabuleiro de xadrez.

Dependendo do valor de β , configurações com conglomerados de valores iguais podem ser estimulados recebendo mais massa de probabilidade (caso de $\beta > 0$) ou, alternativamente, desestimulados (se $\beta < 0$). Quando $\beta = 0$, a probabilidade de todas as configurações é a mesma e os valores nos diferentes sítios são independentes.

Neste problema não é óbvio qual é a distribuição conjunta das variáveis $Y(i, j)$. Nós especificamos $128^2 = 16384$ distribuições condicionais, uma para cada um dos sítios. Especificamos a distribuição de, digamos, $Y(10, 10)$, condicionada aos valores do sítio (9, 10) e dos demais vizinhos. E especificamos ao mesmo tempo a distribuição de $Y(9, 10)$ condicionada aos valores do sítio (10, 10) e dos outros vizinhos.

Está presente neste caso aquele mesmo tipo de circularidade do exemplo das distribuições condicionais uniformes (1.2.4) e (1.2.5). Como garantir que existe uma distribuição conjunta compatível com todas as $128^2 = 16384$ distribuições condicionais? Como garantir que esta conjunta é única? Este é o grande problema que o Teorema de Hammersley-Clifford resolve e que vamos procurar estudar neste texto.

Capítulo 2

Caso Bivariado

Como foi dito anteriormente, não é qualquer especificação de duas distribuições condicionais que implica numa distribuição conjunta válida. Neste capítulo, vamos considerar apenas o caso bivariado para que o leitor possa se familiarizar com o assunto. No restante do livro, vamos tratar do caso geral de n variáveis aleatórias, um caso mais difícil e mais importante para a análise de dados.

Seja (X, Y) um vetor bivariado de variáveis aleatórias. Suponha que (X, Y) seja absolutamente contínua com respeito à medida produto $\mu_1 \times \mu_2$ em $S_x \times S_y$ onde S_x e S_y são os suportes de X e Y , respectivamente.

Se você não está familiarizado com esta linguagem, simplesmente assuma que o vetor (X, Y) seja composto por duas variáveis, ambas contínuas ou ambas discretas. Além disso, S_x e S_y são os conjuntos em \mathbb{R} em que a densidade de probabilidade de X é positiva (no caso contínuo) ou em que a função de probabilidade é positiva (no caso discreto). Analogamente, para S_y .

2.1 Condicionais determinam conjunta

Veremos inicialmente a situação mais simples em que existe uma distribuição conjunta $f_{XY}(x, y)$ e as duas distribuições condicionais são calculadas a partir desta conjunta. Portanto, não existe dúvida de que pelo menos um a conjunta existe. A questão é: se apenas as duas distribuições condicionais $f_{X|Y}(x|y)$ e $f_{Y|X}(y|x)$ forem conhecidas, é possível recuperar a distribuição

conjunta de onde elas foram calculadas? A resposta é sim e a prova é bem simples.

Exemplo 2.1.1: Antes de apresentar o teorema, vamos ver um caso simples que vai ilustrar como o teorema funciona. Suponha que (X, Y) siga uma distribuição normal bivariada em que cada marginal seja uma $N(0, 1)$ e que o coeficiente de correlação linear seja ρ . Então sabemos que

$$X|Y = y \sim N(\rho y, 1 - \rho^2)$$

e que

$$Y|X = x \sim N(\rho x, 1 - \rho^2).$$

Calculando a razão entre as duas densidades condicionais encontramos

$$\frac{f_{X|Y}(x|y)}{f_{Y|X}(y|x)} = e^{-x^2/2} e^{y^2/2} = u(x)v(y).$$

Isto é, a razão é o produto de duas funções, uma dependendo apenas de x e outra dependendo apenas de y . Além disso, a menos de uma constante de integração, $u(x)$ e $v(y)$ são as densidades marginais de X e Y , respectivamente. Isto é, $f_X(x) = cu(x)$.

Portanto, se quisermos recuperar a densidade conjunta

$$f_{XY}(x, y) = f_{Y|X}(y|x)f_X(x),$$

basta escrever a função

$$h(x, y) = f_{Y|X}(y|x)u(x)$$

e integrá-la no plano para obter uma constante de integração e assim uma densidade de probabilidade que será a densidade conjunta $f_{XY}(x, y)$. É exatamente isto o que o teorema a seguir faz. ♠

Teorema 2.1.1. *Se as distribuições condicionais $f_{X|Y}(x|y)$ e $f_{Y|X}(y|x)$ calculadas a partir de uma distribuição conjunta forem dadas, é possível obter $f_{XY}(x, y)$.*

Demonstração. Para evitar complicações com uma prova rigorosa (mas simples) que envolveria teoria da medida, vamos assumir que as funções envolvidas são todas contínuas por partes com um número finito de saltos.

Considere um ponto (x, y) onde $f_{XY}(x, y) > 0$. Pela continuidade da densidade, $f_X(x) > 0$ e $f_Y(y) > 0$. Assim,

$$f_{X|Y}(x|y) > 0 \text{ e } f_{Y|X}(y|x) > 0. \quad (2.1.1)$$

Considere então um ponto (x, y) onde $f_{XY}(x, y) > 0$. Todas as razões abaixo estão bem definidas pois os denominadores são todos positivos:

$$\frac{f_{X|Y}(x|y)}{f_{Y|X}(y|x)} = \frac{f_{XY}(x, y)/f_Y(y)}{f_{XY}(x, y)/f_X(x)} = \frac{f_X(x)}{f_Y(y)} = u(x)v(y).$$

Desta forma, a razão entre as densidades condicionais deve ser um produto de duas funções, $u(x)$ função apenas de x e $v(y)$ função apenas de y . Além disso, estas funções serão proporcionais às densidades marginais de X e Y .

Portanto,

$$h(x, y) = f_{Y|X}(y|x)u(x)$$

é proporcional à densidade conjunta $f_{XY}(x, y)$. Basta agora obter a constante de integração para encontrar a densidade conjunta. ♠ □

Exemplo 2.1.2: Suponha que (X, Y) siga uma distribuição uniforme no triângulo formado pelos pontos $(0, 0)$, $(0, 1)$ e $(1, 1)$. As distribuições condicionais serão também uniformes: $f_{X|Y}(x|y) = 1/y$, se $x \in (0, y)$ e $y \in (0, 1)$, e $f_{Y|X}(y|x) = 1/(1-x)$, se $y \in (x, 1)$ e $x \in (0, 1)$.

Então, para pontos (x, y) no triângulo,

$$\frac{f_{X|Y}(x|y)}{f_{Y|X}(y|x)} = (1-x) \frac{1}{y}$$

e assim

$$f_{XY}(x, y) \propto h(x, y) = \frac{1}{1-x} (1-x) = 1.$$

Portanto a densidade conjunta é uma uniforme no triângulo. ♠

Existe uma outra forma de obter a distribuição conjunta a partir de $f_{X|Y}(x|y)$ e $f_{Y|X}(y|x)$, quando estas condicionais de fato foram obtidas de uma conjunta.

No entanto, esta outra maneira, chamada expansão de Brook, requer uma condição adicional da distribuição conjunta, chamada de condição de positividade. Esta condição é muito importante para o caso multivariado geral.

2.2 Positividade no caso bivariado

Definição 2.2.1. *No caso bivariado, dizemos que a condição de positividade é satisfeita se toda vez em que $f_{XY}(x_1, y_1) > 0$ e $f_{XY}(x_2, y_2) > 0$ implicar que $f_{XY}(x_1, y_2) > 0$ e $f_{XY}(x_2, y_1) > 0$.*

Esta condição é equivalente a dizer que se $f_X(x) > 0$ e $f_Y(y) > 0$ então $f_{XY}(x, y) > 0$. Isto é, o suporte \mathcal{S}_{xy} do vetor (X, Y) é o produto cartesiano $\mathcal{S}_x \times \mathcal{S}_y$ dos suportes de cada variável individual. No caso discreto, isto quer dizer que se o evento $X = x$ e o evento $Y = y$ têm probabilidades positivas de ocorrerem então o evento $X = x, Y = y$ em que ambos ocorrem simultaneamente também tem probabilidade positiva de ocorrer.

Exercício 2.2.1. Prove a equivalência matemática entre as duas definições de positividade dadas acima. ♠

Um exemplo em que esta condição não é satisfeita é aquele em que o vetor (X, Y) possui distribuição uniforme dentro do triângulo formado pelos pontos $(0, 0)$, $(0, 1)$ e $(1, 1)$. Neste caso, $f_{XY}(0.1, 0.2) > 0$ e $f_{XY}(0.8, 0.9) > 0$ mas $f_{XY}(0.8, 0.2) = 0$ violando a definição de positividade. Usando a definição equivalente, vemos que $f_X(0.9) > 0$ e $f_Y(0.1) > 0$ mas $f_{XY}(0.9, 0.1) = 0$.

Vimos na seção anterior como esta distribuição uniforme era obtida a partir de suas condicionais. Isto não é possível com a expansão de Brook. Assim, o método da expansão de Brook é menos geral que aquele apresentado na seção anterior. No entanto, apesar de menos geral, esta expansão de Brook é importante pois a condição de positividade é encontrada com frequência na prática e porque, no caso multivariado, é somente através da expansão de Brook que obtemos resultados relevantes.

2.3 Expansão de Brook: caso bivariado

Teorema 2.3.1. *Seja (X, Y) um vetor aleatório satisfazendo a condição de positividade e seja (x^*, y^*) um ponto qualquer do suporte \mathcal{S}_{xy} . Então*

$$\frac{f_{XY}(x, y)}{f_{XY}(x^*, y^*)} = \frac{f_{X|Y}(x|y^*)}{f_{X|Y}(x^*|y^*)} \frac{f_{Y|X}(y|x)}{f_{Y|X}(y^*|x)}. \quad (2.3.2)$$

Demonstração. Simples substituição no lado direito da igualdade das densidades condicionais pela sua definição como conjunta dividida pela marginal produz o resultado:

$$\begin{aligned} \frac{f_{X|Y}(x|y^*)}{f_{X|Y}(x^*|y^*)} \frac{f_{Y|X}(y|x)}{f_{Y|X}(y^*|x)} &= \frac{f_{XY}(x, y^*)/f_Y(y^*)}{f_{XY}(x^*, y^*)/f_Y(y^*)} \frac{f_{YX}(y, x)/f_X(x)}{f_{YX}(y^*, x)/f_X(x)} \\ &= \frac{f_{XY}(x, y)}{f_{XY}(x^*, y^*)}. \end{aligned}$$



Fica claro acima porque precisamos da condição de positividade: o valor $f_{XY}(x, x^*)$ precisa ser positivo para que a expressão acima faça sentido.

O lado direito da expansão de Brook em (2.3.2) fornece a expressão da densidade conjunta a menos de uma constante. Por exemplo, suponha que $(x^*, y^*) = (0, 0)$ faça parte do suporte de (X, Y) . Assim,

$$f_{XY}(x, y) = c \frac{f_{X|Y}(x|0)}{f_{X|Y}(0|0)} \frac{f_{Y|X}(y|x)}{f_{Y|X}(0|x)}$$

onde a constante de integração c é simplesmente o valor $f_{XY}(0, 0)$. Assim, basta integrar sobre o plano \mathbb{R}^2 para obter a densidade conjunta.

Observe que a expansão de Brook também pode ser feita numa ordem diferente daquela mostrada em (2.3.2). Podemos escrever:

$$\frac{f_{XY}(x, y)}{f_{XY}(x^*, y^*)} = \frac{f_{Y|X}(y|x^*)}{f_{Y|X}(y^*|x^*)} \frac{f_{X|Y}(x|y)}{f_{X|Y}(x^*|y)}.$$

A demonstração é como acima, basta substituir as densidades condicionais pelas suas expressões em termos da conjunta e marginais.

2.4 Existência de conjunta bivariada compatível

Nesta seção, nosso problema vai ficar um pouco diferente. Considere duas famílias de densidades condicionais $a(x, y)$ e $b(x, y)$. Nós queremos saber quando existe uma densidade conjunta $f_{XY}(x, y)$ tal que

$$f_{X|Y}(x|y) = a(x, y)$$

e

$$f_{Y|X}(y|x) = b(x, y).$$

Se esta densidade existe nós dizemos que a e b são famílias de densidades condicionais compatíveis (com a existência de uma conjunta). Além disso, quando existir, queremos saber se ela é única e também como encontrar esta distribuição conjunta. Estes resultados foram demonstrados da forma como vamos fazer aqui por Arnold e Press (1989). Uma discussão mais extensa por ser encontrada em Arnold, Castillo e Sarabia (2001).

Observe que a razão para escrevemos $a(x, y)$ e $b(x, y)$ ao invés de $f_{X|Y}(x|y)$ e $f_{Y|X}(y|x)$ é que não sabemos se elas são condicionais legítimas. Isto é nós ainda não sabemos se existe de fato uma conjunta que tenha como condicionais estas duas famílias $a(x, y)$ e $b(x, y)$ de distribuições.

Sejam $N_a = \{(x, y) : a(x, y) > 0\}$ e $N_b = \{(x, y) : b(x, y) > 0\}$.

Teorema 2.4.1. *Existe pelo menos uma densidade conjunta $f_{XY}(x, y)$ compatível com $a(x, y)$ e $b(x, y)$ se, e somente se,*

(i) $N_a = N_b = N$

(ii) existem funções $u(x)$ e $v(y)$ tais que

$$\frac{a(x, y)}{b(x, y)} = u(x) v(y)$$

para todo $(x, y) \in N$ com $\int_{S_x} u(x) d(x) < \infty$.

Demonstração. A condição (i) é simples: se existe alguma conjunta compatível, o resultado (2.1.1) diz que $N_a = N_b = N$. Se existir uma conjunta, para todo $(x, y) \in N$, temos

$$\frac{a(x, y)}{b(x, y)} = \frac{f_{X|Y}(x|y)}{f_{Y|X}(y|x)} = \frac{f_{XY}(x, y)/f_Y(y)}{f_{XY}(x, y)/f_X(x)} = \frac{f_X(x)}{f_Y(y)} = u(x) v(y).$$

e parte da condição (ii) é satisfeita. Além disso, como $f_X(x) \propto u(x)$, então $u(x)$ integrável.

A volta do teorema, a parte do “se”, é demonstrada de forma similar. Se a condição (ii) é válida, então defina $h(x, y) = b(x, y)u(x)$. Basta agora verificar que, a menos de uma constante multiplicativa, $h(x, y)$ é uma densidade conjunta cujas densidades condicionais são $a(x, y)$ e $b(x, y)$. De fato, $h(x, y) \geq 0$ para todo ponto no plano. Além disso, $h(x, y)$ é integrável pois, assumindo que lidamos com variáveis contínuas,

$$\int_{\mathcal{S}_x} \int_{\mathcal{S}_y} h(x, y) dy dx = \int_{\mathcal{S}_x} u(x) \left(\int_{\mathcal{S}_y} b(x, y) dy \right) dx .$$

Mas como $b(x, y)$ é a nossa candidata a distribuição condicional de Y dado que $X = x$, temos

$$\int_{\mathcal{S}_y} b(x, y) dy = 1 \quad (2.4.3)$$

para todo $x \in \mathcal{S}_x$ fixo. Portanto

$$\int_{\mathcal{S}_x} \int_{\mathcal{S}_y} h(x, y) dy dx = \int_{\mathcal{S}_x} u(x) dx < \infty .$$

pela hipótese. Finalmente, se $f_{XY}(x, y) = c h(x, y)$, as densidades condicionais num ponto arbitrário (x^*, y^*) serão iguais a $a(x^*, y^*)$ e $b(x^*, y^*)$, como requerido:

$$\begin{aligned} f_{Y|X}(y^*|x^*) &= \frac{c h(x^*, y^*)}{\int_{\mathcal{S}_y} c h(x^*, y) dy} \\ &= \frac{b(x^*, y^*) u(x^*)}{u(x^*) \int_{\mathcal{S}_y} b(x^*, y) dy} \\ &= b(x^*, y^*) , \end{aligned}$$

usando (2.4.3) na última igualdade. Com relação à outra densidade, usando a

igualdade na condição (ii), temos:

$$\begin{aligned} f_{X|Y}(x^*|y^*) &= \frac{b(x^*, y^*) u(x^*)}{\int_{S_x} b(x, y^*) u(x) dx} \\ &= \frac{a(x^*, y^*)/v(y^*)}{1/v(y^*) \int_{S_x} a(x, y^*) dx} \\ &= a(x^*, y^*) \end{aligned}$$

pois

$$\int_{S_x} a(x, y^*) dx = 1$$

por ser uma distribuição condicional para $X|Y = y^*$. ♠

□

2.5 Exemplos no caso bivariado

Exemplo 2.5.1: No Capítulo 1, vimos o exemplo de duas distribuições condicionais uniformes:

$$[Y|X = x] \sim U(0, x) \quad (2.5.4)$$

e

$$[X|Y = y] \sim U(0, y). \quad (2.5.5)$$

É claro que a condição (i) do teorema falha neste exemplo e portanto não pode existir densidade conjunta para o vetor (X, Y) satisfazendo estas duas distribuições condicionais ao mesmo tempo. ♠

Exemplo 2.5.2: No Capítulo 1, vimos um exemplo em que

$$(X|Y = y) \sim N(\mu_1(y), \sigma_1^2) \quad (2.5.6)$$

$$(Y|X = x) \sim N(\mu_2(x), \sigma_2^2). \quad (2.5.7)$$

Isto é, cada condicional é uma normal. Vamos verificar quando pode existir uma distribuição normal bivariada compatível com estas duas propostas de distribuições condicionais. Vamos descobrir que restrições devem ser colocadas

nas expressões das médias $\mu(y)$ e $\mu(x)$ como função dos valores condicionados x e y .

Suponha que o vetor (X, Y) possua uma distribuição normal bivariada com vetor de médias (μ_x, μ_y) e com matriz de covariância dada por

$$\Sigma = \begin{bmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{bmatrix}$$

onde $\sigma_x > 0$, $\sigma_y > 0$ e $\rho \in (-1, 1)$. A densidade conjunta de (X, Y) é dada por

$$f_{XY}(x, y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)}\left[\frac{x^2}{\sigma_x^2} + \frac{y^2}{\sigma_y^2} - \frac{2\rho xy}{(\sigma_x\sigma_y)}\right]\right).$$

Um resultado bem conhecido em probabilidade diz que as distribuições condicionais de uma normal mutivariada também são normais. No caso bivariado temos:

$$X|Y=y \sim N(\mu_x + \sigma_x/\sigma_y\rho(y - \mu_y), \sigma_x^2(1 - \rho^2))$$

e

$$Y|X=x \sim N(\mu_y + \sigma_y/\sigma_x\rho(x - \mu_x), \sigma_y^2(1 - \rho^2)).$$

Isto é, as médias das distribuições condicionais devem ser funções lineares dos valores nos quais condicionamos. Portanto, se existir uma conjunta seguindo uma distribuição normal bivariada, é necessário que

$$\mu_1(y) = a_1 + b_1y \text{ e que } \mu_2(x) = a_2 + b_2x$$

e também que

$$b_1\sigma_2^2 = b_2\sigma_1^2$$

com $b_1 b_2 < 1$. Vemos assim que, se $\mu_1(y) = y^2$ e $\mu_2(x) = x$, não existe uma distribuição conjunta normal bivariada compatível com essa especificação de distribuição condicional. ♠

Exemplo 2.5.3: Como um caso particular do exemplo anterior, assuma que

$$(X|Y=y) \sim N(\beta y, \sigma_1^2)$$

$$(Y|X = x) \sim N(\gamma x, \sigma_2^2).$$

Então, só existe uma conjunta se $\beta\gamma < 1$. Dessa forma, não é possível ter ao mesmo tempo

$$(X|Y = y) \sim N(y, 1) \text{ e } (Y|X = x) \sim N(x, 1)$$

Também não é possível ter

$$(X|Y = y) \sim N(y, 1)$$

(distribuição de X é centrada ao redor de y) e, ao mesmo tempo,

$$(Y|X = x) \sim N(3x, 1)$$

(intuitivamente, distribuição de Y é jogada 3 vezes mais “longe” que x) ♠

Exemplo 2.5.4: Vamos ver um exemplo de especificação condicional bivariada usando a regressão logística. Assuma que Y é binária com $P(Y = 1|X = x) = \pi_x$ e que

$$\log\left(\frac{\pi_x}{1 - \pi_x}\right) = \alpha + \beta x.$$

Suponha que $(X|Y = y) \sim N(\mu_y, \sigma_y^2)$. Então, pelo teorema, existe uma conjunta compatível com estas condicionais se, e somente se, $\sigma_0^2 = \sigma_1^2 = \sigma^2$ e $\beta = (\mu_1 - \mu_0)/\sigma^2$.

Para ver isto, vamos aplicar o teorema. Temos

$$\begin{aligned} a(x, 0) &= \frac{1}{\sqrt{2\pi}\sigma_0} \exp\left(-\frac{1}{2\sigma_0^2}(x - \mu_0)^2\right) \\ a(x, 1) &= \frac{1}{\sqrt{2\pi}\sigma_1} \exp\left(-\frac{1}{2\sigma_1^2}(x - \mu_1)^2\right) \\ b(x, 1) &= \pi_x = 1/(1 + \exp(-\alpha - \beta x)) \\ b(x, 0) &= 1 - \pi_x = 1 - 1/(1 + \exp(-\alpha - \beta x)). \end{aligned}$$

Sabemos que para a especificação ser válida precisamos ter a razão

$a(x, y)/b(x, y) = u(x)v(y)$. Isto é, precisamos ter

$$\frac{a(x, 0)}{b(x, 0)} = \frac{\frac{1}{\sqrt{2\pi}\sigma_0} \exp\left(-\frac{1}{2\sigma_0^2}(x - \mu_0)^2\right)}{1 - 1/(1 + \exp(-\alpha - \beta x))} = u(x)v(0) \quad (2.5.8)$$

$$\frac{a(x, 1)}{b(x, 1)} = \frac{\frac{1}{\sqrt{2\pi}\sigma_1} \exp\left(-\frac{1}{2\sigma_1^2}(x - \mu_1)^2\right)}{1/(1 + \exp(-\alpha - \beta x))} = u(x)v(1). \quad (2.5.9)$$

A razão entre (2.5.8) e (2.5.9) deve ser $u(x)v(0)/u(x)v(1) = v(0)/v(1)$, uma constante, e protanto não pode depender de x . Pedir que a razão entre (2.5.8) e (2.5.9) seja constante é equivalente a pedir que

$$\exp\left(-\frac{1}{2\sigma_0^2}(x - \mu_0)^2 + \frac{1}{2\sigma_1^2}(x - \mu_1)^2 + \alpha + \beta x\right)$$

seja constante em x . Reagrupando os termos no expoente, temos um polinômio de segundo grau em x que deve ser constante em x :

$$\left(-\frac{1}{2\sigma_0^2} + \frac{1}{2\sigma_1^2}\right)x^2 + \left(\frac{\mu_0}{\sigma_0^2} - \frac{\mu_1}{\sigma_1^2} + \beta\right)x = \text{cte.}$$

Isto é, os coeficientes dos termos x^2 e x devem ser nulos. Anular o coeficiente de x^2 implica que $\sigma_1^2 = \sigma_0^2 = \sigma^2$. Impondo esta restrição no coeficiente de X e anulando-o, temos $\beta = (\mu_1 - \mu_0)/\sigma$. Não há nenhuma restrição no termo α da regressão. ♠

Exemplo 2.5.5: Exemplo da impossibilidade de ter uma conjunta bivariada tal que suas duas condicionais sejam Poisson. Suponha que

$$(X|Y = y) \sim \text{Poisson}(\beta y)$$

$$(Y|X = x) \sim \text{Poisson}(\alpha x).$$

Nós assumimos que uma Poisson com média igual a zero é uma distribuição com massa de probabilidade concentrada no valor zero. Usando o teorema, vamos calcular a razão entre as densidades condicionais propostas:

$$\frac{a(x, y)}{b(x, y)} = \frac{(\beta y)^x \exp(-\beta y)/x!}{(\alpha x)^y \exp(-\alpha x)/y!} = \left(\frac{\beta^x e^{\alpha x}}{x!}\right) \left(\alpha^{-y} e^{-\beta y} y!\right) \frac{y^x}{x^y}.$$

Por causa do último fator, não é possível expressar esta razão como um produto $u(x)v(y)$ de funções separáveis. Portanto não existe conjunta neste caso. ♠

2.6 Unicidade

Suponha que $f_{X|Y}(x|y)$ e $f_{Y|X}(y|x)$ sejam compatíveis com pelo menos uma distribuição conjunta $f_{XY}(x,y)$. É natural perguntar se $f_{XY}(x,y)$ é única. Equivalentemente, podemos perguntar: a marginal $f_X(x)$ é única? As perguntas são equivalentes porque se houverem duas distribuições conjuntas $f_{XY}^1(x,y)$ e $f_{XY}^2(x,y)$ distintas e compatíveis, então devemos ter duas densidades marginais para X :

$$f_X^1(x) = f_{XY}^1(x,y)/f_{Y|X}(y|x)$$

e

$$f_X^2(x) = f_{XY}^2(x,y)/f_{Y|X}(y|x).$$

Defina o seguinte núcleo de transição q de uma cadeia de Markov com espaço de estados S_x :

$$q(x|z) = \int_{S_y} f_{X|Y}(x|y) f_{Y|X}(y|z) dy.$$

Então $f_X(x)$ é uma distribuição estacionária dessa cadeia e ela será única se, e somente se, a cadeia é irredutível. Ver Arnold, Castillo e Sarabia (1999) para mais detalhes sobre unicidade no caso bivariado.

2.7 Considerações finais

No caso bivariado, é fácil o estabelecimento da existência e unicidade de uma conjunta a partir das condicionais. No caso multivariado e espacial, os resultados são mais delicados mas muito importantes no contexto atual de modelos com muitos parâmetros via abordagem Bayesiana e MCMC. A solução para o problema geral está no famoso Teorema de Hammersley-Clifford. Besag (1974) é a referência básica para a prova.

Capítulo 3

Campos de Markov

3.1 Grafos

Antes de iniciarmos esse capítulo vamos apresentar alguns conceitos sobre grafos que serão utilizados em seguida. Dados que podem ser representados como grafos são as estruturas ideais para o uso de campos de Markov e distribuições especificadas condicionalmente.

Um grafo $G = (V, E)$ é constituído por um conjunto V de pontos, que chamamos de *vértices* ou de *nós* ou ainda de *sítios*. Os vértices ligam-se entre si por meio de um conjunto E de segmentos de retas, que são denominadas de *arestas*. Dois nós conectados por uma aresta são chamados de *vizinhos*. A Figura 3.1 apresenta um exemplo simples de um grafo com seis vértices, representados pelos círculos, e sete arestas.

A definição de um grafo parece etérea, quase sem conteúdo. Ela não diz o que é um vértice, nem o que uma aresta representa. Este grau de abstração faz com que muitas coisas possam ser vistas como grafos. No contexto espacial, é muito simples representar a estrutura de vizinhança de um mapa em um grafo. Cada área do mapa é associada a um único vértice. Os vértices são conectados por arestas de acordo com algum critério de vizinhança geográfica. O critério mais comum consiste em ligar dois vértices correspondentes a duas áreas quando as mesmas possuem uma fronteira comum. Para exemplificar, a Figura 3.2 apresenta à esquerda o mapa da cidade de Vitória dividido por setor censitário. À direita, encontra-se o grafo correspondente a esse mapa

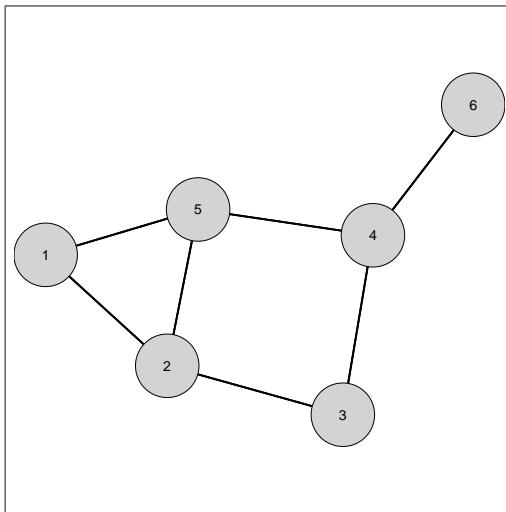


Figura 3.1: Exemplo de grafo com seis vértices e sete arestas.

considerando como critério de vizinhança a presença de fronteira entre as áreas. Cada área é representada por um vértice localizado no seu centro de massa (também chamado centróide).

Muitas situações não-geográficas podem ser representadas como grafos. Redes de transporte ou comunicação constituem o exemplo mais óbvio: tome as cidades como vértices e deixe as arestas conectarem as cidades que possuem alguma estrada ligando-as.

Outro exemplo importante nos dias de hoje são as páginas web sendo nós com arestas representando a existência de algum link entre elas. Podemos também conectar usuários de uma rede social da web. A Figura 3.3 é uma ilustração da conectividade da web na forma de grafos. No grafo da esquerda, cada vértice representa uma pessoa que digitou o termo *SharePoint*, nome de uma plataforma de colaboração Windows voltada para aplicações intranet, no período próximo a 28 de dezembro de 2008.

No grafo da direita na Figura 3.3, temos o resultado de uma *query* no programa TouchGraph com a palavra *ufmg*. Ele cria um grafo com as principais (mais acessadas) páginas em que aparecem a palavra procurada e as páginas similares a ela. Similaridade nesse caso não significa apenas links ligando as

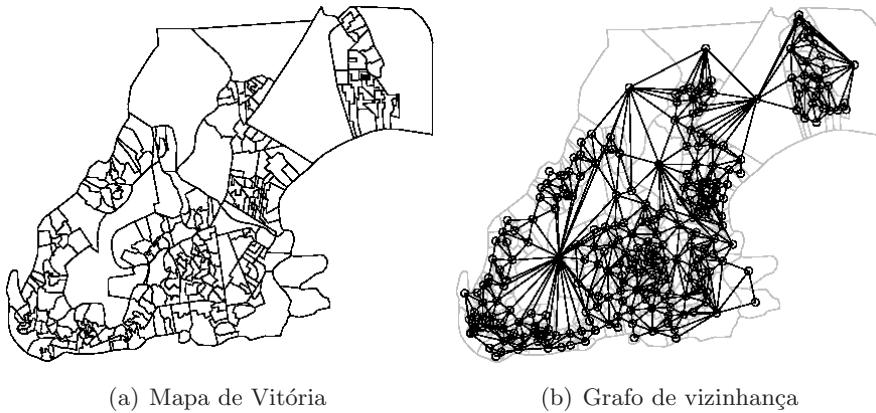


Figura 3.2: Mapa de Vitória e grafo de vizinhança correspondente.

duas páginas: elas também podem ser ligadas no grafo se existirem outros termos em comuns mencionados pelas duas páginas.

Economistas, sociólogos, demógrafos e outros estudiosos têm estudado como a interação entre pessoas (ou entre empresas ou outras entidades complexas) podem explicar vários fenômenos que são observados. Estas interações são representadas como grafos. Elas refletem ligações baseadas em colaboração ou competição que criam redes de ligações sociais. A idéia neste tipo de estudo é que um conjunto de indivíduos são os nós de uma rede (ou grafo). As arestas do grafo refletem as relações entre os indivíduos. Eles fazem escolhas e agem a partir de um conjunto de alternativas disponíveis. Existe incerteza sobre os ganhos obtidos em cada possível ação que o indivíduo pode tomar. Para tomar decisões frente à incerteza, esses indivíduos usam informação própria e informação obtida de seus *vizinhos*, os indivíduos ligados a eles na rede social. Seguindo a teoria da escolha racional, escolhem ação que maximiza a utilidade individual.

O ponto relevante nestes estudos é que a *estrutura* da rede, sua topologia, influencia as decisões individuais. A topologia da rede induz distribuições de probabilidade que levam em conta essa configuração espacial de interrelações. Um exemplo clássico é devido a James S. Coleman, o brilhante sociólogo responsável pelo famoso *Coleman Report* sobre a igualdade de oportunidade na

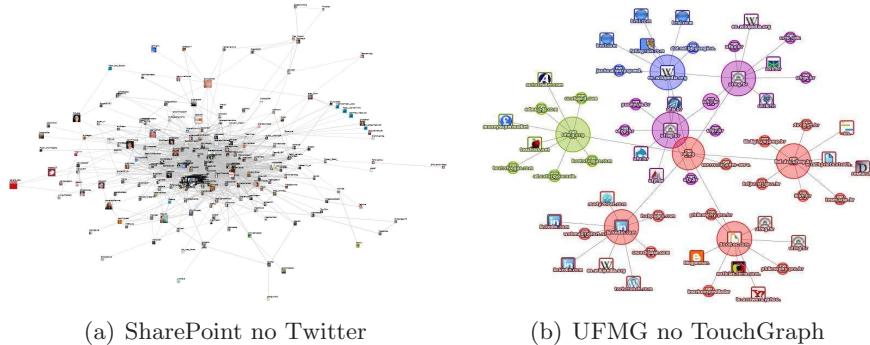


Figura 3.3: Esquerda: Grafo baseado nas conexões do Twitter entre aqueles que mencionaram o termo *SharePoint* no período próximo a 28 de dezembro de 2008. Direita: Grafo de conexões entre as principais páginas em que aparece a palavra *ufmg*.

educação. Em seu outro livro *Medical Innovation: A Diffusion Study*, Coleman, Katz e Menzel (1966) estudaram como inovações médicas ganhavam ampla aceitação após um certo tempo. Médicos decidem recomendar produtos sem um conhecimento completo dos mesmos. Eles buscam informação na literatura profissional e entre os colegas de profissão. *Ceteris paribus*, os médicos mais conectados são aqueles que passam a recomendar produtos melhores mais rapidamente. O objetivo principal de Coleman e colegas foi quantificar esta associação, como a posição de um indivíduo na sua rede de relações sociais determinava quando ele ia adotar a inovação. Eles também procuraram estabelecer a dinâmica induzida pelas diversas topologias de rede social na difusão da inovação médica.

Outro exemplo clássico de análise de redes, também é devido a James Coleman, está no seu livro *The Adolescent Society* (Coleman, 1961). Ele estava preocupado em entender como a organização da vida adolescente levava a uma cultura anti-aprendizagem e o que podia ser feito a respeito disso. Na época, a guerra fria estava em pleno curso e os Estados Unidos estavam preocupados com seu desempenho pífio em ciências e com o crescimento dos soviéticos. Entre outras ferramentas de análise, Coleman usou os grafos de relacionamento entre os estudantes de diversas escolas para explorar estas

questões. A Figura 3.4 mostra o grafo de amizade recíproca entre meninos de uma escola americana. Se A dissesse que era amigo de B e B também dissesse ser amigo de A, eles eram conectados por uma aresta.

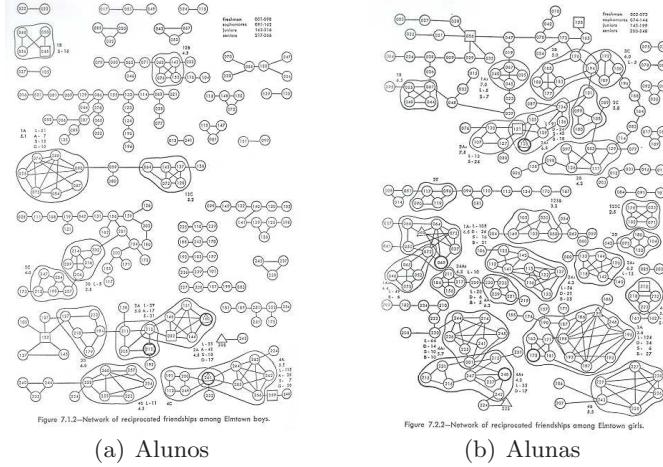


Figura 3.4: Grafos de amizade recíproca entre meninos (esquerda) e meninas (direita) de uma escola americana.

Podemos também atribuir pesos w_{ij} às arestas que conectam os vértices i e j . No caso espacial, esses pesos podem depender, por exemplo, da extensão da fronteira dividida entre as áreas ou da distância entre seus centróides ou de alguma outra característica tal como a diferença entre a renda per capita das duas áreas vizinhas.

Muitos problemas práticos estão associados com uma estrutura de grafo. Por exemplo, pode-se desejar encontrar o trajeto de menor custo entre duas cidades. Grafos podem também ser utilizados em problemas mais complexos, como identificar padrões de associação em grandes redes de relacionamento como a internet. De acordo com o problema tratado, as arestas do grafo podem apresentar ou não uma determinada direção. Quando estamos tratando com fluxo de mercadorias entre dois locais, as arestas são direcionadas. Trataremos aqui apenas com grafos não direcionados pois estamos considerando apenas relações de vizinhança simétrica entre determinados locais.

Em grafos mais gerais, pode existir mais de uma aresta ligando dois vértices

distintos ou pode existir uma aresta ligando um vértice a ele mesmo. Serão considerados aqui apenas grafos denominados simples, para os quais existe no máximo uma aresta ligando um par de vértices e nenhum vértice se conecta a ele mesmo. Para obter mais informações sobre esse tópico consultar Wilson(1997). Daqui por diante, vamos à vezes nos referir ao grafo como um mapa, mas o leitor deve lembrar-se que ele pode representar uma estrutura não-geográfica.

3.2 Vizinhança e cliques

Associado a cada vértice i , temos um conjunto de vizinhos, que denotaremos por \mathcal{N}_i . A relação binária de vizinhança é denotada pelo símbolo \sim . Isto é, $i \sim j$ se os vértices i e j são vizinhos. Vamos representar o mapa dos vértices e a sua estrutura de vizinhança através de um grafo $G = (V, E)$ onde as arestas indicam quem é vizinho de quem.

Duas restrições básicas são colocadas na estrutura de vizinhança. Em primeiro lugar,

- i não pertence a \mathcal{N}_i ,

ou seja, um sítio não é vizinho de si mesmo. A outra condição que vamos impor é que

- \mathcal{N}_i não é vazio.

Cada área possui pelo menos um vizinho. Isto significa que o grafo de vizinhança não é desconectado em dois ou mais sub-grafos. Vamos também assumir que as relações de vizinhança são simétricas:

- se $j \in \mathcal{N}_i$, então $i \in \mathcal{N}_j$.

Vamos utilizar o termo *clique* para denotar um conjunto de sítios em que cada um deles é vizinho de todos os demais. Na literatura referente a grafos, essa estrutura recebe o nome de subgrafo completo. Por definição, nós consideramos o conjunto unitário formado por um único sítio como sendo uma *clique*. Também consideramos o conjunto nulo \emptyset como uma clique.

Imagine um mapa formado por um tabuleiro de xadrez com as áreas sendo os quadrados por onde as peças circulam (ver Figura 3.5). Este mapa é representado por uma grade regularmente espaçada. Para uma grade regular como esta, é comum definir estruturas de vizinhanças a partir de um raio r de distância:

$$\mathcal{N}_i = \{j \in G \text{ tais que } d(i, j) \leq r\}$$

onde $d(i, j)$ é a distância euclidiana entre os vértices i e j .

Vamos considerar duas estruturas de vizinhança. A primeira delas é chamada de estrutura de primeira ordem e é dada pelo movimento do peão:

$$\mathcal{N}_i^1 = \{j \in G \text{ tais que } d(i, j) \leq 1\}$$

(casa A na Figura 3.5). Isto é, os vizinhos da área/casa (x, y) são as áreas/casas $(x-1, y), (x+1, y), (x, y-1), (x, y+1)$ imediatamente ao norte, sul, leste e oeste da área em questão. A segunda delas é chamada de estrutura de segunda ordem e é a estrutura de vizinhança é definida a partir do movimento do rei:

$$\mathcal{N}_i^2 = \left\{ j \in G \text{ tais que } d(i, j) \leq \sqrt{2} \right\}$$

(casa B na Figura 3.5). Quais são as cliques definidas num sistema de vizinhança definido desta forma?

Na estrutura de vizinhança de primeira ordem, também chamada de sistema de 4-vizinhos, todo vértice no interior do grafo tem quatro vizinhos, como mostrado na Figura 3.6(a), onde x denota o vértice considerado e Ω denota os seus vizinhos. Um sítio na borda possui três vizinhos e cada um dos quatro sítios nos cantos possui dois vizinhos.

Na estrutura de segunda ordem, ou sistema de 8 vizinhos, existem 8 vizinhos para cada vértice no interior do grafo, como mostrado na Figura 3.6(b). Os números $n = 1, \dots, 5$ mostrados na 3.6(c) indicam os sítios mais longínquos num sistema de vizinhança de n -ésima ordem. Todos os sítios com números menores ou iguais a n fazem parte da vizinhança \mathcal{N}_x do sítio x .

As cliques da estrutura de primeira ordem de 4-vizinhos são os subconjuntos unitários formados pelos sítios isolados ou os subconjuntos formados pelos pares de vizinhos horizontais ou verticais (ver (d) e (e) na Figura 3.6). As cliques para a vizinhança de segunda ordem de 8-vizinhos incluem não apenas

A= movimento do peão, B = movimento do rei

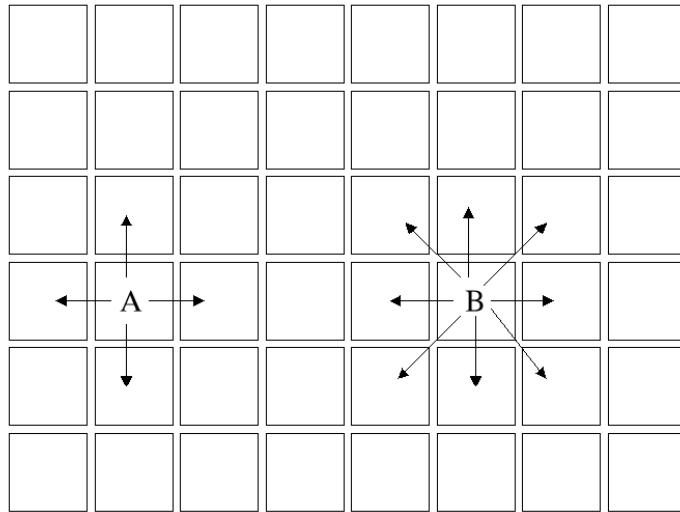


Figura 3.5: Mapa regular apresentando duas possíveis formas de definir vizinhança, \mathcal{N}_A e \mathcal{N}_B , para as áreas A e B . A área A possui a vizinhança dada pelo movimento do peão no jogo de xadrez e a área B possui a vizinhança dada pelo movimento do rei.

aqueles subconjuntos da forma (d) e (e) mas também cliques de pares em diagonal (f), cliques de triplas de vizinhos (g) e cliques de quádruplas de vizinhos (h). A medida que a ordem do sistema de vizinhança cresce, o número de cliques cresce rapidamente.

Cliques para grades irregulares não possuem um formato fixo como aqueles em grades regulares. Considere os quatro sítios f , i , m e n dentro do círculo na Figura 3.7. Os vértices m e n são considerados vizinhos entre si, bem como n e f . Os sítios f e m não são vizinhos por estarem muito longe um do outro. Considerando apenas estes quatro vértices, as cliques formadas por um sítio, pares de sítios e triplas de sítios associadas com este subconjunto de vértices

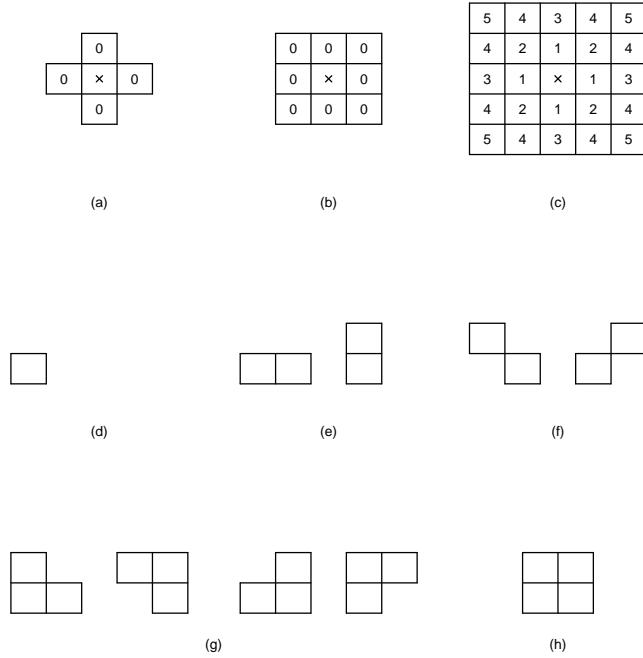


Figura 3.6: Vizinhança e cliques de uma grade regular de vértices.

são mostradas na Figura 3.7. O conjunto $\{m, i, f\}$ não forma uma clique porque f e m não são vizinhos.

3.3 Campos aleatórios de Markov

Os Campos Aleatórios de Markov são definidos em uma estrutura de grafo na qual se define como vizinhos aqueles sítios que estão conectados por uma aresta. Em cada nó ou sítio do grafo, observamos uma ou mais variáveis aleatória. Para efeito de exposição, vamos tratar apenas do caso univariado. Nesse tipo de modelo a distribuição condicional da variável observada em um dado sítio depende apenas de seus vizinhos. Dessa forma, temos um modelo semelhante a uma cadeia de Markov. A diferença reside no fato de estarmos lidando aqui com variáveis aleatórias no espaço (ou num grafo) e não no tempo.

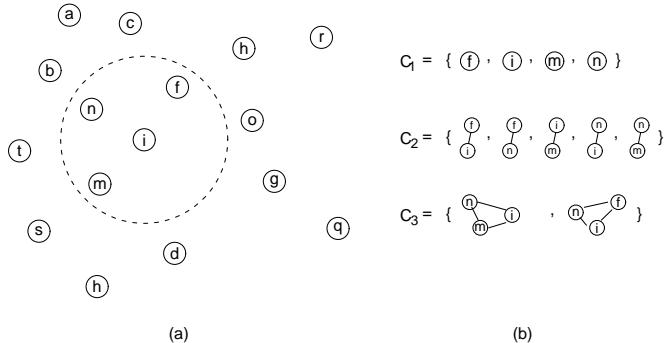


Figura 3.7: Vizinhança e cliques de uma grade irregular de sítios.

Andrei Andreyevich Markov nasceu em 14 de junho de 1856 em Ryazan, Russia, filho de um funcionário público. Mostrou grande talento para a matemática tendo escrito aos 17 anos um artigo sobre a integração de equações diferenciais lineares. Os resultados não eram novos mas serviram para torná-lo conhecido de importantes matemáticos russos. Ele estudou na Universidade de São Petersburgo, foi aluno de Chebyshev e graduou-se em 1878 com medalha de ouro por seu trabalho On the integration of differential equations by means of continued fractions. Durante os dois próximos anos, dedicou-se ao mestrado como o intuito de tornar-se um professor universitário. Em 1880, ele recebe o título apresentando a dissertação On the binary quadratic forms with positive determinant, a qual foi considerada um grande avanço em matemática. Em seguida, ele começa a dar aulas na Universidade São Petersburgo, termina seu doutorado, é promovido e, anos mais tarde, é eleito membro da Academia Russa de Ciências. Seus trabalhos iniciais foram principalmente na área de teoria dos números, análise, funções contínuas algébricas, limites de integrais, teoria de aproximação e convergência de séries. Somente após 1890, seguindo seu orientador Pafnutiy Chebyshev, ele passa a aplicar métodos de frações continuas à Teoria da Probabilidade. Se destacou pelas suas pesquisas relacionadas à Lei dos Grandes Números e método de Mínimos Quadrados. Ele é particularmente lembrado por seus estudos sobre as Cadeias de Markov. Esse trabalho fundou uma área completamente nova em teoria da probabilidade e iniciou a teoria dos Processos Estocásticos. Ao longo de sua vida, esteve envolvido em polêmicas religiosas contra a igreja ortodoxa e em polêmicas políticas contra os czares. Em 1913 solicita, e é atendido, sua excomunhão da igreja ortodoxa em solidariedade a Leon Tolstoi, que também havia sido excomungado. Ele chegou a presenciar os primeiros anos da revolução russa de 1917 quando vai voluntariamente para o interior ensinar matemática para crianças. Morreu em 1921 após meses de grande sofrimento físico.



3.3.1 Markov no tempo

É útil considerar o caso de uma cadeia de Markov no tempo. Tradicionalmente, a cadeia é definida considerando a distribuição de x_t dado todo o passado da cadeia. No modelo de Markov, esta distribuição depende apenas da última observação. Um exemplo clássico de cadeia de Markov com espaço de estados contínuo é o modelo $AR(1)$ de séries temporais com média zero:

$$x_t | x_{t-1}, x_{t-2}, \dots, x_1 \sim N(\phi x_{t-1}, \sigma^2) .$$

O objetivo prático de trabalhar com previsões a partir de dados já observados é a principal motivação para isto. O que é menos comum é apresentar a distribuição de x_t dado todo o passado e o *futuro* da cadeia. Vamos mostrar que, para um tempo t que não seja o primeiro nem o último de uma série temporal, temos

$$x_t | x_n, x_{n-1}, \dots, x_{t+1}, x_{t-1}, x_{t-2}, \dots, x_1 \sim N\left(\frac{\phi(x_{t-1} + x_{t+1})}{1 + \phi^2}, \frac{1}{1 + \phi^2}\right) . \quad (3.3.1)$$

Assim, num sentido óbvio, o instante de tempo $t-1$ e $t+1$ são os vizinhos do instante de tempo t num grafo de vizinhança linear que conecta sucessivamente os tempos de 1 a n .

Para mostrar (3.3.1), vamos obter inicialmente a distribuição conjunta da série. Dado todos os valores anteriores $x_{t-1}, x_{t-2}, \dots, x_1$, temos

$$x_t = \phi x_{t-1} + \epsilon_t, \quad \epsilon_t \stackrel{iid}{\sim} N(0, \sigma^2) .$$

Isto é,

$$x_t | x_{t-1}, x_{t-2}, \dots, x_1 \sim N(\phi x_{t-1}, \sigma^2) . \quad (3.3.2)$$

Vamos assumir que a distribuição marginal de x_1 é normal com média zero e variância $1/(1 - \phi^2)$, que é a distribuição estacionária do processo.

Utilizando a representação da distribuição conjunta na forma (1.6.7) e a propriedade markoviana (1.6.9) ou (3.3.2) acima, a distribuição conjunta

dessas variáveis pode ser calculada como

$$\begin{aligned} f(\mathbf{x}) &= f(x_1)f(x_2|x_1)f(x_3|x_2)\dots f(x_n|x_{n-1}) \\ &= \left(\frac{1-\phi^2}{2\pi\sigma^2}\right)^{1/2} \exp\left(-\frac{1-\phi^2}{2\sigma^2}x_1^2\right) \prod_{t=2}^n \frac{1}{2\pi\sigma^2} \exp\left(-\frac{1}{2\sigma^2}(x_t - \phi x_{t-1})^2\right) \\ &\propto \exp\left(-\frac{1}{2\sigma^2}\left((1-\phi^2)x_1^2 + \sum_{t=2}^n (x_t - \phi x_{t-1})^2\right)\right). \end{aligned}$$

Para obter a condicional de x_t dado todos os demais valores da série, nós podemos ignorar todos os termos multiplicativos que não envolvem x_t na expressão acima:

$$\begin{aligned} f(x_t|x_n, \dots, x_{t+1}, x_{t-1}, \dots, x_1) &\propto \exp\left(-\frac{1}{2\sigma^2}[(x_t - \phi x_{t-1})^2 + (x_{t+1} - \phi x_t)^2]\right) \\ &\propto \exp\left(-\frac{1+\phi^2}{2\sigma^2}\left(x_t^2 - 2x_t \frac{\phi}{1+\phi^2}(x_{t-1} + x_{t+1})\right)\right) \\ &\propto \exp\left(-\frac{1+\phi^2}{2\sigma^2}\left(x_t - \frac{\phi}{1+\phi^2}(x_{t-1} + x_{t+1})\right)^2\right) \end{aligned}$$

onde concluímos que a distribuição condicional é

$$x_t|x_n, \dots, x_{t+1}, x_{t-1}, \dots, x_1 \sim N\left(\frac{\phi}{1+\phi^2}(x_{t-1} + x_{t+1}), \frac{1}{1+\phi^2}\right).$$

Fazendo este mesmo tipo de cálculo, encontramos a distribuição nos dois extremos da série:

$$x_t|\mathbf{x}_{-t} = \begin{cases} N(\phi x_2, 1) & \text{se } t = 1 \\ N\left(\frac{\phi}{1+\phi^2}(x_{t-1} + x_{t+1}), \frac{1}{1+\phi^2}\right) & \text{se } 1 < t < n \\ N(\phi x_{n-1}, 1) & \text{se } t = n \end{cases}.$$



Essa estrutura de vizinhança é simples por causa da uni-direcionalidade do tempo, o que não é o caso do espaço.

3.3.2 Caso Markov em duas dimensões

Vamos assumir daqui para frente que temos n nós ou sítios, os quais iremos denotar por i, j, k, \dots . As medidas aleatórias feitas em cada um desses pontos

serão denotadas por X_i, X_j, X_k, \dots . Seja $(X_1, \dots, X_n) = \mathbf{X}$ o vetor para o qual queremos definir uma distribuição de probabilidade com um caráter espacial. Suponha que alguma estrutura de vizinhança foi estabelecida pelo usuário. Vamos denotar por \mathbf{X}_{-i} o vetor de dimensão $n - 1$ formado por todas as variáveis exceto a i -ésima:

$$\mathbf{X}_{-i} = (X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_k).$$

Definição 3.3.1. *A distribuição de \mathbf{X} possui a propriedade markoviana com respeito ao sistema de vizinhança representado por $\{\mathcal{N}_i, i = 1, \dots, n\}$, e é chamada de campo aleatório de Markov, se, para todo $i, i = 1, \dots, n$, a distribuição conjunta satisfaçá à seguinte propriedade:*

$$f(x_i | \mathbf{X}_{-i} = \mathbf{x}_{-i}) = f(x_i | \{x_j \text{ tais que } j \in \mathcal{N}_i\}). \quad (3.3.3)$$

Isto é, \mathbf{X} possui uma distribuição de Markov se a distribuição de X_i condicionada em todo o restante do mapa for a mesma que a distribuição de X_i condicionada apenas nos valores de áreas vizinhas de i . É importante notar que a expressão (3.3.3) engloba n diferentes distribuições de probabilidade que devem ser compatíveis entre si. Afinal, a distribuição de i é definida dependendo dos valores de um vizinho j mas a distribuição deste vizinho j também é definida em termos de i . Isto não ocorre em séries temporais (ver Capítulo 1).

Quem é definido como vizinho de i é, em princípio, deixado a critério do usuário destes modelos. Além disso, duas áreas j e k , ambas pertencentes a \mathcal{N}_i , poderão ter contribuições ou pesos diferentes na determinação do valor de X_i . Estes pesos distintos vão aparecer na expressão da densidade condicional (3.3.3), como veremos nos exemplos nos próximos capítulos.

Estabelecido um sistema de vizinhança, obter uma distribuição conjunta que satisfaça à propriedade (3.3.3) não parece ser uma tarefa trivial, como já observamos no Capítulo 1. Afinal de contas, a propriedade (3.3.3) refere-se a cada uma das áreas i do mapa e portanto uma distribuição conjunta markoviana $f(\mathbf{x})$ terá que satisfazer a um sistema de n distribuições condicionais da forma (3.3.3) simultaneamente. A solução para esse problema foi apresentada no Teorema de Hammersley e Clifford.

3.4 Energia e densidades

Toda densidade conjunta $f(\mathbf{x})$ pode ser escrita na forma de uma função exponencial $\exp(-U(\mathbf{x}))$ para alguma função $U(\mathbf{x})$ nos pontos \mathbf{x} do suporte de f (isto é, nos pontos em que $f(\mathbf{x}) > 0$). Isto é trivial já que sempre podemos escrever $f(\mathbf{x}) = \exp(\log(f(\mathbf{x})))$ e portanto, $U(\mathbf{x}) = -\log(f(\mathbf{x}))$.

O que se ganha ao escrever a densidade desta forma? Os físicos e probabilistas preferem esta forma exponencial de escrever as densidades quando estão lidando com distribuições para variáveis identificadas com nós de uma rede. A situação típica é que em cada ponto da rede existe uma partícula (átomo ou carga elétrica, por exemplo) e elas interagem entre si (via a força induzida pelos seus campos magnéticos ou elétricos, por exemplo).

O sistema como um todo possui uma energia $U(\mathbf{x})$ associada com a configuração do mesmo. Esta energia depende das posições relativas das partículas, do valor que cada partícula possui (sua carga ou sua magnetização), do campo criado por cada uma delas, da possível existência de uma força externa ao sistema, etc.

A chance de observar o sistema numa determinada configuração depende dessa energia $U(\mathbf{x})$. Algumas configurações são muito instáveis e dificilmente são observadas. Outras configurações, sendo mais estáveis, são observadas com mais frequência. De qualquer modo, físicos preferem pensar em termos da energia associada com um sistema e esta energia está associada com a chance de observar o sistema numa dada configuração. Por isto, é comum escrever a densidade de probabilidade da configuração \mathbf{x} como

$$f(\mathbf{x}) = \frac{1}{Z} \exp(-U(\mathbf{x})) \propto \exp(-U(\mathbf{x})).$$

onde Z é uma constante com respeito a \mathbf{x} . Estas densidades colocam mais massa de probabilidade em configurações com baixos níveis de energia. Quanto menor a energia $U(\mathbf{x})$ do sistema numa dada configuração específica \mathbf{x} , maior a probabilidade de observação desta configuração num momento qualquer. O sistema tende a preferir configurações com baixo estado de energia.

A constante de normalização Z é chamada de função de partição. Quando o número de possíveis estados x_i para os vértices $i = 1, \dots, n$ é finito, a função de partição sempre existe. Isto é verdade porque se as variáveis aleatórias

X_1, \dots, X_n assumem valores no conjunto E com um número k finito de elementos ($k = 2$ se os $E = \{0, 1\}$, por exemplo), então $\Omega = E^n = E \times \dots \times E$ também é finito com k^n elementos. Assim,

$$Z = \sum_{\mathbf{x} \in \Omega} \exp(-U(\mathbf{x}))$$

é uma soma finita de k^n elementos.

Se E é infinito (enumerável ou não-enumerável), a constante $Z(\beta)$ acima é uma série ou uma integral, caso E seja um intervalo da reta. Neste caso, Z pode ser infinito e portanto a densidade não vai existir. Nestes casos, uma função de energia $U(\mathbf{x})$ não foi bem escolhida.

Quando os n vértices não possuem nenhuma interação, o vetor aleatório $\mathbf{X} = (X_1, \dots, X_n)$ é constituído de variáveis aleatórias independentes e neste caso a energia $U(\mathbf{x})$ é simplesmente uma soma de n funções, uma para cada vértice. Para ver isto, basta notar que, sob independência, a densidade conjunta é o produto das densidades marginais e portanto podemos escrever

$$\begin{aligned} f(\mathbf{x}) &= \prod_{i=1}^n f_{X_i}(x_i) = \exp\left(\sum \log(f_{X_i}(x_i))\right) \\ &= \exp\left(-\sum_i U(x_i)\right) = \exp(-U(\mathbf{x})). \end{aligned} \quad (3.4.4)$$

Neste caso, a energia total é simplesmente a soma das energias de cada partícula. No entanto, quando existe interação entre as partículas, isto não é razoável. Imagine, por exemplo, que temos um conjunto de vértices dispostos sequencialmente ao longo de uma linha reta nas posições $1, 2, \dots, 2n$. Cada vértice pode assumir aleatoriamente o valor $+1$ ou -1 . Considere duas configurações possíveis:

- (i) $\mathbf{x} = (+1, +1, \dots, +1, -1, -1, \dots, -1)$
- (ii) $\mathbf{x} = (+1, -1, +1, -1, +1, -1, \dots, +1, -1)$.

Ambas possuem o mesmo número total n de $+1$'s e -1 's. Se a energia total é simplesmente a soma das energias individuais, então as duas configurações possuem energias idênticas e portanto, probabilidades idênticas de ocorrer.

Entretanto, se este é um sistema em que as partículas são como cargas elétricas, elas tendem a repelir valores iguais. Portanto, cargas de mesmo sinal e muito próximas seriam improváveis e deveríamos esperar que uma configuração alternada como aquela em (ii) é mais provável que aquela outra em (i).

Um modelo ligeiramente mais complicado que este supõe que as partículas possuem forças de curto alcance de forma que uma partícula só é capaz de influenciar diretamente as outras partículas na sua vizinhança imediata. Isto leva a introdução de efeitos de pares de vizinhos na expressão (3.4.4) de forma que escrevemos:

$$f(\mathbf{x}) \propto \exp \left(- \sum_i U_1(x_i) - \sum_{i \sim j} U_2(x_i, x_j) \right). \quad (3.4.5)$$

Por exemplo, nós poderíamos ter um sistema em que os valores de dois sítios vizinhos são os mesmos (ambos +1 ou ambos -1), a energia é igual a $\alpha > 0$ e quando os dois sítios vizinhos são diferentes, a energia é igual a $\alpha < 0$:

$$U_2(x_i, x_j) = \begin{cases} \alpha & \text{se } x_i = x_j \\ -\alpha & \text{se } x_i \neq x_j \end{cases}. \quad (3.4.6)$$

Assim,

$$f(\mathbf{x}) \propto \exp \left(- \sum_i U_1(x_i) - \alpha \sum_{i \sim j} x_i x_j \right) = \exp \left(- \sum_i U_1(x_i) - \alpha(n_s - n_d) \right)$$

onde n_s e n_d são os números de pares de sítios com valores iguais e com valores diferentes, respectivamente. Note que no caso das partículas disposta em linha existem apenas $n-1$ pares de vizinhos ao todo e que $n_s + n_d = n-1$. Portanto, $n_s - n_d = 2n_s - (n-1)$.

É claro que várias outras modificações são possíveis a partir deste exemplo simples. Uma possibilidade seria admitir que a força de interação entre as partículas (o valor de α) muda a medida que nos deslocamos ao longo da reta onde elas estão. Assim, pares localizados num extremo do segmento teriam um valor de α para seu estado (x_i, x_{i+1}) diferente daquele de pares localizados

no outro extremo. Isto é, a função $U_2(x_i, x_j)$ em (3.4.4) dependeria do par (i, j) de sítios e deveria ser mais apropriadamente escrita como $U_{ij}(x_i, x_j)$.

Isto poderia valer também para a energia individual das partículas. De alguma forma, o valor de +1 num extremo do segmento tem um impacto diferente do mesmo valor de +1 no outro extremo. Dessa forma, a maneira mais geral de escrever a densidade (3.4.5) onde somente interações de pares entram seria a seguinte:

$$f(\mathbf{x}) \propto \exp \left(- \sum_i U_i(x_i) - \sum_{i \sim j} U_{ij}(x_i, x_j) \right). \quad (3.4.7)$$

Outra modificação é permitir que cada partícula tenha uma influência direta em partículas mais distantes que apenas uma unidade. Por exemplo, poderímos modificar (3.4.7) introduzindo triplas de sítios consecutivos:

$$f(\mathbf{x}) \propto \exp \left(- \sum_i U_i(x_i) - \sum_{i \sim j} U_{ij}(x_i, x_j) - \sum_{i \sim j \sim k} U_{ijk}(x_i, x_j, x_k) \right). \quad (3.4.8)$$

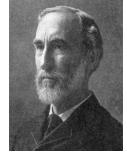
Isto pode continuar em ordens mais elevadas.

3.5 Distribuições de Gibbs

Nós vamos generalizar a noção de energia. Seja A um subconjunto não-vazio dos índices $\{1, \dots, n\}$ dos sítios. Seja $\mathbf{x}_A = (x_i, i \in A)$ o vetor que contém apenas as coordenadas correspondentes aos sítios em A . Vamos denotar por $\mathbf{x}_{-A} = (x_i, i \notin A)$ o vetor com as demais coordenadas, aquelas que não estão em A .

Suponha que \mathcal{A} seja uma coleção de subconjuntos distintos e não-nulos, cada subconjunto composto por elementos do conjunto dos índices $\{1, \dots, n\}$. Suponha também que, para cada cada subconjunto $A \in \mathcal{A}$, temos uma função U_A tal que $U_A(\mathbf{x}) = U_A(\mathbf{x}_A)$. Isto é, U_A só depende dos valores x_i dos sítios $i \in A$.

Josiah Willard Gibbs nasceu em 11 de Fevereiro de 1839 e fez sua carreira nos Estados Unidos quando ele ainda não era uma potência científica. Ele foi um dos principais responsáveis pelo estabelecimento dos fundamentos teóricos da termodinâmica estatística, o estudo do comportamento macroscópico de gases (via variáveis tais como pressão, volume e temperatura) a partir do comportamento microscópicos de suas moléculas. Ele iniciou sua vida universitária em 1854 na Universidade de Yale, recebendo de lá o título de Ph.D. em engenharia em 1863, o primeiro dos EUA. Ele passou toda sua carreira em Yale. Em 1876, ele publicou a primeira parte de seu trabalho de maior destaque: On the Equilibrium of Heterogeneous Substances. Na época, os EUA não tinham grande interesse científico e ele não teve grande reconhecimento em vida no seu país. Em sua biografia de Gibbs, Crowther diz que ele "remained a bachelor, living in his surviving sister's household. In his later years he was a tall, dignified gentleman, with a healthy stride and ruddy complexion, performing his share of household chores, approachable and kind (if unintelligible) to students. Gibbs was highly esteemed by his friends, but U.S. science was too preoccupied with practical questions to make much use of his profound theoretical work during his lifetime. He lived out his quiet life at Yale, deeply admired by a few able students but making no immediate impress on U.S. science commensurate with his genius." Ele faleceu em 28 de April de 1903 na casa que seu pai construiu e em que ele passou toda sua vida.



Por exemplo, na seção anterior, a densidade (3.4.7) possui funções U_A para subconjuntos A formados por um único índice (isto é, do tipo $A = \{i\}$) e por subconjuntos A formados por pares de áreas vizinhas (isto é, do tipo $A = \{i, i + 1\}$ para i entre 1 e $n - 1$). Assim, neste caso particular,

$$\mathcal{A} = \{\{1\}, \{2\}, \dots, \{n\}, \{1, 2\}, \dots, \{n - 1, n\}\}$$

Definição 3.5.1. Um potencial de interação é uma família

$$\mathcal{U} = \{U_A \text{ tal que } A \in \mathcal{A}\}$$

de funções $U_A : \mathbf{x} \rightarrow \mathbb{R}$ tal que a $U_A(\mathbf{x}) = U_A(\mathbf{x}_A)$ e a soma

$$U(\mathbf{x}) = \sum_{A \in \mathcal{A}} U_A(\mathbf{x}_A) \tag{3.5.9}$$

existe para \mathbf{x} . U_A é o potencial em A .

Quando a energia $U(\mathbf{x})$ puder ser escrita da forma (3.5.9), nós estaremos quebrando a energia total de uma configuração em componentes associados com diferentes subconjuntos de sítios.

Definição 3.5.2. *Uma medida de probabilidade de Gibbs induzida por um potencial de interação \mathcal{U} é definida por*

$$f(\mathbf{x}) = \frac{1}{Z} \exp \left(- \sum_{A \in \mathcal{A}} U_A(\mathbf{x}_A) \right) = \frac{1}{Z} \prod_{A \in \mathcal{A}} \exp(-U_A(\mathbf{x}_A)) . \quad (3.5.10)$$

Especificar uma distribuição de Gibbs implica ter de especificar os subconjuntos $A \in \mathcal{A}$ de nós que aparecem na densidade conjunta (3.5.10).

3.6 Distribuições de Gibbs e campos de Markov

Dada uma distribuição de Gibbs (3.5.10) com respeito a um potencial de interação, podemos encontrar a densidade condicional de um sítio dados todos os demais:

$$\begin{aligned} f(x_i | \mathbf{x}_{-i}) &= \frac{f(\mathbf{x})}{f(\mathbf{x}_{-i})} \\ &= \frac{f(\mathbf{x})}{\int f(x_1, \dots, x_{i-1}, y, x_{i+1}, \dots, x_n) dy} . \end{aligned}$$

Na razão acima, todos os fatores do produto (3.5.10) que não envolvem a i -ésima coordenada são cancelados. Isto é, todos os termos $\exp(-U_A(\mathbf{x}_A))$ em que o sítio $i \notin A$ são cancelados. Assim, a densidade condicional derivada de uma distribuição de Gibbs pode depender apenas de sítios j tais que i e j compartilhem pelo menos um conjunto A do potencial de interação. Os demais sítios, que não fazem parte de nenhum conjunto A no qual i também esteja presente, não podem ser vizinhos de i se a distribuição de Gibbs é um campo de Markov com respeito a alguma estrutura de vizinhança.

Suponha que a família \mathcal{A} de subconjuntos de índices de vértices possui todos os conjuntos unitários formados pelos vértices isolados. Isto é, suponha que $\{i\} \in \mathcal{A}$ para todo $i = 1, \dots, n$. Associamos um grafo de vizinhança a

uma família \mathcal{A} desse tipo da seguinte forma: conecte os sítios i e j por uma aresta se existir pelo menos um $A \in \mathcal{A}$ tal que $\{i, j\} \subseteq A$. Dizemos que esta é a estrutura de vizinhança induzida por \mathcal{A} .

O Teorema de Hammersley-Clifford esclarece qual a relação entre os campos de Markov e as distribuições de Gibbs com potenciais de interação em que \mathcal{A} contém todos os nós isolados. Ele será provado no Capítulo 6 e diz o seguinte: uma distribuição de Gibbs com potencial de interação \mathcal{A} é um campo de Markov com estrutura de vizinhança induzida por \mathcal{A} se, e somente se, todo subconjunto $A \in \mathcal{A}$ for uma clique.

Capítulo 4

Campos de Marvok Gaussianos

4.1 Definição

Neste capítulo, nós vamos apresentar um tipo específico de campo de Markov. São aqueles em que as variáveis aleatórias possuem uma distribuição conjunta normal multivariada, conhecidos como GMRF (*Gaussian Markov Random Fields*). Este tipo de modelo é importante porque, entre outras razões, ele é o modelo favorito para representar efeitos aleatórios com estrutura de correlação induzida por grafos. Em particular, seu uso é fundamental em modelos bayesianos. Como a sua matriz de precisão (inversa da matriz de covariâncias) é uma matriz esparsa (ou seja, cheia de zeros), métodos de simulação Monte Carlo via cadeias de Markov são implementados mais facilmente.

Considere um grafo G com um conjunto de n vértices V e um conjunto de arestas E . Esse grafo será denotado por $G = (V, E)$. Seja \mathbf{Q} uma matriz $n \times n$ simétrica e definida positiva tal que o elemento Q_{ij} é igual a zero se, e somente se, os vértices i e j não estiverem conectados por uma aresta. Pensando em grafos que representam mapas, cada área será conectada a poucas outras áreas vizinhas. Dessa forma, uma matriz \mathbf{Q} associada a um mapa desses deverá ter a maioria de seus elementos iguais a zero. Por exemplo, numa grade regular com n vértices, cada um com 4 vizinhos (ignorando as bordas), teremos uma matriz

de dimensão $n \times n$ com uma fração de elementos nulos aproximadamente igual a $(n^2 - 4n)/n^2 = 1 - 4/n$. Se n for grande, isto implica que a quase totalidade da matriz será composta de zeros. Ela será uma matriz esparsa.

Definição 4.1.1. Um vetor aleatório $\mathbf{X} \in R^n$ é denominado GMRF com respeito a um grafo $G = (V, E)$ com média μ e matriz de precisão $\mathbf{Q} > 0$ (definida positiva), se e somente se, a distribuição conjunta do vetor é dada por

$$f_{\mathbf{X}}(\mathbf{x}) = (2\pi)^{-\frac{n}{2}} |\mathbf{Q}|^{\frac{1}{2}} \exp\left(-\frac{1}{2}(\mathbf{x} - \mu)^T \mathbf{Q}(\mathbf{x} - \mu)\right) \quad (4.1.1)$$

onde

$$Q_{ij} \neq 0 \Leftrightarrow i, j \in E \text{ para todo } i \neq j. \quad (4.1.2)$$

A definição acima define um campo de Markov gaussiano sem fazer referência à propriedade fundamental (3.3.3) dos campos de Markov em geral. Isto quer dizer que a propriedade (3.3.3) não é válida? Se ela não for válida, então (4.1.1) não é um campo gaussiano como definido por (3.3.3). Na verdade, veremos na seção 4.4 que a propriedade markoviana fundamental (3.3.3) é válida.

Entretanto, a definição (4.1.1) tem uma consequência estranha à primeira vista. Ela implica que *qualquer* distribuição normal multivariada com matriz de covariância simétrica definida positiva é um campo de Markov gaussiano com respeito a *algum* grafo de vizinhança. Isto é verdade porque, se $\Sigma > 0$ é a matriz de covariância de uma distribuição normal multivariada qualquer com média μ , sabemos que sua densidade conjunta é dada por

$$f(\mathbf{x}) = (2\pi)^{-n/2} |\Sigma|^{-1/2} \exp\left(-\frac{1}{2}(\mathbf{x} - \mu)^T \Sigma^{-1}(\mathbf{x} - \mu)\right).$$

Para obter 4.1.1 basta tomar $\mathbf{Q} = \Sigma^{-1}$. Assim, qualquer normal multivariada com matriz de covariância $\Sigma > 0$ definida positiva é um Campo de Markov com respeito ao grafo de vizinhança induzido por $\mathbf{Q} = \Sigma^{-1}$. Isto é, se $[\Sigma^{-1}]_{ij} > 0$, então trace uma aresta conectando i e j . Se $[\Sigma^{-1}]_{ij} = 0$, então não conecte os dois vértices.

Se todo campo de Markov gaussiano é uma normal multivariada e se toda normal multivariada pode ser vista como um campo de Markov gaussiano com respeito a algum grafo de vizinhança, para que criar a definição de um campo

de Markov gaussiano? Afinal os dois conjuntos de distribuições são os mesmos e assim temos dois nomes para a mesma coisa.

A vantagem de se definir um modelo como um GMFR está no fato de sua matriz de precisão ser esparsa. Isto é, na prática só usamos o nome GMRF para designar uma distribuição normal multivariada se a matriz de precisão $\mathbf{Q} = \boldsymbol{\Sigma}^{-1}$ for esparsa. Geralmente, em modelos usuais de análise de dados, em que temos várias medições em um indivíduo, tanto a matriz de covariância $\boldsymbol{\Sigma}$ quanto sua inversa $\mathbf{Q} = \boldsymbol{\Sigma}^{-1}$ terão muito poucos ou nenhuma entrada igual a zero. Um GMRF em que $\mathbf{Q} = \boldsymbol{\Sigma}^{-1}$ tem todos os seus elementos diferentes de zero induz um grafo em todos os possíveis pares de vértices estão conectados por arestas. Isto é, todo vértice é vizinho de todo outro vértice. Estamos interessados em modelos com uma estrutura de vizinhança mais restrita, com cada vértice conectando-se a alguns poucos outros vértices.

Dessa forma, os modelos GMRF interessantes são aqueles em que $\mathbf{Q} = \boldsymbol{\Sigma}^{-1}$ tem muitos zeros. No próximo capítulo vamos ver vários exemplos de modelos GMRF usados na modelagem de dados. Neste capítulo, vamos explorar as propriedades decorrentes da matriz de precisão \mathbf{Q} ser esparsa. Entre estas propriedades, vamos verificar que as relações de independência condicional estão relacionadas com o grafo de vizinhança induzido por \mathbf{Q} . Isto é, os zeros de \mathbf{Q} vão determinar as relações de independência condicional entre as variáveis.

Antes de prosseguirmos, vamos apresentar uma maneira simples de se encontrar distribuições condicionais a partir da conjunta, visto que essa será uma ferramenta usada recorrentemente daqui em diante. Vamos também simplificar a notação das densidades omitindo o subscrito que identifica as variáveis aleatórias. Assim, de agora em diante vamos escrever $f(x, y)$ para a densidade conjunta $f_{XY}(x, y)$ e $f(y|x)$ para a densidade condicional $f_{Y|X}(y|x)$. Da mesma forma, vamos escrever simplesmente $f(\mathbf{x})$ para a densidade conjunta $f_{\mathbf{X}}(\mathbf{x})$ do vetor \mathbf{X} e $f(\mathbf{x})$ para a densidade conjunta $f(x_1|x_2, \dots, x_n)$ para a densidade condicional ao invés de $f_{X_1|X_2, \dots, X_n}(x_1|x_2, \dots, x_n)$. Deve ficar claro a partir dos argumentos da função quais as variáveis envolvidas na densidade em questão. Quando este não for o caso, nós vamos retornar para a notação um pouco mais explícita que usamos até agora.

Seja \mathbf{x}_A o conjunto de observações do vetor aleatório \mathbf{X} restritas ao subconjunto de vértices $A \subset G$. Assim, se A possui k vértices i_1, \dots, i_k , o vetor \mathbf{x}_A possui dimensão k e tem os elementos x_{i_j} em suas entradas: $\mathbf{x}_A =$

$(x_{i_1}, \dots, x_{i_k})$. Vamos denotar por \mathbf{x}_{-A} o vetor formado pelas demais observações do grafo. Sabemos então que

$$f(\mathbf{x}_A | \mathbf{x}_{-A}) = \frac{f(\mathbf{x}_A, \mathbf{x}_{-A})}{f(\mathbf{x}_{-A})} \propto f(\mathbf{x}) \quad (4.1.3)$$

visto que $f(\mathbf{x}_{-A})$ é uma constante em relação \mathbf{x}_A .

4.2 Independência Condicional

Definição 4.2.1. *Dizemos que duas variáveis X e Y são condicionalmente independentes dado Z (e denotamos isto por $X \perp Y | Z$), se, e somente se,*

$$f(x, y | z) = f(x | z)f(y | z).$$

Em outras palavras, dizemos que X e Y são condicionalmente independentes dado Z , se conhecido o valor de Z , o conhecimento de Y não nos diz nada em relação a X . Então $f(x|y, z) = f(x|z)$. Sob essa condição, a distribuição conjunta é dada por

$$f(x, y, z) = f(x, y | z)f(z) = f(x | z)f(y | z)f(z).$$

Vamos mostrar um resultado muito importante sobre independência condicional. Ele será usado diversas vezes para verificar essa propriedade de independência condicional.

Teorema 4.2.1. *Duas variáveis X e Y são condicionalmente independentes dado Z se, e somente se*

$$f(x, y, z) = h(x, z)g(y, z)$$

para alguma função h e g e para todo Z com $f(z) > 0$.

Demonastração. Vamos provar primeiro a ida, ou seja, se $X \perp Y | Z$, então a conjunta dessas três variáveis aleatórias pode ser fatorada em duas partes, uma que depende só de x e z outra que depende só de y e z . Sabemos que

$$f(x, y, z) = f(x, y | z)f(z) = f(x | z)f(y | z)f(z)$$

sendo que para a segunda igualdade estamos usando a hipótese de X e Y são condicionalmente independentes dado Z . Fazendo, então,

$$h(x, z) = f(x|z) \quad \text{e} \quad g(y, z) = f(y|z)f(z)$$

temos que

$$f(x, y, z) = h(x, z)g(y, z)$$

como queríamos demonstrar.

Vamos mostrar a volta do teorema. Nossa hipótese agora é de que

$$f(x, y, z) = h(x, z)g(y, z).$$

Usando isso, vamos encontrar $f(z)$, $f(x|z)$ e $f(y|z)$. Sabemos que

$$f(z) = \int \int f(x, y, z) dx dy = \int \int h(x, z)g(y, z) dx dy = \int h(x, z) dx \int g(y, z) dy \quad (4.2.4)$$

e que

$$f(x|z) = \frac{f(x, z)}{f(z)}.$$

Usando a igualdade (4.2.4) e a hipótese de que a conjunta pode ser fatorada, temos que

$$f(x|z) = \frac{\int h(x, z)g(y, z) dy}{\int h(x, z) dx \int g(y, z) dy} = \frac{h(x, z)}{\int h(x, z) dx}.$$

Procedendo de maneira análoga, obtemos também

$$f(y|z) = \frac{g(y, z)}{\int g(y, z) dy}.$$

Dessa maneira, utilizando os resultados apresentados acima, percebemos que a conjunta de X e Y dado Z pode ser escrita como

$$f(x, y|z) = \frac{f(x, y, z)}{f(z)} = \frac{h(x, z)g(y, z)}{\int h(x, z) dx \int g(y, z) dy} = \underbrace{\left(\frac{h(x, z)}{\int h(x, z) dx} \right)}_{f(x|z)} \underbrace{\left(\frac{g(y, z)}{\int g(y, z) dy} \right)}_{f(y|z)}.$$

Portanto, de acordo com a definição (4.2.1), X e Y são condicionalmente independentes dado Z . E, portanto, está demonstrado o teorema. ♠ □

Exercício 4.2.1. Para cada um dos casos abaixo verifique se X e Y dado Z são condicionalmente independentes.

- $f(x, y, z) \propto \exp\{xyz\}$
- $f(x, y, z) \propto \exp\{-(x+z)^2 - (y+z)^2\}$

Resposta: O primeiro conjunto de variáveis são condicionalmente dependentes e o segundo são condicionalmente independentes. ♠

Vamos agora apresentar um resultado importante relacionado a esse conceito de independência condicional para o caso de variáveis aleatórias normais. Ele garante que a matriz de precisão $\mathbf{Q} = \boldsymbol{\Sigma}^{-1}$ de campos de Markov gaussianos está cheia de zeros.

Teorema 4.2.2. Seja \mathbf{X} um vetor de variáveis aleatórias normais com média μ e matriz de covariância \mathbf{Q}^{-1} ou matriz de precisão \mathbf{Q} . Temos então que para $i \neq j$

$$X_i \perp X_j | X_{-ij} \Leftrightarrow Q_{ij} = 0$$

Demonstração. Vamos dividir o vetor \mathbf{x} em três partes (x_i, x_j, x_{-ij}) e usar o critério da fatoração apresentado em 4.2.1. Sem perda de generalidade, podemos supor que $\mu = 0$. Dessa maneira, de 4.1.3 temos que

$$\begin{aligned} f(x_i | x_j, x_{-ij}) &\propto f(x_i, x_j, x_{-ij}) \propto \exp\left(-\frac{1}{2}\mathbf{x}^T \mathbf{Q} \mathbf{x}\right) = \exp\left(-\frac{1}{2} \sum_k \sum_l x_k Q_{kl} x_l\right) \\ &= \exp\left(\underbrace{-\frac{1}{2} x_i x_j (Q_{ij} + Q_{ji})}_{A} - \underbrace{\frac{1}{2} \sum_{k:k \neq j} \sum_{l:l \neq i} x_k Q_{kl} x_l}_{B}\right) \end{aligned}$$

percebe-se, portanto, que o termo B não apresenta o produto $x_i x_j$ e o termo A possui esse produto se, e somente se, $Q_{ij} = Q_{ji} \neq 0$. Isso significa que a conjunta $f(x_i, x_j, x_{-ij})$ pode ser fatorada como um produto de funções do tipo $f(x_i, x_{-ij})$ e $g(x_j, x_{-ij})$ se, e somente se, $Q_{ij} = 0$. E, portanto, está demonstrado o teorema. ♠ □

Dessa maneira, podemos definir, a partir do padrão de zeros da matriz de precisão, quais sítios são ou não condicionalmente independentes. A situação inversa também é possível: num GMRF com grafo de vizinhança $G = (V, E)$ sabemos que $Q_{ij} = 0$ se não existir uma aresta conectando os vértices i e j . Isso é importante quando estamos trabalhando com um modelo que tenha como objetivo captar dependência espacial entre observações.

Exercício 4.2.2. Dentre as matrizes apresentadas abaixo, diga qual delas é a matriz de precisão do grafo que se encontra na Figura 4.1.

$$\begin{array}{ll}
 a) \begin{bmatrix} 1 & 0.5 & 0 & 0 & 5 \\ 2 & 1 & 3 & 0 & 0 \\ 0 & 3 & 0.3 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 \\ -0.6 & 0 & 0.1 & 1 & 1 \end{bmatrix} & b) \begin{bmatrix} 2 & -0.8 & 0 & 0 & -0.8 \\ -0.8 & 2 & -0.8 & 0 & 0 \\ 0 & -0.8 & 3 & -0.8 & -0.8 \\ 0 & 0 & -0.8 & 2 & -0.8 \\ -0.8 & 0 & -0.8 & -0.8 & 3 \end{bmatrix} \\
 c) \begin{bmatrix} 1 & -0.2 & 0 & 0 & -1 \\ -1 & -2 & 0.9 & 0 & 0 \\ 0 & -1 & 1 & 1 & -1 \\ 0 & 0 & -1 & 1 & -1 \\ -0.3 & 0 & -1 & -0.5 & -3 \end{bmatrix} & d) \begin{bmatrix} 1 & -1 & 0 & -1 & -1 \\ -1 & -2 & 1 & 0 & -1 \\ 0 & -1 & 1 & 1 & -1 \\ -1 & 0 & -1 & 1 & -1 \\ -1 & -1 & -1 & -1 & -3 \end{bmatrix}
 \end{array}$$

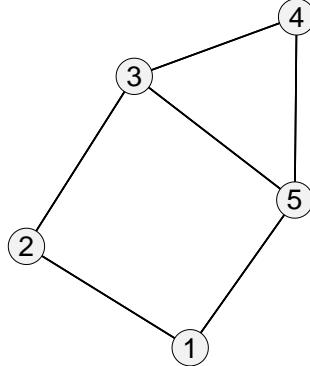


Figura 4.1: Grafo de vizinhança: como exercício, identifique a matriz de precisão \mathbf{Q} associada a este grafo.

Resposta: matriz B. ♠

Exercício 4.2.3. Dentre os gráficos apresentados na Figura 4.2 diga qual deles corresponde a seguinte matriz de precisão.

$$\begin{bmatrix} 5 & -0.5 & -0.5 & -0.5 & -0.5 & -0.5 \\ -0.5 & 3 & -0.5 & 0 & 0 & -0.5 \\ -0.5 & -0.5 & 3 & -0.5 & 0 & 0 \\ -0.5 & 0 & -0.5 & 3 & -0.5 & 0 \\ -0.5 & 0 & 0 & -0.5 & 3 & -0.5 \\ -0.5 & -0.5 & 0 & 0 & -0.5 & 3 \end{bmatrix}$$

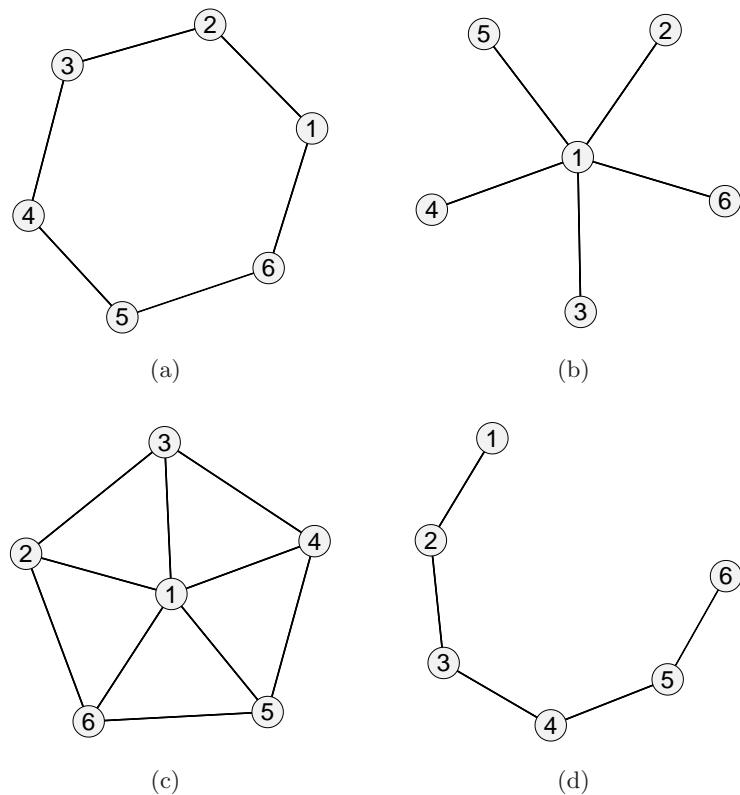


Figura 4.2: Possíveis grafos correspondentes à matriz de precisão \mathbf{Q} .

Resposta: grafo C. ♠

4.3 Condicionais obtidas a partir da conjunta

Quando estamos sob a suposição de normalidade, temos ainda alguns resultados que permitem obter facilmente os parâmetros das distribuições condicionais a partir dos termos da matriz de precisão. Os mesmos são apresentados no teorema a seguir.

Teorema 4.3.1. *Seja \mathbf{X} um GMRF com respeito a um grafo $G = (V, E)$ com média μ e matriz de precisão $\mathbf{Q} > 0$ então*

$$E(X_i|\mathbf{X}_{-i}) = \mu_i - \frac{1}{Q_{ii}} \sum_{j \in \mathcal{N}_i} Q_{ij}(x_j - \mu_j)$$

$$Prec(X_i|\mathbf{X}_{-i}) = Q_{ii}$$

$$Corr(X_i, X_j|\mathbf{X}_{-ji}) = -\frac{Q_{ij}}{\sqrt{Q_{ii}Q_{jj}}} \text{ para } i \neq j$$

lembrando que \mathcal{N}_i denota o conjunto de vizinhos de i .

Demonstração. Para uma única variável aleatória, se $\mathbf{X} \sim N(\gamma, k^{-1})$ (ou seja, uma normal com média γ e precisão k) então

$$\begin{aligned} f(x) &\propto \exp\left(-\frac{k}{2}(x - \gamma)^2\right) = \exp\left(-\frac{kx^2}{2} + kx\gamma - \frac{\gamma^2k}{2}\right) \\ &\propto \exp\left(-\frac{kx^2}{2} + kx\gamma\right). \end{aligned} \quad (4.3.5)$$

Voltando ao caso multivariado mas considerando o vetor de médias $\mu = 0$, temos que

$$f(x_i|\mathbf{x}_{-i}) \propto \exp\left(-\frac{1}{2}\mathbf{x}^T \mathbf{Q} \mathbf{x}\right) = \exp\left(-\frac{1}{2} \sum_{l=1}^n \sum_{k=1}^n x_k x_l Q_{kl}\right).$$

Descartando os termos que não dependem de x_i chegamos a

$$f(x_i|\mathbf{x}_{-i}) \propto \exp\left(-\frac{1}{2}x_i^2 Q_{ii} - x_i \sum_{j \in \eta_i} Q_{ij} x_j\right). \quad (4.3.6)$$

Comparando (4.3.5) e (4.3.6) percebemos que

$$Q_{ii} = k \quad \text{e} \quad k\gamma = - \sum_{j \in \eta_i} Q_{ij} x_j$$

o que implica que

$$\gamma = -\frac{1}{Q_{ii}} \sum_{j \in \mathcal{N}_i} Q_{ij} x_j ,$$

ou seja,

$$E(X_i | \mathbf{x}_{-i}) = -\frac{1}{Q_{ii}} \sum_{j \in \mathcal{N}_i} Q_{ij} x_j \quad \text{e} \quad \text{Prec}(X_i | x_{-i}) = Q_{ii} .$$

Podemos estender facilmente esse resultado para o caso em que \mathbf{X} tem média μ diferente de zero. Basta notar que se \mathbf{X} tem média μ , então $\mathbf{X} - \mu$ tem média zero. Usando os resultados que acabamos de provar, temos então

$$E(X_i - \mu_i | \mathbf{x}_{-i}) = -\frac{1}{Q_{ii}} \sum_{j \in \mathcal{N}_i} Q_{ij}(x_j - \mu_j) \rightarrow E(X_i | \mathbf{x}_{-i}) = \mu_i - \frac{1}{Q_{ii}} \sum_{j \in \mathcal{N}_i} Q_{ij}(x_j - \mu_j)$$

e

$$\text{Prec}(x_i - \mu_i | \mathbf{x}_{-i}) = \text{Prec}(x_i | \mathbf{x}_{-i}) = Q_{ii} .$$

Vamos olhar agora para a correlação condicional. Sabemos que se o vetor X_i, X_j tem distribuição normal bivariada com matriz de variância Σ e média zero, então sua densidade conjunta pode ser escrita como

$$f(x_i, x_j) \propto \exp \left\{ -\frac{1}{2} \begin{bmatrix} x_i & x_j \end{bmatrix} \begin{bmatrix} \Sigma_{ii} & \Sigma_{ij} \\ \Sigma_{ji} & \Sigma_{jj} \end{bmatrix}^{-1} \begin{bmatrix} x_i \\ x_j \end{bmatrix} \right\} . \quad (4.3.7)$$

Vamos mostrar que $f(x_i, x_j | \mathbf{x}_{-ij})$ é escrita desta forma e portanto é uma normal bivariada e com parâmetros facilmente identificáveis. Temos

$$f(x_i, x_j | \mathbf{x}_{-ij}) \propto \exp \left\{ -\frac{1}{2} \sum_{k=1}^n \sum_{l=1}^n x_k x_l Q_{kl} \right\}$$

e, descartando os termos que não dependem de x_i nem de x_j , temos

$$\begin{aligned} f(x_i, x_j | \mathbf{x}_{-ij}) &\propto \exp \left\{ -\frac{1}{2} (x_i^2 Q_{ii} + x_i x_j Q_{ij} + x_i x_j Q_{ji} + x_j^2 Q_{jj}) \right\} = \\ &\exp \left\{ -\frac{1}{2} \begin{bmatrix} x_i & x_j \end{bmatrix} \begin{bmatrix} Q_{ii} & Q_{ij} \\ Q_{ji} & Q_{jj} \end{bmatrix} \begin{bmatrix} x_i \\ x_j \end{bmatrix} \right\}. \end{aligned} \quad (4.3.8)$$

Comparando (4.3.7) e (4.3.8) percebemos que

$$\begin{bmatrix} Q_{ii} & Q_{ij} \\ Q_{ji} & Q_{jj} \end{bmatrix} = \begin{bmatrix} \Sigma_{ii} & \Sigma_{ij} \\ \Sigma_{ji} & \Sigma_{jj} \end{bmatrix}^{-1} = \begin{bmatrix} \Sigma_{jj}/\Delta & -\Sigma_{ij}/\Delta \\ -\Sigma_{ji}/\Delta & \Sigma_{ii}/\Delta \end{bmatrix},$$

onde

$$\Delta = \det \begin{bmatrix} \Sigma_{ii} & \Sigma_{ij} \\ \Sigma_{ji} & \Sigma_{jj} \end{bmatrix}.$$

Isso implica que

$$Q_{ii} = \frac{\Sigma_{jj}}{\Delta}, \quad Q_{jj} = \frac{\Sigma_{ii}}{\Delta}, \quad Q_{ij} = -\frac{\Sigma_{ij}}{\Delta} \quad \text{e} \quad Q_{ji} = -\frac{\Sigma_{ji}}{\Delta},$$

ou seja

$$\text{Corr}(x_i, x_j | \mathbf{x}_{-ij}) = -\frac{Q_{ij}}{\sqrt{Q_{ii}Q_{jj}}}.$$

Portanto, está demonstrado o Teorema 4.3.1. ♠

□

Observe que $\text{Cov}(x_i, x_j | \mathbf{x}_{-ij}) = 0$ se, e somente se, $Q_{ii} = 0$. Portanto, a matriz \mathbf{Q} indica quais os pares de vértices são condicionalmente independentes. Dado um grafo $G = (V, E)$, podemos de imediato dizer quais os elementos de $\mathbf{Q} = \boldsymbol{\Sigma}^{-1}$ que devem ser nulos: são aqueles correspondentes aos pares de vértices i e j que não são conectados por uma aresta no grafo de vizinhança.

Sabemos que a matriz de covariância $\boldsymbol{\Sigma}$ fornece a informação para calcular a correlação *marginal* entre i e j pois

$$\text{Corr}(X_i, X_j) = \frac{\Sigma_{ij}}{\sqrt{\Sigma_{ii}\Sigma_{jj}}}.$$

Já a matriz \mathbf{Q} está associada com a correlação entre i e j condicionalmente nos valores das demais variáveis:

$$\text{Corr}(x_i, x_j | \mathbf{x}_{-ij}) = -\frac{Q_{ij}}{\sqrt{Q_{ii}Q_{jj}}}.$$

Existe uma certa simetria nas duas fórmulas, a da correlação marginal e a da correlação condicional. Elas são quase idênticas: existe um sinal de menos que não existe na correlação marginal.

Além disso, a diagonal da matriz \mathbf{Q} fornece o inverso da variância de X_i condicionado nos valores da demais variáveis.

4.4 Conjunta obtida a partir das condicionais

Na seção anterior mostramos como chegamos às condicionais a partir da distribuição conjunta. Vamos fazer agora o oposto: definir as distribuições condicionais e a partir delas chegaremos à distribuição conjunta. Na verdade, este é o principal interesse na modelagem de dados espaciais. As distribuições condicionais vão determinar o grafo de vizinhança e portanto vão também determinar quais elementos da matriz de precisão $\mathbf{Q} = \boldsymbol{\Sigma}^{-1}$ são nulos e quais são não-nulos.

A dificuldade é saber quando um conjunto de densidades condicionais realmente determina a distribuição conjunta. A resposta a esta pergunta está no Teorema de Hammersley-Clifford, a ser visto no Capítulo 6. No capítulo atual, nós vamos adotar um conjunto de densidade condicionais que são compatíveis entre si sem nos preocupar em demonstrar este fato. Queremos apenas apresentar um exemplo de densidades condicionais normais que determinam uma conjunta que é uma normal multivariada.

Para obter a conjunta a partir das densidades condicionais, vamos utilizar um resultado que foi apresentado no Capítulo 2 para o caso bivariado, a expansão de Brook (ver (2.3.2) na seção 2.3). Esse resultado será utilizado aqui no caso multivariado geral e sua demonstração será feita no Capítulo 6, na seção 6.1.

A expansão de Brook diz que se temos um conjunto de condicionais compatíveis e se a densidade conjunta satisfaz a condição de positividade (isto é, se $f(x_i) > 0$ para cada i então $f(x_1, x_2, \dots, x_n) > 0$), então

$$\frac{f(\mathbf{x})}{f(\mathbf{y})} = \prod_{i=1}^n \frac{f(x_i|x_1, \dots, x_{i-1}, y_{i+1}, \dots, y_n)}{f(y_i|x_1, \dots, x_{i-1}, y_{i+1}, \dots, y_n)} = \prod_{i=1}^n \frac{f(x_i|y_1, \dots, y_{i-1}, x_{i+1}, \dots, x_n)}{f(y_i|y_1, \dots, y_{i-1}, x_{i+1}, \dots, x_n)} \quad (4.4.9)$$

onde \mathbf{y} é uma configuração de referência.

Considerando a primeira igualdade, temos

$$\begin{aligned} \frac{f(\mathbf{x})}{f(\mathbf{y})} &= \frac{f(x_1|y_2, \dots, y_n)}{f(y_1|y_2, \dots, y_n)} \frac{f(x_2|x_1, y_3, \dots, y_n)}{f(y_2|x_1, y_3, \dots, y_n)} \\ &\quad \frac{f(x_3|x_1, x_2, y_4, \dots, y_n)}{f(y_3|x_1, x_2, y_4, \dots, y_n)} \cdots \frac{f(x_n|x_1, \dots, x_{n-1})}{f(y_n|x_1, \dots, x_{n-1})} \end{aligned}$$

e, considerando a segunda igualdade, temos

$$\begin{aligned} \frac{f(\mathbf{x})}{f(\mathbf{y})} &= \frac{f(x_1|x_2, \dots, x_n)}{f(y_1|x_2, \dots, x_n)} \frac{f(x_2|y_1, x_3, \dots, x_n)}{f(y_2|y_1, x_3, \dots, x_n)} \\ &\quad \frac{f(x_3|y_1, y_2, x_4, \dots, x_n)}{f(y_3|y_1, y_2, x_4, \dots, x_n)} \cdots \frac{f(x_n|y_1, \dots, y_{n-1})}{f(y_n|y_1, \dots, y_{n-1})}. \end{aligned}$$

Considere um grafo $G = (V, E)$ determinando uma estrutura de vizinhança entre os vértices. Seja β_{ij} pesos tais que $\beta_{ij} = 0$ se não existir uma aresta conectando i e j . Assim, os pesos β_{ij} refletem a estrutura de vizinhança do grafo G . Baseado neste grafo e nos pesos β_{ij} , vamos usar as seguintes distribuições condicionais:

$$X_i|\mathbf{X}_{-i} = \mathbf{x}_{-i} \sim N\left(\mu_i - \sum_{j:j \neq i} \beta_{ij}(x_j - \mu_j), k_i\right) \quad (4.4.10)$$

com $k_i > 0$.

O objetivo de definir o modelo (4.4.10) é colocar a maioria dos β_{ij} 's iguais a zero. Apenas áreas vizinhas à a área i teriam $\beta_{ij} \neq 0$. Dessa forma, a distribuição condicional (4.4.10) de um sítio i dado o resto do mapa depende apenas dos valores x_j observados em seus vizinhos estabelecidos pelo grafo $G = (V, E)$. Isto acontece porque

$$\sum_{j:j \neq i} \beta_{ij}(x_j - \mu_j) = \sum_{j:j \sim i} \beta_{ij}(x_j - \mu_j).$$

Esta igualdade implica na desejada propriedade markoviana com respeito ao grafo G .

A média condicional em (4.4.10) é um modelo de regressão linear nos desvios $x_j - \mu_j$ das observações x_j das áreas vizinhas em relação às suas médias μ_j . Se os β_{ij} forem positivos, estaremos dizendo que se os vizinhos forem maiores que suas médias. Isto é, se $x_j - \mu_j > 0$ então podemos antecipar que o evento $X_i > \mu_i$ tem uma chance maior de ocorrer que o evento $X_i < \mu_i$. Assim, se os $\beta_{ij} > 0$ e se as médias forem constantes ($\mu_i = \mu$ para todo i), devemos esperar áreas vizinhas com valores similares, clusters espaciais de valores altos entremeados com clusters de valores baixos.

Este modelo linear nos vizinhos permite a compatibilidade desse conjunto de condicionais com uma conjunta, como será demonstrado pelo teorema de Hammersley e Clifford no Capítulo 6. Utilizando a igualdade (4.4.9) vamos mostrar que a densidade conjunta é a de uma normal multivariada com matriz de precisão \mathbf{Q} dada por

$$Q_{ij} = \begin{cases} k_i \beta_{ij} & \text{se } i \neq j \\ k_i & \text{se } i = j \end{cases}$$

sujeito à restrição $k_i \beta_{ij} = k_j \beta_{ji}$.

Vamos escolher aqui o vetor n -dimensional de zeros $\mathbf{0} = (0_1, \dots, 0_n)$ como nosso estado de referência \mathbf{y} . Podemos supor ainda, sem perda de generalidade, que o vetor de médias μ também é o vetor zero. Dessa maneira, usando a primeira forma de escrever a razão entre as probabilidades como em (4.4.9) chegamos a

$$\begin{aligned} \frac{f(\mathbf{x})}{f(\mathbf{0})} &= \prod_{i=1}^n \frac{f(x_i | x_1, \dots, x_{i-1}, 0_{i+1}, \dots, 0_n)}{f(0_i | x_1, \dots, x_{i-1}, 0_{i+1}, \dots, 0_n)} \\ &= \frac{\prod_{i=1}^n \exp\left(-\frac{k_i}{2}(x_i + \sum_{j=1}^{i-1} \beta_{ij} x_j)^2\right)}{\prod_{i=1}^n \exp\left(-\frac{k_i}{2}(\sum_{j=1}^{i-1} \beta_{ij} x_j)^2\right)} \\ &= \frac{\prod_{i=1}^n \exp\left\{-k_i x_i^2/2 - x_i k_i \sum_{j=1}^{i-1} \beta_{ij} x_j - k_i/2 (\sum_{j=1}^{i-1} \beta_{ij})^2\right\}}{\prod_{i=1}^n \exp\left\{-k_i/2 (\sum_{j=1}^{i-1} \beta_{ij} x_j)^2\right\}}. \end{aligned}$$

Tirando o logaritmo dos dois lados da igualdade e cancelando alguns termos da divisão temos que

$$\log \left(\frac{f(\mathbf{x})}{f(\mathbf{0})} \right) = -\frac{1}{2} \sum_{i=1}^n k_i x_i^2 - \frac{1}{2} \sum_{i=2}^n \sum_{j=1}^{i-1} k_i \beta_{ij} x_i x_j . \quad (4.4.11)$$

Note que o segundo somatório começa em $i = 2$.

Usando agora a segunda igualdade, temos

$$\frac{f(\mathbf{x})}{f(\mathbf{0})} = \prod_{i=1}^n \frac{f(x_i | 0_1, \dots, 0_{i-1}, x_{i+1}, \dots, x_n)}{f(0_i | 0_1, \dots, 0_{i-1}, x_{i+1}, \dots, x_n)} .$$

Portanto,

$$\begin{aligned} \frac{f(\mathbf{x})}{f(\mathbf{0})} &= \frac{\prod_{i=1}^n e^{-\frac{k_i}{2}(x_i + \sum_{j=i+1}^n \beta_{ij} x_j)^2}}{\prod_{i=1}^n e^{-\frac{k_i}{2}(\sum_{j=i+1}^n \beta_{ij} x_j)^2}} \\ &= \frac{\prod_{i=1}^n \exp \left\{ -\frac{k_i x_i^2}{2} - x_i k_i \sum_{j=i+1}^n \beta_{ij} x_j - \frac{k_i}{2} (\sum_{j=i+1}^n \beta_{ij})^2 \right\}}{\prod_{i=1}^n \exp \left\{ -\frac{k_i}{2} (\sum_{j=i+1}^n \beta_{ij} x_j)^2 \right\}} . \end{aligned}$$

Tirando novamente o logaritmo e cancelando os termos da divisão temos

$$\log \left(\frac{f(\mathbf{x})}{f(\mathbf{0})} \right) = -\frac{1}{2} \sum_{i=1}^n k_i x_i^2 - \sum_{i=1}^{n-1} \sum_{j=i+1}^n k_i \beta_{ij} x_i x_j . \quad (4.4.12)$$

Nesse caso, o segundo somatório vai apenas até $i = n - 1$.

Como as duas formas de obter a conjunta devem levar ao mesmo resultado, os termos em (4.4.11) e (4.4.12) devem ser iguais. Isto implica dizer que

$$-\frac{1}{2} \sum_{i=1}^n k_i x_i^2 - \frac{1}{2} \sum_{i=2}^n \sum_{j=1}^{i-1} k_i \beta_{ij} x_i x_j = -\frac{1}{2} \sum_{i=1}^n k_i x_i^2 - \sum_{i=1}^{n-1} \sum_{j=i+1}^n k_i \beta_{ij} x_i x_j .$$

Cancelando o primeiro termo da igualdade e multiplicando por -1 encontramos

$$\sum_{i=1}^{n-1} \sum_{j=i+1}^n x_i x_j k_i \beta_{ij} = \sum_{i=2}^n \sum_{j=1}^{i-1} x_i x_j k_i \beta_{ij}$$

Vamos trocar os índices dos somatórios de forma que os dois lados da igualdade fiquem da mesma forma. Somar em $2 \leq i \leq n$ e $1 \leq j \leq i - 1$ é o mesmo que somar em $j + 1 \leq i \leq n$ e $1 \leq j \leq n - 1$. Portanto, a igualdade acima pode ser escrita como

$$\sum_{i=1}^{n-1} \sum_{j=i+1}^n x_i x_j k_i \beta_{ij} = \sum_{j=1}^{n-1} \sum_{i=j+1}^n x_i x_j k_i \beta_{ij} = \sum_{i=1}^{n-1} \sum_{j=i+1}^n x_i x_j k_j \beta_{ji} \dots$$

Como esta igualdade é válida para todo \mathbf{x} , isso implica que

$$k_i \beta_{ij} = k_j \beta_{ji},$$

o que mostra a necessidade da restrição acima estabelecida. Além disso,

$$\log(f(\mathbf{x})) = \text{const} - \frac{1}{2} \sum_{i=1}^n k_i x_i^2 - \frac{1}{2} \sum_{i \neq j} k_i x_i x_j \beta_{ij}$$

e portanto

$$f(\mathbf{x}) \propto \exp \left\{ -\frac{1}{2} \sum_{i=1}^n k_i x_i^2 - \frac{1}{2} \sum_{i \neq j} k_i x_i x_j \beta_{ij} \right\} = \exp \left\{ -\frac{1}{2} \mathbf{x}' \mathbf{Q} \mathbf{x} \right\}$$

onde

$$Q_{ij} = \begin{cases} k_i \beta_{ij} & \text{se } i \neq j \\ k_i & \text{se } i = j \end{cases}.$$

Ou seja, \mathbf{X} tem uma distribuição normal multivariada com vetor de médias $\mathbf{0}$ e matriz de precisão \mathbf{Q} definida acima. Com isto, fica demonstrado o resultado.



Capítulo 5

Exemplos de Campos de Markov Gaussianos

Vamos agora apresentar alguns exemplos de campos aleatórios de Markov gaussianos que são muito utilizados na análise de dados em estatística espacial.

5.1 Modelo CAR

O modelo CAR é um exemplo de um Campo de Markov que é definido a partir de distribuições condicionais como aquelas apresentadas na equação (4.4.10). O que se faz nesse caso é colocar $\beta_{ij} = \rho/n_i$, se $i \sim j$, e $\beta_{ij} = 0$, caso i e j não sejam vizinhos, onde ρ é uma parâmetro que mede a correlação espacial e n_i é o número de vizinhos do sítio i . Além disso, toma-se $k_i = n_i/\sigma^2$. Portanto as distribuições condicionais são definidas da seguinte forma

$$Y_i | \mathbf{Y}_{-i} = \mathbf{y}_{-i} \sim N\left(\rho\bar{y}_{-i}, \frac{\sigma^2}{n_i}\right) \quad (5.1.1)$$

onde \bar{y}_{-i} denota a média dos vizinhos do sítio i .

A média de um sítio, dado o resto do mapa, é proporcional à média de seus vizinhos. A constante de proporcionalidade é ρ . Se este parâmetro estiver entre -1 e 1 , temos um modelo semelhante ao modelo autoregressivo de séries temporais. Já falaremos sobre os valores possíveis para ρ . A precisão da distribuição condicional (5.1.1) é diretamente proporcional ao total de vizinhos

74 CAPÍTULO 5. EXEMPLOS DE CAMPOS DE MARKOV GAUSSIANOS

da área. Isto é intuitivo visto que, quanto mais vizinhos temos para uma determinada região, mais informação temos a respeito da mesma.

Utilizando os resultados apresentados anteriormente, é fácil perceber que a distribuição conjunta do vetor aleatório \mathbf{Y} será uma normal multivariada com média zero e com uma matriz de precisão \mathbf{Q} dada por

$$\mathbf{Q} = \frac{1}{\sigma^2} (\text{diag}(\mathbf{n}) - \rho \mathbf{A}) \quad \text{ou} \quad \sigma^2 Q_{ij} = \begin{cases} n_i & \text{se } i = j \\ -\rho & \text{se } i \sim j \\ 0 & \text{caso contrário} \end{cases}$$

onde \mathbf{A} é a matriz de adjacência. Ou seja, seu termos a_{ij} são iguais a 1 se i e j são vizinhos, e iguais a 0 caso contrário. O vetor \mathbf{n} tem na sua i -ésima entrada o número de vizinhos n_i do sítio i .

Outra forma de escrever a densidade conjunta de \mathbf{Y} envolve a matriz de pesos espaciais padronizada \mathbf{W} . Esta matriz é simplesmente a matriz de adjacência \mathbf{A} padronizada por linhas de forma que suas linhas somem 1. Isto é, $w_{ij} = a_{ij}/n_i$ onde $n_i = \sum_j a_{ij}$. A matriz \mathbf{W} é obtida por meio de uma simples multiplicação matricial:

$$\mathbf{W} = \text{diag}(\mathbf{n})^{-1} \mathbf{A}.$$

Esta matriz é importante e vamos descrevê-la um pouco mais por meio de um exemplo. Considere o mapa da Figura 5.1 com o grafo de vizinhança superposto. A matriz de adjacência \mathbf{A} é dada por

$$\mathbf{A} = \left(\begin{array}{ccccccc} & \text{L} & \text{BN} & \text{C} & \text{A} & \text{Cur} & \text{Col} \\ \text{Lapa} & 0 & 1 & 1 & 0 & 0 & 0 \\ \text{Balsa Nova} & 1 & 0 & 1 & 1 & 0 & 0 \\ \text{Contenda} & 1 & 1 & 0 & 1 & 0 & 0 \\ \text{Araucária} & 0 & 1 & 1 & 0 & 1 & 0 \\ \text{Curitiba} & 0 & 0 & 0 & 1 & 0 & 1 \\ \text{Colombo} & 0 & 0 & 0 & 0 & 1 & 0 \end{array} \right)$$

enquanto a matriz \mathbf{W} de pesos espaciais padronizados é dada por

$$\mathbf{W} = \begin{pmatrix} & L & BN & C & A & Cur & Col \\ Lapa & 0 & 1/2 & 1/2 & 0 & 0 & 0 \\ Balsa Nova & 1/3 & 0 & 1/3 & 1/3 & 0 & 0 \\ Contenda & 1/3 & 1/3 & 0 & 1/3 & 0 & 0 \\ Araucária & 0 & 1/3 & 1/3 & 0 & 1/3 & 0 \\ Curitiba & 0 & 0 & 0 & 1/2 & 0 & 1/2 \\ Colombo & 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix}$$

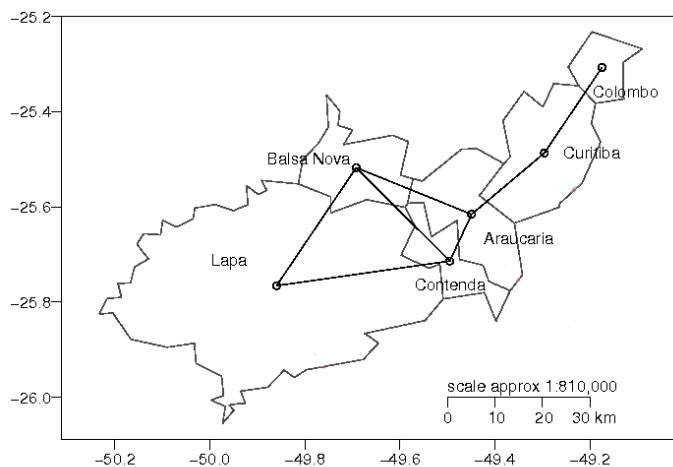


Figura 5.1: Mapa de municípios próximos a Curitiba, Paraná, com o grafo de vizinhança por adjacência superposto

Observe que a matriz \mathbf{W} é uma matriz estocástica, com seus elementos maiores ou iguais a zero e com suas linhas somando 1.

Como podemos escrever

$$\mathbf{Q} = \frac{1}{\sigma^2} \text{diag}(\mathbf{n}) (\mathbf{I} - \rho \mathbf{W}) ,$$

a distribuição conjunta do vetor \mathbf{Y} é dada por

$$\mathbf{Y} \sim N_n (\mathbf{0}, \sigma^2 (\mathbf{I} - \rho \mathbf{W})^{-1} \text{diag}(\mathbf{n})^{-1}) . \quad (5.1.2)$$

76 CAPÍTULO 5. EXEMPLOS DE CAMPOS DE MARKOV GAUSSIANOS

Para que possa existir a densidade conjunta do modelo CAR, é necessário que a matriz de covariância seja simétrica e definida positiva. Essa condição só é satisfeita se o parâmetro ρ estiver entre $1/\lambda_1$ e $1/\lambda_n$, onde λ_1 e λ_n são respectivamente o menor e o maior autovalor da matriz \mathbf{W} . Como esta é uma matriz estocástica, o seu maior autovalor é sempre 1 e portanto o intervalo em que ρ está definido deve ser do tipo $(a, 1)$ ou $(1, a)$. O segundo caso ocorrerá se o menor autovalor de \mathbf{W} for positivo e menor que 1. Entretanto, isto é impossível: como a matriz \mathbf{W} tem traço nulo (pois $w_{ii} = 0$) e 1 é um dos autovalores, pelo menos um de seus outros autovalores deve ser negativo. Isso significa que $\lambda_1 < 0$ e, portanto, o parâmetro $\rho \in (1/\lambda_1, 1) = (a, 1)$ onde $a < 0$. Note que, se $\lambda_1 \in (-1, 0)$ então ρ pode ser menor que -1.

Dessa forma, satisfeita tal condição, se um vetor de variáveis aleatórias $(Y_1, Y_2, \dots, Y_n) \sim CAR(\rho)$, a função de densidade conjunta dessas variáveis é proporcional a

$$f(y_1, y_2, \dots, y_n) \propto \exp\left(-\frac{1}{2\sigma^2} \mathbf{y}' \text{diag}(\mathbf{n}) [\mathbf{I} - \rho \mathbf{W}] \mathbf{y}\right).$$

Pelo Teorema 4.3.1 podemos obter a correlação entre dois sítios, dado o resto do mapa:

$$\text{Cor}(y_i, y_j | \mathbf{y}_{-ij})_{CAR} = \frac{\rho}{\sqrt{n_i n_j}} w_{ij}.$$

A correlação condicional entre duas áreas é proporcional ao parâmetro de correlação espacial ρ . Quanto maior o valor de ρ , mais correlacionadas serão as áreas. Além disso, ela é inversamente proporcional ao número de vizinhos de cada uma das áreas. Isso significa que duas áreas vizinhas, cada uma delas conectada com muitos vizinhos, terão uma correlação condicional bem menor que outras duas áreas vizinhas mas relativamente isoladas no mapa.

Esse modelo, apesar de amplamente utilizado, apresenta alguns problemas. Primeiramente, a partir da equação 5.1.1, nota-se que, mesmo sob a hipótese de independência espacial ($\rho = 0$), a variância condicional depende do número de vizinhos da área. Além disso, essa quantidade não depende do parâmetro de correlação espacial ρ , o que não faz muito sentido. Se estamos interessados em predizer o valor y_i dado o resto do mapa, a variância dessa estimativa deve depender não só do número de vizinhos da área, mas também do grau de correlação entre elas.

Além disso, como foi mencionado por Besag (1991), para se obter uma dependência espacial moderada entre as áreas, é necessário que o parâmetro de correlação espacial fique muito próximo de seu limite superior, que é igual a 1. Quando $\rho = 1$, a distribuição torna-se imprópria e caímos em um caso específico desse modelo que é denominado CAR intrínseco, denotado por ICAR, e que será apresentado na próxima seção.

5.2 Modelo ICAR

O modelo ICAR (*Intrinsic CAR*) consiste em um caso particular do modelo CAR, quando $\rho = 1$. Entretanto, este valor realmente não é possível pois, como vimos, $\rho \in (1/\lambda_1, 1)$ onde $1/\lambda_1 < 0$. Se $\rho = 1$, a matriz de precisão $\mathbf{Q} = 1/\sigma^2 \text{diag}(\mathbf{n}) (\mathbf{I} - \rho \mathbf{W})$ não é invertível e portanto não existe a matriz de covariância. Isto que dizer que não temos, de fato, uma distribuição de probabilidade. Entretanto, esse modelo ICAR, com $\rho = 1$, é utilizado como priori em modelos hierárquicos bayesianos. O fato do modelo ICAR não ser uma distribuição de probabilidade própria não é um problema desde que se garanta que a posteriori é própria e este é geralmente o caso nestas análises bayesianas.

Dessa maneira, a função de densidade (imprópria) conjunta do modelo ICAR assume o seguinte formato

$$f(y_1, y_2, \dots, y_n) \propto \exp\left(-\frac{1}{2\sigma^2} \mathbf{y}^T [\text{diag}(\mathbf{n}) - \mathbf{A}] \mathbf{y}\right) = \exp\left(-\frac{1}{2\sigma^2} \sum_{i \sim j} (y_i - y_j)^2\right).$$

Besag(1991) propôs uma abordagem segundo a qual o erro aleatório é dividido em duas partes: uma não estruturada espacialmente, à qual se atribui distribuições normais independentes, e uma parte espacialmente estruturada, à qual se atribui uma distribuição ICAR. Dessa maneira, a primeira componente representa características intrínsecas de cada área, ao passo que a segunda parte é formada por características que apresentam uma estrutura espacial.

Esse modelo proposto por Besag é um dos mais utilizados atualmente para mapeamento de doenças. Uma de suas dificuldades, porém, consiste em separar o efeito espacial e o não espacial. Uma maneira geralmente utilizada para se ter uma estimativa desse efeito consiste em dividir o desvio padrão de cada

uma das componentes pela soma dos desvios de ambas. Leroux (1999) apresentou uma alternativa a isso incluindo um parâmetro λ que consegue medir o efeito de cada uma dessas partes. Esse modelo será apresentado a seguir.

5.3 Modelo de Leroux

Como foi dito anteriormente, o modelo proposto por Leroux (1999) é bem similar aquele proposto por Besag (1991). A diferença entre os dois modelos está no fato de que para o modelo proposto por Besag existe apenas um parâmetro σ^2 (variância do erro estruturado) responsável por medir a dispersão e a correlação espacial. Por outro lado, no modelo de Leroux é incluído um termo λ , que irá medir a correlação espacial entre as áreas e um parâmetro σ^2 que deve mensurar a variância. O que ele propõe, então, é um Campo Markoviano Gaussiano com a seguinte matriz de precisão

$$\mathbf{Q} = (\sigma^2)^{-1} ((1 - \lambda)\mathbf{I} + \lambda\mathbf{R})$$

onde \mathbf{I} é a matriz identidade e \mathbf{R} é matriz de precisão do modelo ICAR. Isso significa que a matriz de precisão é uma soma ponderada dessas duas matrizes. Para esse modelo o parâmetro λ está definido no intervalo $(0, 1)$.

Utilizando os resultados apresentados no capítulo anterior, percebe-se que as distribuições condicionais são dadas por

$$Y_i | \mathbf{y}_{-i} \sim N \left(\frac{\lambda n_i}{1 - \lambda + \lambda n_i} \bar{y}_{-i}, \frac{\sigma^2}{1 - \lambda + \lambda n_i} \right).$$

Nota-se, portanto, que sob a hipótese de independência espacial, a variância condicional é igual a σ^2 , ou seja, não depende do número de vizinhos da área. Além disso, essa quantidade depende do parâmetro de correlação espacial λ , o que apresenta um apelo intuitivo, como já foi explicado anteriormente.

A correlação condicional nesse caso é dada por

$$\text{Corr}(y_i, y_j | \mathbf{y}_{-ij})_{\text{Leroux}} = \frac{\lambda}{\sqrt{1 - \lambda + \lambda n_i} \sqrt{1 - \lambda + \lambda n_j}} w_{ij}.$$

Assim como no modelo CAR, essa quantidade é proporcional ao parâmetro de correlação espacial e inversamente proporcional ao número de vizinhos da área. Dessa maneira será sempre menor ou igual a $1/\sqrt{n_i n_j}$.

Esse modelo apresenta algumas vantagens em relação aos anteriormente apresentados. Porém, assim como no modelo CAR, valores alto de correlação entre as áreas só são alcançados apenas se λ se aproxima muito de 1.

Capítulo 6

Teorema de Hammersley-Clifford

Apresentados alguns exemplos de campos de Markov, passaremos agora a resultados mais genéricos sobre esse tópico e a algumas demonstrações de resultados que já foram anunciados nesse texto. Os teoremas apresentados nesse capítulo são aplicáveis a várias áreas, não somente em estatística espacial.

6.1 Expansão de Brook

Brook (1964) apresentou um teorema que permite obter a distribuição conjunta $f(x_1, \dots, x_n)$ a partir das n distribuições condicionais completas $f(x_i|\mathbf{x}_{-i})$, uma para cada das variáveis x_i do vetor \mathbf{x} . O teorema assume que esta densidade conjunta realmente existe: as condicionais devem ser compatíveis entre si. Como determinar se um conjunto de n condicionais é compatível é um resultado muito mais difícil de estabelecer e é fundamentalmente o conteúdo teorema de Hammersley-Clifford, a ser visto mais a frente.

O teorema de Brook exige também a condição de positividade: se cada x_1, x_2, \dots, x_n pode ocorrer individualmente, então eles podem ocorrer conjuntamente. Formalmente isso significa que se $\mathbf{x} = (x_1, \dots, x_n)$ é um ponto tal que a densidade marginal $f(x_i)$ é maior que zero para cada variável i , então a densidade conjunta $f(x_1, x_2, \dots, x_n)$ no ponto \mathbf{x} também deve ser maior que zero.

John Hammersley nasceu em 1920 na Inglaterra e se formou em matemática em 1948 em Cambridge. Ele ficou muito conhecido devido ao seu trabalho na área de percolação, passeios auto-evitantes, processos estocásticos subaditivos e métodos Monte Carlo. Ele deu contribuições fundamentais em todos estes tópicos. Por pouco a matemática não perdeu um de seus talentos. Aos 10 anos, seu professor perguntou "how many blue beans made five". Quando ele não conseguiu responder à charada, o professor disse que a resposta era 5 e que ele era um tolo. Após um começo mediocre em Cambridge, ele se alistou no esforço de guerra inglês e aprendeu métodos numéricos e estatística e durante este tempo. Leu os livros de M.G. Kendall e Fisher que haviam naquela época. De acordo com ele, as técnicas estatísticas tiveram um importante papel em assegurar bom desempenho dos radares na guerra. Volta para Cambridge em 1946 e termina como Wrangler nos Tripos de matemática. Ele começou trabalhando em Oxford com Finney e lá permaneceu até aposentar-se. Ele tinha altas expectativas com seus alunos e orientou apenas 8 doutorandos durante toda sua carreira. Um desses estudantes, ao candidatar-se como assistente de John Hammersley para estudar métodos Monte Carlo em 1955 conta que ele e mais alguns outros foram convocados para um exame. Chegando lá, todos foram postos para fazer uma prova de 4 horas com uma dezena de questões bem difíceis. No dia seguinte, durante as entrevistas, pediram para conhecer as respostas e saber quanto cada um tinha tirado na prova. Souberam então que não havia respostas para as questões. John havia passeado no departamento de física teórica, pediu para ver os problemas em que estavam trabalhando e montou as questões com esses problemas. Não era esperado que os estudantes acertassem, queriam ver apenas como cada candidato tentaria resolver os problemas. Ele faleceu em 2004. Hammersley foi um matemático muito criativo, do tipo resolvedor de problemas e não um construtor de teorias abstratas. Ele tinha o dom de identificar as questões matemáticas fundamentais por trás de problemas científicos e de construir um teoria útil em cima dessas questões. Ele sempre acreditou na ligação entre a matemática e situações do mundo real, negando que existia qualquer distinção entre matemática pura e aplicada. Acreditava também em um método de ensino de matemática segundo o qual a resolução de problemas deveria preceder a exposição da teoria, o que fez com que ele se juntasse ao movimento de New math. Se tornou reader na universidade de Oxford em 1969 e foi eleito Fellow of the Royal Society em 1976.



Teorema 6.1.1. Se $f(x_1|\mathbf{x}_{-1})$, $f(x_2|\mathbf{x}_{-2})$, ..., $f(x_n|\mathbf{x}_{-n})$ são as n distribuições condicionais completas de uma densidade conjunta $f(\mathbf{x})$ e se \mathbf{y} é uma configuração de referência qualquer em que $f(\mathbf{x}) > 0$, então a densidade conjunta

pode ser obtida a menos de uma constante de integração:

$$\frac{f(\mathbf{x})}{f(\mathbf{y})} = \prod_{i=1}^n \frac{f(x_i|x_1, \dots, x_{i-1}, y_{i+1}, \dots, y_n)}{f(y_i|x_1, \dots, x_{i-1}, y_{i+1}, \dots, y_n)}. \quad (6.1.1)$$

Demonstração. A partir da definição de densidade condicional, podemos escrever

$$f(\mathbf{x}) = f(x_n|x_1, x_2, \dots, x_{n-1})f(x_1, x_2, \dots, x_{n-1}).$$

Diferentemente do caso de séries temporais, o termo $f(x_1, x_2, \dots, x_{n-1})$ não pode ser fatorado de uma forma que nos seja conveniente. Além disso, a probabilidade $f(x_{n-1}|x_1, \dots, x_{n-2})$ não pode ser facilmente obtida a partir das condicionais completas que possuímos. Podemos, porém, inserir a observação y_n do estado de referência \mathbf{y} da seguinte maneira

$$\begin{aligned} f(\mathbf{x}) &= f(x_n|x_1, x_2, \dots, x_{n-1})f(x_1, x_2, \dots, x_{n-1}) \\ &= f(x_n|x_1, x_2, \dots, x_{n-1}) \frac{f(x_1, x_2, \dots, x_{n-1})}{f(x_1, x_2, \dots, x_{n-1}, y_n)} f(x_1, x_2, \dots, x_{n-1}, y_n) \\ &= \frac{f(x_n|x_1, x_2, \dots, x_{n-1})}{f(y_n|x_1, x_2, \dots, x_{n-1})} f(x_1, x_2, \dots, x_{n-1}, y_n). \end{aligned} \quad (6.1.2)$$

Vamos tratar de maneira análoga o termo $f(x_1, x_2, \dots, x_{n-1}, y_n)$, incluindo então a observação y_{n-1} :

$$f(x_1, x_2, \dots, x_{n-1}, y_n) = \frac{f(x_{n-1}|x_1, \dots, x_{n-2})}{f(y_{n-1}|x_1, \dots, x_{n-2}, y_n)} f(x_1, x_2, \dots, x_{n-2}, y_{n-1}, y_n).$$

Substituindo na Equação 6.1.2:

$$\frac{f(x_1, x_2, \dots, x_n)}{f(x_1, x_2, \dots, x_{n-2}, y_{n-1}, y_n)} = \frac{f(x_n|x_1, x_2, \dots, x_{n-1})}{f(y_n|x_1, x_2, \dots, x_{n-1})} \frac{f(x_{n-1}|x_1, x_2, \dots, x_{n-2}, y_n)}{f(y_{n-1}|x_1, x_2, \dots, x_{n-2}, y_n)}.$$

Introduzindo as observações do estado de referência \mathbf{y} uma a uma, chegamos à igualdade 6.1.1. ♠ □

Peter Clifford é bacharel pelo University College, Londres, em 1965 e Ph.D. em estatística pela University of California, Berkeley, 1969. De acordo com Clifford (1990), Jerzy Neyman, chefe do departamento de estatística em Berkeley, convidou John Hammersley e vários outros visitantes, incluindo Peter Clifford, no verão de 1971. Foi nesta época que eles provaram o famoso teorema que leva o nome deles. Peter foi professor nas Universidade de Tel Aviv e Bristol. Entre 1975 e 1976 ocupou o cargo de pesquisador visitante na área de estatística, na Universidade da Califórnia, Berkeley. Tornou-se então pesquisador do Radiation and Environmental Research Institute. A partir de 1976 se tornou pesquisador e professor de matemática em Oxford onde ocupa hoje o cargo de reader em Estatística Matemática. Ele produziu mais de 80 artigos numa ampla gama de assuntos, desde artigos de probabilidade e matemática pura até artigos bastante aplicados e outros de cunho mais político sobre educação.



Essa igualdade mostra alguns pontos importantes sobre a especificação da distribuição conjunta a partir de condicionais. Primeiramente, como a inclusão dos sítios de referência y_i foi feita de maneira arbitrária, essa mesma relação é válida para qualquer permutação do vetor \mathbf{y} . Isso significa que poderíamos entrar com as observações desse vetor de um maneira totalmente arbitrária e mesmo assim deveríamos chegar ao mesmo resultado. Essa condição implica algumas restrições na forma funcional das probabilidades condicionais a fim de se conseguir uma distribuição conjunta consistente. No capítulo sobre Campos Gaussianos encontramos uma restrição desse tipo para os parâmetros da matriz de precisão.

Um outro ponto relevante em relação à expansão de Brook é que ela fornece apenas uma razão de densidades com respeito a um vetor de referência \mathbf{y} . Para encontrarmos a densidade conjunta devemos chegar ainda à constante de normalização o que, na maioria dos casos, não é uma tarefa simples. Quando isso não é possível de ser feito analiticamente, recorremos a métodos de integração numérica.

6.2 Algumas definições importantes

Definição 6.2.1. Um sítio j é chamado de vizinho de i (onde $j \neq i$) se, e somente se, a forma funcional da densidade condicional $f(x_i|x_1, x_2, \dots, x_{i-1}, x_{i+1}, \dots, x_n)$ depende da variável x_j .

Vamos fazer duas suposições. A primeira delas é que existe apenas um número finito de estados disponíveis para cada sítio, suposição esta que será relaxada posteriormente. Sem perda de generalidade, vamos também supor que o vetor de referência \mathbf{y} é o vetor $\mathbf{0} = (0, \dots, 0)$ de forma que todo sítio pode assumir o valor zero, ou seja, $f(x_i = 0) > 0 \forall i$. Essa suposição, juntamente com a condição de positividade, mencionada anteriormente, garantem que uma realização formada completamente por zeros seja possível, ou seja, $f(\mathbf{0}) > 0$.

Tendo isso em vista, podemos definir

$$H(\mathbf{x}) = \ln \frac{f(\mathbf{x})}{f(\mathbf{0})}$$

para todo \mathbf{x} no espaço amostral Ω .

Para cada realização \mathbf{x} definiremos \mathbf{x}_i a realização que é igual à \mathbf{x} em todos os sítios, com exceção do i -ésimo, que receberá o valor zero. Dessa maneira a realização \mathbf{x}_i é dada por

$$\mathbf{x}_i = (x_1, x_2, \dots, x_{i-1}, 0, x_{i+1}, \dots, x_n)$$

Besag (1974) percebeu que o problema se reduzia a encontrar a forma mais geral que $H(\mathbf{x})$ poderia assumir, a fim de garantir uma estrutura de probabilidades condicionais válida para o sistema. Realmente basta encontrar esta forma geral de $H(\mathbf{x})$ pois, como

$$\begin{aligned} \exp\{H(\mathbf{x}) - H(\mathbf{x}_i)\} &= \exp\left(\ln \frac{f(\mathbf{x})}{f(\mathbf{0})} - \ln \frac{f(\mathbf{x}_i)}{f(\mathbf{0})}\right) \\ &= \exp\left(\ln \frac{f(\mathbf{x})}{f(\mathbf{x}_i)}\right) = \frac{f(\mathbf{x})}{f(\mathbf{x}_i)} \\ &= \frac{f(x_i | x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n) f(x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n)}{f(0 | x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n) f(x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n)}, \end{aligned}$$

então

$$\exp\{H(\mathbf{x}) - H(\mathbf{x}_i)\} = \frac{f(x_i | x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n)}{f(0 | x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n)}. \quad (6.2.3)$$

O valor $f(0 | x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n)$ da densidade condicional de X_i no ponto 0 é uma constante com respeito ao argumento x_i . Portanto, a menos de uma constante, o conhecimento da forma funcional de $\exp\{H(\mathbf{x}) - H(\mathbf{x}_i)\}$

leva-nos, automaticamente, à forma funcional de $f(x_i|x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n)$. Como esta última é uma densidade de probabilidade, a constante é obtida por integração.

Uma idéia muito criativa de Besag (1974) foi encontrar uma expansão de $H(\mathbf{x})$ que é única no espaço amostral Ω . Essa expansão é a seguinte

$$\begin{aligned} H(\mathbf{x}) = & \sum_{1 \leq i \leq n} x_i G_i(x_i) + \sum \sum_{1 \leq i \leq j \leq n} x_i x_j G_{i,j}(x_i, x_j) + \\ & + \sum \sum \sum_{1 \leq i < j < k \leq n} x_i x_j x_k G_{i,j,k}(x_i, x_j, x_k) + \dots \\ & + \dots + x_1 x_2 \dots x_n G_{1,2,\dots,n}(x_1, x_2, \dots, x_n). \end{aligned} \quad (6.2.4)$$

As funções G 's só podem ser as seguintes:

$$\begin{aligned} x_i G_i(x_i) &= H(0, \dots, 0, x_i, 0, \dots, 0) - H(0, 0, \dots, 0) \\ x_i x_j G_{i,j}(x_i, x_j) &= H(0, \dots, x_i, \dots, x_j, \dots, 0) - H(0, \dots, x_i, \dots, 0) - \\ &\quad - H(0, \dots, x_j, \dots, 0) + 2H(0, 0, \dots, 0) \end{aligned}$$

e assim por diante, e por isto elas são únicas.

Exemplo 6.2.1: Para ver esta expansão de $H(\mathbf{x})$ num caso simples, suponha que (X, Y) é um vetor com distribuição normal bivariada onde cada marginal é uma $N(0, 1)$ e com correlação igual a $\rho \in (-1, 1)$. Portanto,

$$f(x, y) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp \left\{ \frac{-1}{2(1-\rho^2)} [x^2 + y^2 - 2\rho xy] \right\}$$

e

$$H(x, y) = \log \left(\frac{f(x, y)}{f(0, 0)} \right) = -\frac{1}{2(1-\rho^2)} (x^2 + y^2 - 2\rho xy).$$

Dessa forma, temos

$$\begin{aligned}
xG_X(x) &= H(x, 0) - H(0, 0) = -\frac{1}{2(1-\rho^2)}x^2 \\
&\implies G_X(x) = -\frac{x}{2(1-\rho^2)} \\
yG_Y(y) &= H(0, y) - H(0, 0) = -\frac{1}{2(1-\rho^2)}y^2 \\
&\implies G_Y(y) = -\frac{y}{2(1-\rho^2)} \\
xyG_{XY}(x, y) &= H(x, y) - H(x, 0) - H(0, y) + H(0, 0) = \frac{\rho}{1-\rho^2}xy \\
&\implies G_{XY}(x, y) = \frac{\rho}{(1-\rho^2)}.
\end{aligned}$$

É claro que, com as definições acima, temos a identidade

$$\log \left(\frac{f(x, y)}{f(0, 0)} \right) = H(x, y) = xG_X(x) + yG_Y(y) + xyG_{XY}(x, y).$$



Voltando ao caso geral, nós podemos notar que

$$H(0, \dots, 0, x_i, \dots, x_j, 0, \dots, 0) = x_iG_i(x_i) + x_jG_j(x_j) + x_i x_j G_{ij}(x_i, x_j).$$

De maneira análoga, é fácil perceber que

$$\begin{aligned}
H(0, \dots, 0, x_i, \dots, x_j, 0, \dots, x_k, 0, \dots, 0) &= x_iG_i(x_i) + x_jG_j(x_j) + x_kG_k(x_k) + \\
&\quad + x_i x_j G_{ij}(x_i, x_j) + x_i x_k G_{ik}(x_i, x_k) \\
&\quad + x_j x_k G_{jk}(x_j, x_k) \\
&\quad + x_i x_j x_k G_{ijk}(x_i, x_j, x_k)
\end{aligned}$$

e finalmente que a expansão é (6.2.4) válida para qualquer ponto \mathbf{x} .

6.3 O Teorema de Hammersley-Clifford

Apresentaremos agora o resultado principal demonstrado por Besag em seu artigo de 1974.

Teorema 6.3.1. *Para cada $1 \leq i < j < \dots < s \leq n$ a função $G_{i,j,\dots,s}$ será não nula se, e somente se, os sítios i, j, \dots, s formarem uma clique. Sujeita a essa restrição, as funções G podem ser escolhidas arbitrariamente.*

Dessa maneira, conhecidos os vizinhos de cada sítio, podemos escrever a forma mais geral da função $H(\mathbf{x})$ e, portanto, das distribuições condicionais.

Demonstração. Como já vimos,

$$\exp\{H(\mathbf{x}) - H(\mathbf{x}_i)\} = \frac{f(x_i|x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n)}{f(0|x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n)}. \quad (6.3.5)$$

Portanto, pela definição de vizinhança em (6.2.1), a função $H(\mathbf{x}) - H(\mathbf{x}_i)$ só pode depender de x_i e de seus vizinhos.

Sem perda de generalidade, vamos considerar apenas o sítio 1. Então $H(\mathbf{x}) - H(\mathbf{x}_1)$ é dada por

$$\begin{aligned} x_1 G_1(x_1) + x_1 \sum_{2 \leq j \leq n} x_j G_{1,j}(x_1, x_j) + x_1 \sum \sum_{2 \leq j < k \leq n} x_j x_k G_{1,j,k}(x_1, x_j, x_k) + \dots \\ + x_1 x_2 x_3 \dots x_n G_{1,2,\dots,n}(x_1, x_2, \dots, x_n). \end{aligned}$$

Ao final da subtração de $H(\mathbf{x}_1)$, restam apenas os termos que possuem x_1 pois os demais foram cancelados. Considere agora um sítio $l \neq 1$ que não seja vizinho do sítio 1. Fazendo $x_i = 0$ para todo i diferente de l ou de 1, obtemos

$$H(\mathbf{x}) - H(\mathbf{x}_1) = x_1 x_l G_{1,l}(x_1, x_l).$$

Mas, por (6.3.5), $H(\mathbf{x}) - H(\mathbf{x}_1)$ não pode depender de x_l já que ele não é vizinho de 1. Mas a única forma de $H(\mathbf{x}) - H(\mathbf{x}_1)$ não depender de x_l é se $G_{1,l}(x_1, x_l) = 0$

Vamos agora considerar um terceiro sítio j e zerar todos x_i com i diferente de j , l e 1. Temos então que

$$H(\mathbf{x}) - H(\mathbf{x}_1) = x_1 x_l G_{1,l}(x_1, x_l) + x_1 x_j G_{1,j}(x_1, x_j) + x_1 x_l x_j G_{1,l,j}(x_1, x_l, x_j). \quad (6.3.6)$$

Existem três casos possíveis:

- j não é vizinho de l nem de 1. Neste caso, pelo raciocínio para os pares de não vizinhos, sabemos que $G_{j,l}(x_j, x_l) = G_{1,j}(x_1, x_j) = 0$. Logo, teríamos

$$H(\mathbf{x}) - H(\mathbf{x}_1) = x_1 x_l x_j G_{1,l,j}(x_1, x_l, x_j).$$

Mas como l não é vizinho de 1, $H(\mathbf{x}) - H(\mathbf{x}_1)$ não pode depender de x_l . A única forma disso ocorrer é se $G_{1,l,j}(x_1, x_l, x_j) = 0$.

- j é vizinho de l e de 1. Apesar de j ser vizinho de l e de 1, o trio não forma uma clique, pois l e 1 não são vizinhos. Precisamos mostrar que $G_{1,l,j}(x_1, x_l, x_j) = 0$. Como j é vizinho de 1, podemos ter que $G_{1,j}(x_1, x_j) \neq 0$. Logo, (6.3.6) fica reduzida a

$$H(\mathbf{x}) - H(\mathbf{x}_1) = x_j x_1 G_{1,j}(x_1, x_j) + x_1 x_l x_j G_{1,l,j}(x_1, x_l, x_j).$$

Como $x_j x_1 G_{1,j}(x_1, x_j)$ é uma constante com respeito a x_l , a única maneira de $H(\mathbf{x}) - H(\mathbf{x}_1)$ não depender de x_l é se $G_{1,l,j}(x_1, x_l, x_j) = 0$.

- j é vizinho apenas de 1 ou vizinho apenas de l : são situações mais simples que a anterior e a prova é praticamente a mesma.

De maneira análoga ao que foi feito acima, pode-se acrescentar um sítio de cada vez e é fácil perceber que serão nulas todas as funções G com $4, 5, \dots, n$ variáveis envolvendo x_1 e x_l . Obviamente, o mesmo ocorre para quaisquer pares de sítios que não sejam vizinhos.

Portanto, em geral, a função $G_{i,j,\dots,s}(x_i, x_j, \dots, x_s)$ só poderá ser não nula se o conjunto de sítios i, j, \dots, s for uma *clique*. Se dois desses sítios não forem vizinhos, a função deve ser nula, como mostrado acima.

Provamos até aqui que, se a função $G_{i,j,\dots,s}(x_i, x_j, \dots, x_s)$ for não-nula, então os sítios $1 \leq i < j < \dots < s \leq n$ formam uma *clique*. Para concluirmos a prova do teorema, basta notar que $H(\mathbf{x}) - H(\mathbf{x}_i)$ depende de x_l apenas se existir uma função G não-nula que envolva tanto i quanto l . Portanto, x_l também iria aparecer na densidade condicional $f(x_i | \mathbf{x}_{-i})$ e, em consequência, seria vizinho de i (e portanto membro da clique formada pelo par i e l). Finalmente, qualquer conjunto de funções G dão origem a uma distribuição conjunta $f(\mathbf{x})$ válida que satisfaz a propriedade de positividade e isto conclui a prova do teorema. ♠

□

6.4 Extensões

Demonstrado o teorema, vamos agora apresentar algumas extensões do mesmo. Vamos primeiramente ver quais condições devem ser satisfeitas para que possamos relaxar a suposição de que o número de possibilidades para cada sítio é finito. Se esse número for infinito, basta que $\sum \exp\{H(\mathbf{x})\}$ seja finita pois

$$\sum \exp\{H(\mathbf{x})\} = \sum \frac{f(\mathbf{x})}{f(\mathbf{0})} = \frac{1}{f(\mathbf{0})}.$$

Para o caso em que as variáveis são contínuas a restrição é semelhante, bastando substituir o somatório pela integral.

Quanto à suposição de positividade, ela não é necessária para garantir que condicionais podem determinar a conjunta, conforme vimos no capítulo 2. Entretanto, um contra-exemplo apresentado por Moussouris mostrou que a conclusão do teorema de Hammersley-Clifford (campo de Markov se, e só se, a conjunta é produtos de funções que dependem de cliques) depende da condição de positividade.

Porém, talvez isso não seja de muito interesse prático, visto que na maioria das situações reais estamos, de fato, sob condições de positividade.

Um outro problema que pode surgir quando se estuda Campos de Markov é o problema das bordas, visto que elas apresentam uma estrutura de vizinhança diferente do resto do grafo. Uma solução utilizada é fixar os sítios desses locais e encontrar a distribuição conjunta supondo conhecido seus valores.

Para finalizar, vamos agora apresentar um corolário do Teorema 6.3.1. Vamos primeiramente definir a realização \mathbf{x}_{ijk} como sendo

$$(x_1, \dots, x_{i-1}, 0, x_{i+1}, \dots, x_{j-1}, 0, x_{j+1}, \dots, x_{k-1}, 0, x_{k+1})$$

ou seja, substituímos na realização \mathbf{x} os valores dos sítios i, j, k por zero. Dessa maneira

$$H(\mathbf{x}) - H(\mathbf{x}_{ijk}) = \frac{f(X_i = x_i, X_j = x_j, X_k = x_k | \text{os valores dos demais sítios})}{f(X_i = 0, X_j = 0, X_k = 0 | \text{os valores dos demais sítios})}.$$

Utilizando a igualdade (6.3.1), ao tomarmos a diferença $H(\mathbf{x}) - H(\mathbf{x}_{ijk})$, ficamos apenas com os termos que envolvem x_i, x_j, x_k e os seus vizinhos, visto

que as funções-G são nulas para os demais casos. Portanto a densidade

$$f(x_i, x_j, x_k | \text{os valores dos demais sítios})$$

das variáveis X_i, X_j e X_k , condicionada nas demais, só depende dos valores dos sítios i, j, k e de seus vizinhos. Esse resultado pode ser facilmente estendido para o caso mais geral, e portanto para qualquer Campo de Markov

$$f(x_i, x_j, \dots, x_k | \text{os valores dos demais sítios})$$

depende apenas dos sítios i, j, \dots, k e de seus vizinhos. ♠

6.5 Voltando com energia

No Capítulo 3 nós definimos as distribuições de Gibbs induzida por um potencial de interação \mathcal{U} como densidades conjuntas da seguinte forma

$$f(\mathbf{x}) = \frac{1}{Z} \exp \left(- \sum_{A \in \mathcal{A}} U_A(\mathbf{x}_A) \right). \quad (6.5.7)$$

Existe uma certa ambiguidade na forma de escrever a distribuição de Gibbs. Com um potencial de interação qualquer, é sempre possível reescrever (6.5.8) da seguinte maneira:

$$f(\mathbf{x}) = \frac{1}{Z} \exp \left(- \sum_{A \in \mathcal{A}} U_A(\mathbf{x}_A) \right) = \frac{1}{Z^*} \exp \left(- \sum_{A \in \mathcal{A}} (c + U_A(\mathbf{x}_A)) \right) \quad (6.5.8)$$

onde $Z^* = Z \exp(kc)$ onde k é o número de conjuntos na coleção \mathcal{A} do potencial de interação. Como veremos abaixo, a expansão (6.2.4) de Besag vai eliminar este tipo de ambigüidade. Por enquanto, vamos imaginar que a função energia é arbitrária, sujeita apenas à somabilidade, que permite a existência da constante de integração.

Pela demonstração de Besag, sabemos que

$$H(\mathbf{x}) = \log \left(\frac{f(\mathbf{x})}{f(\mathbf{0})} \right) \quad (6.5.9)$$

$$\begin{aligned} &= \sum_{1 \leq i \leq n} x_i G_i(x_i) + \sum \sum_{1 \leq i \leq j \leq n} x_i x_j G_{i,j}(x_i, x_j) + \\ &\quad + \sum \sum \sum_{1 \leq i < j < k \leq n} x_i x_j x_k G_{i,j,k}(x_i, x_j, x_k) + \dots \\ &\quad + \dots + x_1 x_2 \dots x_n G_{1,2,\dots,n}(x_1, x_2, \dots, x_n) \end{aligned} \quad (6.5.10)$$

e que esta expansão é única. Isto é, as funções G_A são únicas, onde A é um subconjunto de índices.

Mas se a distribuição conjunta é uma distribuição de Gibbs do tipo (6.5.8), então

$$\begin{aligned} H(\mathbf{x}) &= \log \left(\frac{\exp(-\sum_A U_A(\mathbf{x}))}{\exp(-\sum_A U_A(\mathbf{0}))} \right) \\ &= c - \sum_A U_A(\mathbf{x}) \end{aligned}$$

onde $c = \sum_A U_A(\mathbf{0})$. Entretanto, pela unicidade e pela forma da expressão (6.5.10), temos de ter $c = 0$. Isto elimina aquela ambigüidade aludida acima. Como resultado final, temos

$$H(\mathbf{x}) = - \sum_{A \in \mathcal{A}} U_A(\mathbf{x}) = - \sum_{A \in \mathcal{A}} U_A(\mathbf{x}_A).$$

Besag demonstrou então que, para que a conjunta (6.5.8) seja um campo de Markov com respeito a uma estrutura de vizinhança \mathcal{N} associada com um grafo, os conjuntos A do potencial tem de ser *cliques* desse grafo.

Exemplo: 6.5.1: Vamos retomar o grafo de vizinhança associado com uma grade regular e com a estrutura de 4-vizinhos (ver grafo a esquerda na Figura 6.1). Então um campo de Markov nesta estrutura de vizinhança só pode ser da seguinte forma:

$$f(\mathbf{x}) = \frac{1}{Z} \exp \left(\sum_{i=1}^n x_i G_i(x_i) + \sum \sum_{i \sim j} x_i x_j G_{i,j}(x_i, x_j) \right). \quad (6.5.11)$$

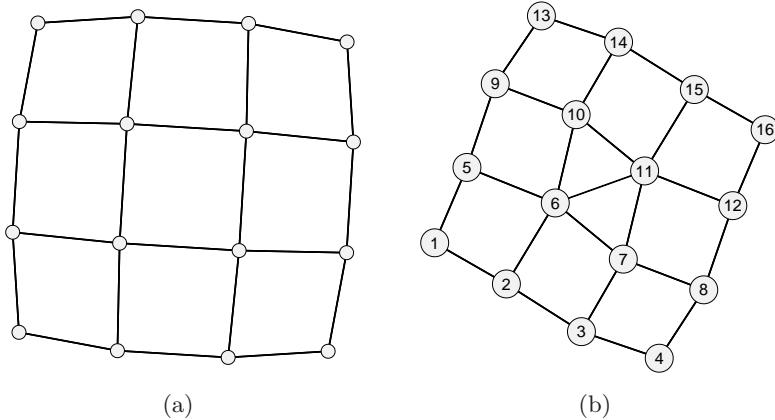


Figura 6.1: Grafo regular com estrutura de 4-vizinhos (esquerda) e grafo com estrutura de 4-vizinhos adicionada de alguns outros pares de vizinhos.

Na densidade conjunta podem aparecer funções que dependem até, no máximo, dos valores x 's nos pares de vizinhos. Não podem aparecer funções envolvendo pares de não-vizinhos nem triplas de nós. Para que triplas pudessem aparecer, o grafo de vizinhança deveria ser modificado. Por exemplo, suponha que a densidade conjunta seja da forma

$$f(\mathbf{x}) = \frac{1}{Z} \exp \left(\sum_{i=1}^n x_i G_i(x_i) + \sum \sum_{i \sim j} x_i x_j G_{i,j}(x_i, x_j) + \beta(x_6 x_7 x_{11} + x_6 x_{10} x_{11}) \right).$$

Esta densidade de Gibbs é um campo de Markov com respeito a um grafo de vizinhança como aquele a direita na Figura (6.1), mas não com respeito a um grafo de vizinhança como aquele da esquerda nesta Figura.

Vamos assumir que as variáveis x_i sejam binárias. Existem dois tipos comuns de valores para x_i : o primeiro assume que $x_i = -1$ ou $x_i = +1$ (como no modelo de Ising do Capítulo 1) enquanto o segundo assume que $x_i = 0$ ou $x_i = 1$ (vamos chamá-lo Bernoulli). A densidade conjunta é igual a

$$f(\mathbf{x}) = \frac{1}{Z} \exp \left(\sum_{i:x_i=+1} G_i(1) - \sum_{i:x_i=-1} G_i(-1) + \sum_{\substack{i \sim j \\ x_i=x_j}} G_{i,j}(x_i, x_j) - \sum_{\substack{i \sim j \\ x_i \neq x_j}} G_{i,j}(x_i, x_j) \right)$$

no caso de Ising, e igual a

$$f(\mathbf{x}) = \frac{1}{Z} \exp \left(\sum_{i:x_i=1} G_i(1) + \sum_{\substack{i \sim j \\ x_i=x_j=1}} G_{i,j}(1, 1) \right)$$

no caso binário.

Se assumimos que $G_i(1) = G_i(-1) = \alpha$ e $G_{i,j}(1, 1) = G_{i,j}(-1, -1) = \beta$ e $G_{i,j}(1, -1) = G_{i,j}(-1, 1) = -\beta$ teremos no caso de Ising a densidade conjunta

$$f(\mathbf{x}) = \frac{1}{Z} \exp (\alpha(2n_+ - n) + \beta(n_s - n_d))$$

onde n_+ é o número de vértices com valor $x_i = +1$, n é o número total de vértices, n_s e n_d é o número de pares de vizinhos com valores iguais e diferentes, respectivamente.

Para o caso binário, a densidade conjunta é fica igual a

$$f(\mathbf{x}) = \frac{1}{Z} \exp (\alpha n_{++} + \beta n_{+-})$$

onde n_{++} é o número de pares de vizinhos em que ambos são iguais a $+1$.

Capítulo 7

Exemplos de Auto-modelos

7.1 Introdução

Vamos agora apresentar uma classe particular de campos de Markov, que são os denominados auto-modelos. Estes modelos foram definidos no mesmo artigo Besag (1974) em que o teorema de Hammersley-Clifford foi provado. Faremos aqui algumas suposições que são apresentadas a seguir.

Suposição 1: A estrutura de probabilidade do sistema satisfaz à condição de positividade e depende apenas de *cliques* com no máximo dois sítios. Isso significa que a expansão de $H(\mathbf{x})$ apresentada em 6.1.2 se reduz a

$$H(\mathbf{x}) = \sum_{1 \leq i \leq n} x_i G_i(x_i) + \sum_{1 \leq i < j \leq n} x_i x_j G_{i,j}(x_i, x_j) \quad (7.1.1)$$

onde $G_{i,j}(x_i, x_j) \neq 0$ se, e somente se, i e j são vizinhos.

Suposição 2: A distribuição de probabilidade condicional associada com cada um dos sítios pertence à família exponencial. Isso significa que

$$\ln p_i(x_i; \dots) = A_i(\cdot)B_i(x_i) + C_i(x_i) + D_i(\cdot) \quad (7.1.2)$$

onde

- $p_i(\cdot)$ irá denotar, daqui pra frente, a distribuição de probabilidade condicional de X_i dado todos os outros sítios

- B_i e C_i dependem apenas de x_i e apresentam uma forma específica
- A_i e D_i são funções dos valores dos sítios vizinhos de i . A escolha de A_i determina o tipo de dependência entre os sítios vizinhos e D_i é a constante de normalização. As funções A_i devem ser válidas no sentido de garantir a existência da constante de normalização.

Julian E. Besag nasceu em 1946 na Inglaterra. Ele iniciou seus estudos em engenharia em Cambridge mas não completou-os. Ele obteve um bacharelado da Universidade de Birmingham em 1968 em estatística. Ele passou a trabalhar como assistente de pesquisa de Maurice Bartlett no Departamento de Biomatemática em Oxford em problemas de processos estocásticos em mais de uma dimensão e, em particular, em campos de Markov. Nesta época, como funcionário do Science Research Council, ele não podia se matricular em cursos de pós-graduação e por isto ele não possui doutorado. Ele passou a trabalhar como professor na Liverpool University. Em 1974 ele teve seu primeiro read paper na Royal Statistical Society. Em 1975 ele visitou Princeton por 6 meses onde aprendeu EDA com John Tukey. Na volta para a Inglaterra, ele passou a trabalhar na Durham University junto com Peter Green. Um outro read paper, de 1986, foi o mais citado durante a década de 80 dentre todos os matemáticos do Reino Unido. O trabalho de Ulf Grenander e dos irmãos Geman levou Julian a adotar os métodos bayesianos e MCMC para a análise de dados espaciais. Ele mudou-se para Seattle, tornando-se full professor na University of Washington em 1991. Julian teve ainda mais dois read papers, em 1993 e 1998. Ele aposentou-se em 2006. Ele publicou um número relativamente pequeno de artigos, mas quase todos eles tiveram enorme impacto em estatística. Ele possui dois artigos na coleção Breakthroughs in Statistics, um feito e tanto. Julian esteve no Brasil participando do XII SINAPE em Caxambu no ano de 1996.



Vamos demonstrar que, como consequência direta dessas duas suposições, a função A_i deve apresentar o seguinte formato:

$$A_i(\cdot) = \alpha_i + \sum_{j=1}^n \beta_{i,j} B_j(x_j) \quad (7.1.3)$$

onde, por definição, $\beta_{i,j} = \beta_{j,i}$ e $\beta_{i,j} = 0$, a não ser que os sítios i e j sejam vizinhos. Isso implica que $G_{i,j}$ na expressão (7.1.1) deve ter a seguinte forma:

$$G_{i,j}(x_i, x_j) = \beta_{i,j} W_i(x_i) W_j(x_j)$$

onde $x_i W_i(x_i) = B_i(x_i) - B_i(0)$. Passemos agora à demonstração de (7.1.3).

Demonstração. Sem perda de generalidade assuma que o estado de referência $\mathbf{0} = (0, \dots, 0)$ faz parte do suporte da densidade conjunta. Pela suposição de positividade, $\ln(p_i(0; \dots)) > 0$ para qualquer configuração nos demais sítios. Vamos escrever A_i e D_i da equação (7.1.2) como função de

$$(x_1, \dots, x_{i-1}, 0, x_{i+1}, \dots, x_n).$$

Sabemos, porém, que A_i e D_i , na verdade, só dependem dos sítios vizinhos de i .

Como $p_i(0; \dots)$ é sempre maior que zero, sob condições de positividade, $H(\mathbf{x})$ está bem definida e pode ser escrita da forma apresentada em (7.1.1).

De (6.2.3) temos que

$$H(\mathbf{x}) - H(\mathbf{x}_i) = \ln f(x_i | x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n) - \ln f(0 | x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n)$$

utilizando, então, (7.1.2) nota-se que

$$\begin{aligned} H(\mathbf{x}) - H(\mathbf{x}_i) &= A_i(x_j \in \mathcal{N}_i)B_i(x_i) + C_i(x_i) + D(x_j \in \mathcal{N}_i) - A_i(x_j \in \mathcal{N}_i)B_i(0) \\ &\quad - C_i(0) - D(x_j \in \mathcal{N}_i) = A_i(x_j \in \mathcal{N}_i)(B_i(x_i) - B_i(0)) + C_i(x_i) - C_i(0) \end{aligned}$$

lembrando que \mathcal{N}_i denota o conjunto de vizinhos de i .

Além disso, de (7.1.1) é fácil perceber que

$$H(\mathbf{x}) - H(\mathbf{x}_i) = x_i G_i(x_i) + \sum_j x_i x_j G_{i,j}(x_i, x_j) \quad (7.1.4)$$

fazendo $x_j = 0$ para todo $j \neq i$ chegamos a

$$x_i G_i(x_i) = A_i(\mathbf{0})(B_i(x_i) - B_i(0)) + C_i(x_i) - C_i(0) \quad (7.1.5)$$

Cabe notar aqui que estamos colocando $x_j = 0$ para todo $j \neq i$, pois como estamos interessados apenas em uma função de x_i , então não importam os valores de x_j para $j \neq i$.

Vamos supor, agora que os sítios 1 e 2 são vizinhos. Procedendo de maneira análoga ao que foi feito anteriormente, mas agora zerando x_j para $j \geq 3$ obtemos para $i = 1$

$$\begin{aligned} H(\mathbf{x}) - H(\mathbf{x}_1) &= x_1 G_1(x_1) + x_1 x_2 G_{1,2}(x_1, x_2) \\ &= A_1(x_2, \dots, 0)(B_1(x_1) - B_1(0)) + C_1(x_i) - C_1(0) \end{aligned} \quad (7.1.6)$$

e para $i = 2$

$$\begin{aligned} H(\mathbf{x}) - H(\mathbf{x}_2) &= x_2 G_2(x_2) + x_1 x_2 G_{1,2}(x_1, x_2) \\ &= A_2(x_1, \dots, 0)(B_2(x_2) - B_2(0)) + C_2(x_2) - C_2(0). \end{aligned} \quad (7.1.7)$$

Substituindo (7.1.5) em (7.1.6) e (7.1.7) chegamos às seguintes igualdades:

$$x_1 x_2 G_{1,2}(x_1, x_2) = [A_1(x_2, \dots, 0) - A(\mathbf{0})][B_1(x_1) - B_1(0)] \quad (7.1.8)$$

$$x_1 x_2 G_{1,2}(x_1, x_2) = [A_2(x_1, \dots, 0) - A(\mathbf{0})][B_2(x_2) - B_2(0)]. \quad (7.1.9)$$

Isso significa que $x_1 x_2 G_{1,2}(x_1, x_2)$ pode ser fatorada em duas funções, uma que depende só de x_1 e outra que depende só de x_2 , ou seja, podemos reescrever (7.1.8) e (7.1.9), respectivamente da seguinte maneira

$$x_1 x_2 G_{1,2}(x_1, x_2) = f_1(x_2)g_1(x_1)$$

$$x_1 x_2 G_{1,2}(x_1, x_2) = f_2(x_1)g_2(x_2)$$

logo

$$\frac{g_1(x_1)}{f_2(x_1)} = \frac{g_2(x_2)}{f_1(x_2)} = \beta_{1,2}$$

onde $\beta_{1,2}$ é uma constante. Essas razões devem ser constantes em relação a x_1 e x_2 pois, se dependessem de uma dessas variáveis, não valeria a igualdade acima.

Dessa maneira, $f_2(x_1) = \beta_{1,2}g_1(x_1)$ e portanto

$$x_1 x_2 G_{1,2}(x_1, x_2) = \beta_{1,2}g_1(x_1)g_2(x_2) = \beta_{1,2}[B_1(x_1) - B_1(0)][B_2(x_2) - B_2(0)].$$

Para dois sítios i e j genéricos teremos então que

$$x_i x_j G_{i,j}(x_i, x_j) = \beta_{i,j} [B_i(x_i) - B_i(0)] [B_j(x_j) - B_j(0)] . \quad (7.1.10)$$

Isso mostra então que podemos escrever

$$G_{i,j}(x_i, x_j) = \beta_{i,j} W_i(x_i) W_j(x_j)$$

para $x_i W_i(x_i) = B_i(x_i) - B_i(0)$. Vamos mostrar agora que como consequência do que foi provado acima, a igualdade (7.1.3) é válida. Sabemos que

$$x_i G_i(x_i) + \sum_j x_i x_j G_{i,j}(x_i, x_j) = A_i(x_j \in \mathcal{N}_i) (B_i(x_i) - B_i(0)) + C_i(x_i) - C_i(0)$$

substituindo (7.1.5) e (7.1.10) chegamos a

$$\begin{aligned} A_i(\mathbf{0})[B_i(x_i) - B_i(0)] + \sum_j \beta_{i,j} [B_i(x_i) - B_i(0)] [B_j(x_j) - B_j(0)] \\ = A_i(\cdot)[B_i(x_i) - B_i(0)] . \end{aligned}$$

Portanto, $A_i(\cdot)$ pode ser escrito como

$$A_i(\cdot) = \alpha_i + \sum_j \beta_{i,j} B_j(x_j) \quad (7.1.11)$$

onde $\alpha_i = A_i(\mathbf{0}) + \sum_j \beta_{i,j} B_j(0)$ e está demonstrada a igualdade. ♠ □

7.2 Modelos autonormal

Esse modelo faz parte da classe dos campos de Markov gaussianos apresentados anteriormente. Quando é razoável supor que as variáveis observadas seguem uma distribuição conjunta normal multivariada, estamos no caso de um modelo auto-normal. Em particular, podemos considerar

$$H(\mathbf{x}) = \sum_i -\frac{1}{2\sigma^2} (x_i - \mu_i)^2 + \frac{1}{\sigma^2} \sum_i \sum_j \beta_{i,j} (x_i - \mu_i)(x_j - \mu_j)$$

onde $\beta_{i,j} = \beta_{j,i}$ e $\beta_{i,j} = 0$, a não ser que os sítios i e j sejam vizinhos. Ignorando os termos multiplicativos que não envolvem x_i encontramos

$$\frac{f(x_i|x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n)}{f(0|x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n)} = \exp \left\{ -\frac{1}{2\sigma^2} \left((x_i - \mu_i)^2 + 2(x_i - \mu_i) \sum_j \beta_{i,j} (x_j - \mu_j) \right) \right\}.$$

Ignorando o denominador e completando o quadrado temos

$$p_i(x_i; \dots) \propto \exp \left\{ -\frac{1}{2\sigma^2} \left((x_i - \mu_i) - \sum_j \beta_{i,j} (x_j - \mu_j) \right)^2 \right\}.$$

Isto significa que, condicionada em seus vizinhos, a variável aleatória X_i tem distribuição normal com média $\mu_i + \sum_{j \in \eta_i} \beta_{i,j} (x_j - \mu_j)$ e variância σ^2 . Usando (6.1.1) podemos chegar à distribuição conjunta, que apresenta o seguinte formato:

$$f(\mathbf{x}) = (2\pi\sigma^2)^{-\frac{n}{2}} |\mathbf{Q}|^{\frac{1}{2}} \exp \left\{ -\frac{1}{2\sigma^2} (\mathbf{x} - \boldsymbol{\mu})^T \mathbf{Q} (\mathbf{x} - \boldsymbol{\mu}) \right\}$$

onde $\boldsymbol{\mu}$ é um vetor de médias de tamanho $n \times 1$ e \mathbf{Q} é uma matriz $n \times n$ cuja diagonal é formada por 1's e cujos elementos fora da mesma são $-\beta_{i,j}$. A matriz \mathbf{Q} é simétrica, pois $\beta_{i,j} = \beta_{j,i}$. Precisamos ainda que ela seja positiva definida para que a distribuição seja própria.

Modelos SAR

Vamos agora fazer uma distinção entre os modelos CAR, definidos de acordo com (5.1.1), e um outro modelo, definido por equações simultâneas, e conhecido como SAR, (*Simultaneous Autoregressive Models*). Este outro modelo é definido da seguinte maneira:

$$X_i = \mu_i + \sum_j \beta_{i,j} (X_j - \mu_j) + \epsilon_i \tag{7.2.12}$$

onde $\epsilon_1, \dots, \epsilon_n$ são variáveis aleatórias i.i.d. normalmente distribuídas com média zero e variância σ^2 . Neste modelo, $\beta_{ii} = 0$ como antes, mas não é mais necessário fazer a restrição de que $\beta_{ij} = \beta_{ji}$ para $i \neq j$. Esta especificação não é baseada em distribuições condicionais. São n equações estocásticas que dizem como as variáveis estão relacionadas. Se denotarmos $X_i - \mu_i$ por Z_i , então:

$$\begin{aligned} Z_1 &= \beta_{11}Z_1 + \beta_{12}Z_2 + \beta_{13}Z_3 + \cdots + \beta_{1n}Z_n + \epsilon_1 \\ Z_2 &= \beta_{21}Z_1 + \beta_{22}Z_2 + \beta_{23}Z_3 + \cdots + \beta_{2n}Z_n + \epsilon_2 \\ Z_3 &= \beta_{31}Z_1 + \beta_{32}Z_2 + \beta_{33}Z_3 + \cdots + \beta_{3n}Z_n + \epsilon_3 \\ \dots &\quad \dots \\ Z_n &= \beta_{n1}Z_1 + \beta_{n2}Z_2 + \beta_{n3}Z_3 + \cdots + \beta_{nn}Z_n + \epsilon_n \end{aligned}$$

As equações acima deixam claro que Z_i é definido em função dos outros elementos do vetor \mathbf{Z} . Entretanto, esses outros elementos $j \neq i$ terão sua definição envolvendo Z_i . Assim, não existe uma sequência recursiva que gere os Z_i . Pelo contrário, eles são determinados simultaneamente, dando origem ao nome do modelo.

Não é óbvio que o sistema de equações acima possua solução. Isto é, não é óbvio que exista uma distribuição conjunta para o vetor \mathbf{Z} tal que cada coordenada possa ser expressa por (7.2.12). Assim, somos levados à pergunta: em que condições as n equações representadas por (7.2.12) definem uma distribuição de probabilidade conjunta para o vetor \mathbf{Z} ? Caso exista uma distribuição satisfazendo todas as n equações, esta distribuição é única? Para responder a estas perguntas, vamos reescrever o sistema de equações simultâneas em notação matricial.

Em notação matricial, o modelo SAR é expresso da seguinte forma:

$$\mathbf{Z} - \mu = \mathbf{S}(\mathbf{Z} - \mu) + \epsilon \quad (7.2.13)$$

onde a matriz S de dimensão $n \times n$ é composta pelos elementos β_{ij} . Note que S não precisa ser simétrica mas $S_{ii} = 0$ para $i = 1, \dots, n$. A partir de 7.2.13, obtemos

$$(\mathbf{I} - \mathbf{S})(\mathbf{Z} - \mu) = \epsilon$$

onde, se $(\mathbf{I} - \mathbf{S})$ for inversível, então

$$\mathbf{Z} = \mu + (\mathbf{I} - \mathbf{S})^{-1} \epsilon \quad (7.2.14)$$

concluindo-se que

$$\mathbf{Z} \sim N_n(\mu, \sigma^2(\mathbf{I} - \mathbf{S})^{-1} (\mathbf{I} - \mathbf{S}^t)^{-1}) \quad (7.2.15)$$

Isto é, para o modelo SAR, a densidade conjunta assume o seguinte formato:

$$f(\mathbf{x}) = (2\pi\sigma^2)^{\frac{-n}{2}} |\beta| \exp \left\{ \frac{-1}{\sigma^2} (\mathbf{x} - \mu)^T \mathbf{Q}^T \mathbf{Q} (\mathbf{x} - \mu) \right\}$$

onde $\mathbf{Q} = \mathbf{I} - \mathbf{S}$. Note que, como já foi dito, não é necessário que $\beta_{ij} = \beta_{ji}$ pois a matriz $\mathbf{Q}^T \mathbf{Q}$ é sempre simétrica. Basta assegurar que ela não seja singular para que a densidade conjunta exista.

7.3 Modelos autologístico

Para modelos binários, nos quais todas as variáveis assumem valores zero ou um, as funções-G apresentadas em (7.1.1) só irão contribuir para $H(\mathbf{x})$ quando todos seus argumentos forem iguais a 1. Caso contrário, o fator que multiplica cada uma dessas funções será igual a zero. Dessa maneira, todas as funções que contribuem para essa soma dependerão apenas de seus índices e podemos substituir $G_i(x_i) = G_i(1) = \alpha_i$ e $G_{i,j}(x_i, x_j) = G_{i,j}(1, 1) = \beta_{i,j}$. Temos, então, a seguinte represemtação de $H(\mathbf{x})$

$$H(\mathbf{x}) = \sum_i \alpha_i x_i + \sum_i \sum_j \beta_{i,j} x_i x_j .$$

Mas de (6.2.3) e (7.1.4) temos que

$$\frac{f(x_i|x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n)}{f(0|x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n)} = \exp\{x_i G_i(x_i) + \sum_j x_i x_j G_{i,j}(x_i, x_j)\}$$

dessa forma

$$\frac{f(x_i|x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n)}{1 - f(x_i = 1|x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n)} = \exp\{\alpha_i x_i + \sum_j \beta_{i,j} x_i x_j\}$$

e portanto

$$p_i(x_i; \dots) = \frac{\exp\{x_i(\alpha_i + \sum_j \beta_{i,j} x_j)\}}{1 + \exp\{\alpha_i + \sum_j \beta_{i,j} x_j\}} \quad (7.3.16)$$

que é muito semelhante ao modelo logístico clássico de Cox. A diferença aqui é que as variáveis explicativas são elas mesmas variáveis observadas no processo e não variáveis explicativas.

7.4 Modelo autobinomial

Suponha que X_i tem uma distribuição condicional binomial com parâmetros m_i e θ_i , sendo que este último depende apenas dos valores dos sítios vizinhos. Dessa maneira, a razão de chances é dada por

$$\ln\left(\frac{\theta_i}{1-\theta_i}\right) = \alpha_i + \sum_j \beta_{i,j}x_j.$$

Quando $m_i = 1$ para todo i , caímos novamente no caso do modelo autologístico.

7.5 Modelo autoPoisson

Suponha agora que X_i tem distribuição Poisson com uma média μ_i que depende apenas dos valores dos sítios vizinhos. Dessa forma, μ_i pode ser expressa da seguinte maneira

$$\mu_i = \exp\left(\alpha_i + \sum_j \beta_{i,j}x_j\right).$$

Vamos ver agora que precisamos de uma restrição nos parâmetros para que esse modelo seja válido. Sabemos que

$$\exp\{H(\mathbf{x})\} = \frac{f(\mathbf{x})}{f(\mathbf{0})} \quad \text{e portanto} \quad \sum_{\mathbf{x}} \exp\{H(\mathbf{x})\} = \frac{1}{f(\mathbf{0})}$$

isso significa que precisamos garantir que $\sum_{\mathbf{x}} \exp\{H(\mathbf{x})\}$ converge. Vamos primeiro encontrar a forma de $H(\mathbf{x})$. Sabemos que

$$f(\mathbf{x}) = \frac{\prod_i (\alpha_i x_i + \sum_{j=i+1}^n \beta_{i,j} x_j)^{x_i} \exp\{-\sum_i (\alpha_i x_i + \sum_{j=i+1}^n \beta_{i,j} x_j)\}}{\prod_i x_i!}$$

e

$$f(\mathbf{0}) = \exp\left\{-\sum_i (\alpha_i x_i + \sum_{j=i+1}^n \beta_{i,j} x_j)\right\}$$

então

$$H(\mathbf{x}) = \sum_i x_i \log\{\mu_i\} - \sum_i \ln(x_i!) = \sum_i (\alpha_i x_i - \ln(x_i!)) + \sum_i \sum_{j:j>i} \beta_{i,j} x_i x_j$$

Dessa maneira

$$\exp\{Q(\mathbf{x})\} = \frac{\exp\{\sum_i x_i \alpha_i + \sum_i \sum_j \beta_{i,j} x_i x_j\}}{\prod_i x_i!}.$$

Temos então que se $\beta_{i,j} = 0$ a expressão se reduz a

$$\exp\{H(\mathbf{x})\} = \prod_i \frac{e^{x_i}}{x_i!}. \quad (7.5.17)$$

Vamos olhar agora para a soma sobre cada x_i . Usando a decomposição em Série de Taylor de e^e temos que

$$\sum_{x_i=0}^{\infty} \frac{e^{x_i}}{x_i!} = e^e$$

então somando $\exp\{H(\mathbf{x})\}$ em (7.5.17) em todos os x'_i s temos

$$\sum_{x_1=0}^{\infty} \dots \sum_{x_n=0}^{\infty} \exp\{H(\mathbf{x})\} = e^{ne}$$

ou seja, $\exp\{H(\mathbf{x})\}$ é somável nesse caso. Para o caso em que $\beta_{i,j} < 0$, é fácil perceber que

$$\frac{\exp\{\sum_i x_i \alpha_i + \sum_i \sum_j \beta_{i,j} x_i x_j\}}{\prod_i x_i!} < \prod_i \frac{e^{x_i}}{x_i!}$$

e como acabamos de ver, a segunda série converge. Logo, pelo critério da comparação, concluímos que $\exp\{H(\mathbf{x})\}$ também é somável quando $\beta_{i,j} < 0$.

Para o caso em que $\beta_{i,j} > 0$ não podemos garantir que a soma converge, como será mostrado a partir de um contra-exemplo. Considere distribuição

de um par de variáveis aleatórias X_1 e X_2 , dado que todos os outros sítios possuem valor zero. O termo $\exp\{H(\mathbf{x})\}$ se reduz então a

$$\begin{aligned}\exp Q(x_1, x_2, 0, \dots, 0) &= \frac{\exp(\alpha_1 x_1 + \alpha_2 x_2 + \beta_{1,2} x_1 x_2)}{x_1! x_2!} = \\ &\frac{\exp(\alpha_1 x_1 + \alpha_2 x_2)}{x_1! x_2!} \exp(\beta_{1,2} x_1 x_2).\end{aligned}$$

Vamos olhar agora só para o termo $\exp(\beta_{1,2} x_1 x_2)$. A soma desses termos em x_1 e x_2 claramente não converge, pois o termo geral da série não tende a zero. Isso implica que $\exp(H(x_1, x_2, 0, \dots, 0))$ não pode ser somável se $\beta_{i,j} > 0$. Percebe-se, portanto, que o modelo de poisson só é valido quando existe “repulsão” entre áreas vizinhas, o que o torna muito restritivo.

Exercício 7.5.1. Usando a expansão de Brook mostre que se

$$p(x_i|x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n) = \frac{\exp\{x_i(\alpha_i + \sum_j \beta_{i,j} x_j)\}}{1 + \exp\{\alpha_i + \sum_j \beta_{i,j} x_j\}}$$

então

$$\frac{f(\mathbf{x})}{f(\mathbf{0})} = \exp\left(\sum_i \alpha_i x_i + \sum_{i>j} \beta_{i,j} x_i x_j\right)$$

onde $\beta_{i,j} = 0$ a menos que i e j sejam vizinhos.

Exercício 7.5.2. Considere um auto modelo formado por duas variáveis aleatórias com distribuição poisson:

$$Y_1|Y_2 \sim Poisson(\mu_1) \quad \text{e} \quad Y_2|Y_1 \sim Poisson(\mu_2)$$

tais que

$$\mu_1 = 1.5 + 2.2Y_1 \quad \text{e} \quad \mu_2 = 3.5 + 2.2Y_2.$$

Encontre a conjunta $f(\mathbf{Y})$ e a razão dada por $H(\mathbf{Y}) = f(\mathbf{Y})/f(\mathbf{0})$ e mostre que a soma $\sum_y \exp(H(Y))$ não converge.

Capítulo 8

Simulação de Auto-modelos

Nesta seção apresentamos códigos para simular dados de alguns modelos apresentados nos capítulos anteriores. Dados com distribuição CAR podem ser simulados facilmente considerando a distribuição conjunta. Para os demais modelos, podemos usar o algoritmo amostrador de Gibbs.

8.1 Modelo CAR

Para simular dados da distribuição CAR, usamos um algoritmo padrão para simular dados da distribuição normal multivariada já que a distribuição conjunta do modelo CAR é normal multivariada. Para simular uma amostra da distribuição normal n -variada $N(\mu, \Sigma)$, com matriz Σ definida positiva, usamos o seguinte algoritmo:

1. obtenha a decomposição de Choleski \mathbf{L} da matriz de covariância, tal que $\mathbf{L}'\mathbf{L} = \Sigma$;
2. simule n amostras aleatórias da distribuição normal padrão (univariada) e guarde num vetor \mathbf{z} ,
3. multiplique \mathbf{z} por \mathbf{L} e some com o vetor μ .

O vetor resultante do último passo é uma amostra aleatória da distribuição normal n -variada $N(\mu, \Sigma)$.

A seguir, nós apresentamos um código em **R**, R Core Team (2009), para simular dados com distribuição CAR determinada pelas densidades condicionais (5.1.1). Vamos fixar os parâmetros $\sigma^2 = 1$ e $\rho = 0.7$. Considerando o grafo da figura 3.1, vamos construir a matriz de precisão **Q** e simular dados dessa distribuição. Neste exemplo, vamos usar algumas funções do pacote **spdep**, Bivand et. al (2010).

```

### lista de vizinhança do grafo da Figura 3.1
nb <- list(c(2,5), c(1,3,5), c(2,4),
           c(3,5,6), c(1,2,4), c(4))

### tipa o objeto para a classe 'nb' do pacote 'spdep'
for (i in 1:6)
  nb[[i]] <- as.integer(nb[[i]])
class(nb) <- "nb"

### carrega pacote 'spdep'
require(spdep)

### obtem matriz de adjacencia a partir de nb
A <- nb2mat(nb,
            style="B")

# vetor com numero de vizinhos
ni <- sapply(nb, length)
### constroi matriz Q
rho <- 0.7
Q <- diag(ni) - rho*A

### matriz de covariancia
covar <- solve(Q)

### simulacao da distribuicao conjunta
set.seed(1)
z <- rnorm(6)
L <- chol(covar)

```

```
### vetor de dados com distribuicao CAR com media zero:
round((z%*%L), 3)
[,1]   [,2]   [,3]   [,4]   [,5]   [,6]
[1,] -0.511 -0.066 -0.666  0.803  0.243 -0.259
```

Opcionalmente, podemos obter a matriz de correlação:

```
correl <- cov2cor(covar)
round(correl,2)
[,1] [,2] [,3] [,4] [,5] [,6]
[1,] 1.00 0.42 0.18 0.17 0.42 0.08
[2,] 0.42 1.00 0.37 0.23 0.41 0.10
[3,] 0.18 0.37 1.00 0.38 0.23 0.17
[4,] 0.17 0.23 0.38 1.00 0.34 0.45
[5,] 0.42 0.41 0.23 0.34 1.00 0.15
[6,] 0.08 0.10 0.17 0.45 0.15 1.00
```

Na matriz de correlação, podemos notar que a correlação entre os nós 1 e 6, as áreas mais distantes no grafo, é a menor de todas. A maior correlação é aquela entre os nós 4 e 6. Pela Figura 3.1, vemos que o nó 4 é o único vizinho do nó 6. Assunção e Krainski (2009) explicam o comportamento da matriz de correlação em função da estrutura do grafo.

Vamos simular alguns conjuntos de dados considerando diferentes valores de ρ e visualizar num mapa. Vamos usar o mapa do estado de Minas Gerais dividido em 853 municípios de acordo com a divisão municipal existente em 2001. O mapa está disponível em formato *shapefile* no site do IBGE, em ftp://geoftp.ibge.gov.br/mapas/malhas_digitais/municipio_2001/MG/. O conjunto de três arquivos necessários são: *31mu2500g.shx*, *31mu2500g.shp* e *31mu2500g.dbf*. O código 31 no início do nome de cada arquivo refere-se ao código do IBGE para o estado de Minas Gerais. O número 2500 refere-se á escala, de 1:2.500.000 e a projeção cartográfica adotada é a policônica.

Como vamos simular mais de um conjunto de dados, vamos criar uma função para simular dados da distribuição CAR a partir de uma lista de vizinhança (argumento *nb*) e dos parâmetros do modelo.

```
rcar <- function(nb, rho, sd=1, mean=0) {
  A <- nb2mat(nb, style="B")
  prec <- (diag(rowSums(A)) - rho*A)/(sd^2)
  L <- chol(solve(prec))
  return(drop(rnorm(length(nb))%*%L) + mean)
}
```

O próximo passo é ler o shapefile de Minas Gerais e obter a lista de vizinhança a partir dos polígonos que compõem o mapa. Para isto, usamos as funções `readShapePoly()` e `poly2nb()`, respectivamente, ambas do pacote **spdep** (que deve ser carregado antes):

```
mg <- readShapePoly("31mu2500g")
mg.nb <- poly2nb(mg)
```

A partir da lista de vizinhança, vamos usar usar a função `rcar` definida anteriormente e simular dados da distribuição CAR, considerando quatro valores de ρ : -0.9, 0, 0.75 e 0.99. Ou seja, veremos no mapa dados simulados sob o cenário de forte correlação negativa entre áreas vizinhas, independência entre os valores gerados, e com correlação positiva moderada e forte.

Na Figura 8.1 nós vemos três simulações para cada um desses 4 valores de ρ . Cada coluna da figura tem um valor diferente de ρ . Observamos que na primeira coluna os mapas contêm uma mistura de áreas claras circundadas por áreas escuras, um cenário de correlação espacial negativa. No outro extremo, a quarta coluna, com $\rho = 0.95$, mostra que áreas próximas tendem a possuir cores iguais.

8.2 Os outros auto-modelos

Para simular dados dos demais modelos apresentados nos capítulos anteriores, vamos utilizar o algoritmo amostrador de Gibbs embora outros algoritmos possam também ser utilizados (ver Cressie, pág. 569, para mais detalhes).

Nós geramos uma configuração inicial atribuindo valores independentes a cada sítio. Estes valores devem fazer parte do suporte da distribuição marginal de cada variável. Por exemplo, devem ser iguais a 0 ou 1 no caso do modelo autologístico ou devem ser inteiros não-negativos no caso de uma Poisson. A partir desse conjunto de dados inicial, arbitrariamente escolhido e sem

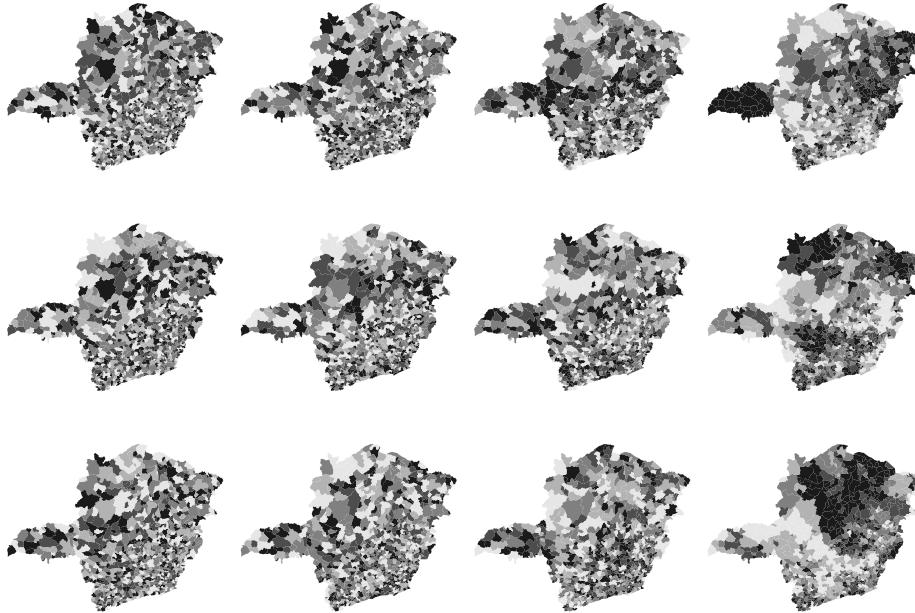


Figura 8.1: Mapa de dados simulados com $\rho = -0.9$ (primeira coluna), $\rho = 0$ (segunda coluna), $\rho = 0.75$ (terceira coluna) e $\rho = 0.95$ (quarta coluna).

dependência espacial alguma, atualizamos iterativamente o valor de cada x_i segundo a distribuição condicional $f(X_i|\mathbf{x}_{-i})$, usando os seguintes passos:

- para b de 1 até B faça
 - para i de 1 até n faça
 - * encontre $f(X_i|\mathbf{x}_{-i})$ usando os valores atuais de \mathbf{x}_{-i}
 - * simule x_{i*} de $f(X_i|\mathbf{x}_{-i})$
 - * atualize x_i fazendo $x_i = x_{i*}$

Ao final de B iterações, com B suficientemente grande, \mathbf{x} é aproximadamente uma configuração gerada pela distribuição conjunta determinada pelo conjunto de distribuições condicionais $f(X_i|\mathbf{x}_i)$.

Dados binários são bastante comuns em grades regulares, tais como imagens preto e branco e dados de presença de doença em plantações comerciais (que são regularmente espaçadas). Dados contínuos e de contagem geralmente são mais comuns em grades irregulares tais como o número de casos de doença por município, número de crimes por município, PIB dos municípios, IDH municipal, etc.

O modelo de Ising ou autologístico é usado em modelos para dados binários. Quando eles estão situados em uma estrutura de grade regular, podemos atualizar o status do dado simulado em blocos, usando o método de codificação proposto por Julian Besag (Besag, 1972; Besag, 1974). Esta facilidade de simulação em blocos é difícil de ser feita em grade irregulares. Portanto, vamos dividir esta seção em duas partes, uma para simular dos modelos de Ising e autologístico em grades regulares e a outra para simular dos modelos auto-normal e de Poisson e grades irregulares.

8.3 Modelo autologístico e de Ising

No Capítulo 1, nós identificamos uma imagem quadrada com uma matriz $N \times M$ onde cada célula da matriz continha um valor numérico indicando sua cor. Para simplificar a notação, ao invés da denotar um sítio genérico como (i, j) , vamos assumir que as células foram ordenadas de 1 até $n = NM$ e denotar um elemento da matriz por um único inteiro genérico i (ou, às vezes, j) entre 1 e n .

Inicialmente vamos detalhar o procedimento de *coding* proposto por Besag (1972, 1974). Observando a figura 8.3, notamos que o lattice regular está dividido meio a meio entre pontos e asteriscos. Considere a estrutura de vizinhança do movimento de peão descrito na Figura (3.5). No esquema de codificação adotado na Figura 8.3, os vizinhos de primeira ordem dos pontos são asteriscos, e vice-versa. Com esta estrutura de vizinhança, é fácil ver que, dados os valores nas posições marcadas por $*$, os valores nas posições marcadas por pontos são variáveis aleatórias independentes e podem ser geradas de uma única vez como um vetor. O mesmo vale para a geração dos valores nas posições marcadas por $*$ condicionalmente me todos os valores nas posições marcadas por pontos.

Dessa forma, usando o esquema de codificação de Besag, podemos simpli-

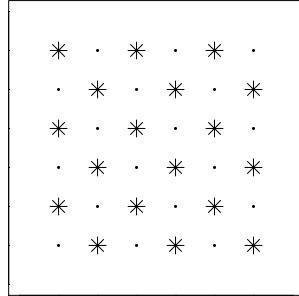


Figura 8.2: Sistema de codificação usado para atualizar o status de x .

ficar o algoritmo amostrador de Gibbs da seguinte forma:

- para b de 1 até B , faça:
 - alternadamente para cada um dos dois conjuntos de nós do grid (pontos ou asteriscos), faça:
 - * encontre $f(X_i|\mathbf{x}_{-i})$ usando os valores de \mathbf{x}_{-i} atuais
 - * simule x_{i*} de $f(X_i|\mathbf{x}_{-i})$
 - * atualize x_i fazendo $x_i = x_{i*}$

Conforme vimos no Capítulo 1, a probabilidade p_i do píxel i no interior da imagem ser preto é uma função no número de pixels vizinhos pretos e do parâmetro β . Ou seja,

$$p_i = \mathbb{P}(X_i = 1 | x_j \text{ com } j \in \mathcal{N}_i) = \frac{\exp(\beta s_i)}{1 + \exp(\beta s_i)}$$

onde s_i é a diferença do número de vizinhos pretos menos o número de vizinhos brancos.

Com essa expressão para p_i , podemos implementar a seguinte função para simular do modelo de Ising:

```
rising <- function(n, beta, burnin=10) {
  ## n - dimensao do lattice (quadrado)
  ## beta - parametro do modelo de Ising
  ## burnin - numero de simulacoes iniciais descartadas
```

```

## x - valores iniciais {-1,1} simulados com
## probabilidade 0.5 para cada 1
x <- matrix(2*(0.5>runif(n*n))-1, n)
## cod1 - indice das linhas e colunas da matriz
## cujo codigo e' ponto
cod1 <- seq(2, n, 2)
## cod2 - indice das linhas e colunas da matriz
## cujo codigo e' asterisco
cod2 <- seq(1, n, 2)
## loop para cada iteracao do algoritmo amostrador
## de Gibbs
for (i in 1:burnin) {
  ## soma os vizinhos
  s <- cbind(x[,-1], 0) + cbind(0, x[,-n]) +
    rbind(x[-1,], 0) + rbind(0, x[-n,])
  ## calcula exponencial
  e <- exp(2*beta*s)
  ## atualiza localizacoes codificadas como pontos
  x[cod1, cod2] <- 2*(e[cod1,cod2]/(1+e[cod1,cod2]))
  >runif(n*n/4))-1
  x[cod2, cod1] <- 2*(e[cod2,cod1]/(1+e[cod2,cod1]))
  >runif(n*n/4))-1
  ## atualiza a soma dos vizinhos
  s <- cbind(x[,-1], 0) + cbind(0, x[,-n]) +
    rbind(x[-1,], 0) + rbind(0, x[-n,])
  e <- exp(2*beta*s)
  ## atualiza localizacoes codificadas como asteriscos
  x[cod2, cod2] <- 2*(e[cod2,cod2]/(1+e[cod2,cod2]))
  >runif(n*n/4))-1
  x[cod1, cod1] <- 2*(e[cod1,cod1]/(1+e[cod1,cod1]))
  >runif(n*n/4))-1
}
return(x)
}

```

Na figura 8.3 temos as imagens geradas para quatro conjunto de dados simulados com essa função, considerando diferentes valores de β . Nessa figura, os pontos pretos correspondem à $x_i = 1$ e os pontos brancos a $x_i = 0$.

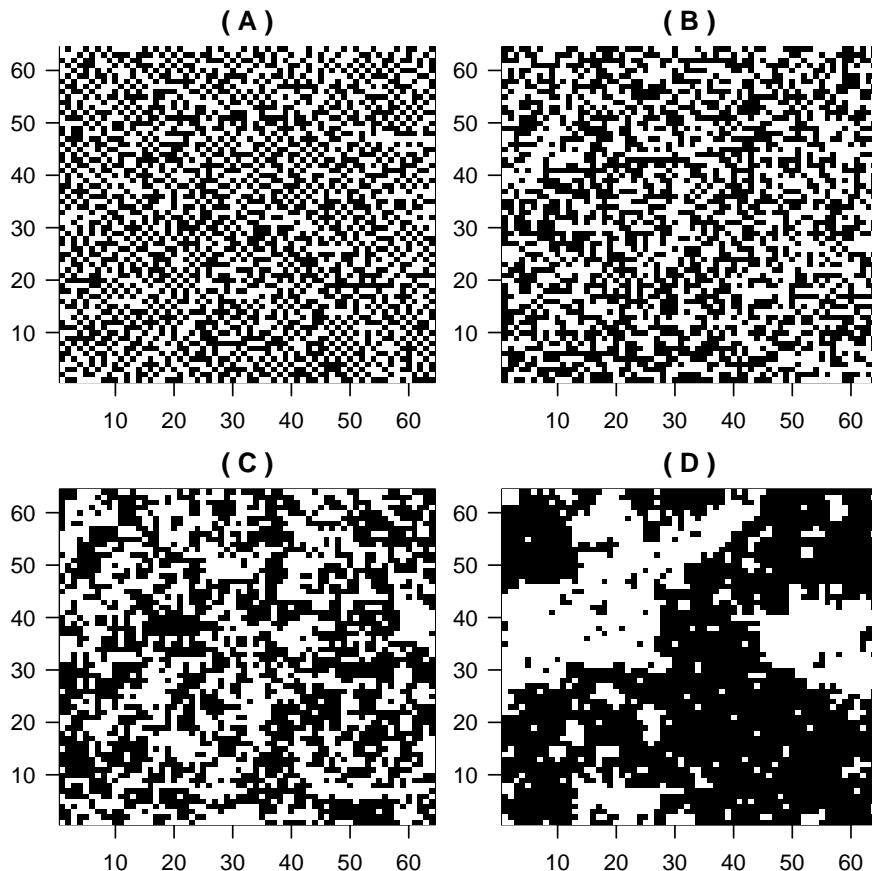


Figura 8.3: Dados simulados do modelo de Ising, com $\beta = -0.3$ (A), $\beta = 0$ (B), $\beta = 0.25$ (C) e $\rho = 0.45$ (D).

O modelo Autologístico é equivalente ao modelo de Ising e sua simulação é quase idêntica. Basta trocar as expressões para p_i e simular zeros e uns ao

invés de -1 's e $+1$'s. A probabilidade p_{ij} no modelo autologístico é dado por

$$p_i = \frac{\exp(x_i(\alpha_i + \beta S_i))}{1 + \exp(\alpha_i + \beta S_i)}$$

com $x_i \in \{0, 1\}$ e $S_i = \sum_{j \sim i} x_j$ é o número de vizinhos do nó i que são iguais a 1. No modelo autologístico, α_i é um parâmetro associado o percentual de valores em x que são iguais a 1 quando não há vizinhos iguais a 1, $P(X_i = 1|S_i = 0)$, podendo ser um parâmetro associado à incidência ou podendo ser associado com covariáveis. Aqui, nós vamos considerar esse parâmetro constante em todo o lattice, ou seja, $\alpha_i \equiv \alpha$.

A função pode ser definida então por:

```
rautologicistic <- function(n, alpha, beta, burnin=10) {
  ## n - dimensao do lattice (quadrado)
  ## alpha - parametro de incidencia
  ## beta - parametro do modelo de autologistico
  ## burnin - numero de simulacoes iniciais descartadas

  ## x - valores iniciais {0,1} simulados
  ## com probabilidade em funcao de alpha
  x <- matrix(rbinom(n*n, 1, exp(alpha)/(1+exp(alpha))), n)
  ## cod1 - indice das linhas e colunas da matriz
  ## cujo codigo e' ponto
  cod1 <- seq(2, n, 2)
  ## cod2 - indice das linhas e colunas da matriz
  ## cujo codigo e' asterisco
  cod2 <- seq(1, n, 2)
  ## loop para cada iteracao do algoritmo amostrador de Gibbs
  for (i in 1:burnin) {
    ## soma os vizinhos
    s <- cbind(x[,-1], 0) + cbind(0, x[,-n]) +
      rbind(x[-1,], 0) + rbind(0, x[-n,])
    ## calcula probabilidade
    pred <- exp(alpha + beta*s)
    prob <- pred/(1+pred)
    ## atualiza localizacoes codificadas
```

```

## como pontos
x[cod1, cod2] <- rbinom(n*n/4, 1,
prob[cod1,cod2])
x[cod2, cod1] <- rbinom(n*n/4, 1,
prob[cod2,cod1])
## atualiza a soma dos vizinhos
s <- cbind(x[,-1], 0) + cbind(0, x[,-n]) +
rbind(x[-1,], 0) + rbind(0, x[-n,])
## atualiza probabilidade
pred <- exp(alpha + beta*s)
prob <- pred/(1+pred)
## atualiza localizacoes codificadas como asteriscos
x[cod2, cod2] <- rbinom(n*n/4, 1, prob[cod2,cod2])
x[cod1, cod1] <- rbinom(n*n/4, 1, prob[cod1,cod1])
}
return(x)
}

```

Na Figura 8.4 temos algumas figuras de dados simulados do modelo autologístico, combinando os valores de α , (-3, -2 e -1), e β , (-0.5, 0 e 0.5). Nessa figura, os pontos pretos correspondem à $x_i = 1$ e os pontos brancos a $x_i = 0$.

Como podemos observar, as imagens dos dados simulados com $\alpha = -3$ são as que apresentam o menor número de pontos pretos. Neste caso, a probabilidade condicional de um pixel ser preto é dada por $\exp(-3 + \beta * s)/(1 + \exp(-3 + \beta * s))$ onde s é o número de vizinhos. Se $\beta = 0$, a probabilidade é igual a 0.05, aproximadamente. Essa probabilidade aumenta se β aumenta. Por isso, a imagem mais inferior à direita apresenta o maior número de pontos pretos.

8.4 Simulando dos modelos ICAR e autoPoisson

Agora nós vamos definir funções para simular dos modelos CAR intrínseco e autoPoisson considerando grades irregulares. O modelo de Ising e o modelo autologístico também podem ser usados neste contexto, assim como esses modelos também podem ser usados no contexto de grades regulares.

No modelo ICAR não temos uma distribuição conjunta porque a matriz de precisão deste modelo não é invertível. Porém, este modelo é dado por um conjunto de distribuições condicionais:

$$(y_i - \mu_i) | \mathbf{y}_{-i} \sim N(\bar{y}_i, \sigma^2/n_i)$$

onde n_i é o número de vizinhos de i e $\bar{y}_i = \sum_{j \sim i} y_j / n_i$ é a média de seus vizinhos. Portanto, se $\mu_i = 0$, $i = 1, 2, \dots, n$, a média de y_i é a média dos seus vizinhos e sua variância é inversamente proporcional ao número de vizinhos.

Não é correto usarmos estas distribuições condicionais no amostrador de Gibbs pois não existe distribuição conjunta e portanto não existe distribuição estacionária para a cadeia de Markov do amostrador. Entretanto, se introduzirmos uma restrição, forçando ser zero a soma dos y_i 's ao final de cada ciclo Gibbs, teremos uma distribuição normal multivariada de dimensão $n - 1$ num vetor de dimensão n .

A seguir, definimos uma função cujos argumentos são uma lista de vizinhança, a média, o desvio-padrão e o número de simulações iniciais descartadas:

```
ricar <- function(nb, mean=0, sd=1, burnin=10)
{
  n <- length(nb)
  ni <- sapply(nb, length)
  ni.s <- sqrt(ni)
  x <- rnorm(n)
  for (k in 1:burnin)
    for (i in 1:n) {
      x[i] <- rnorm(1, mean(x[nb[[i]]]), ni.s[i])
      x <- x - mean(x)
    }
  return(x + mean)
}
```

Usando esta função e o mapa do estado de Minas Gerais subdividido em municípios, nós geramos duas amostras da distribuição ICAR e plotamos na Figura 8.5. Observamos nessa figura que há regiões com vários municípios próximos e com valores parecidos.

Para simular do modelo autoPoisson, consideramos que

$$\mu_i = \exp \left(\alpha + \beta \sum_{j \neq i} x_j \right)$$

ou seja, α e β são constantes. A função para simular do modelo autoPoisson pode ser da forma:

```
rautopoisson <- function(nb, alpha, beta, burnin=10) {
  if (beta>0)
    stop("'beta' must be non positive in auto-poisson!")
  n <- length(nb)
  x <- rpois(n, exp(alpha))
  for (k in 1:burnin)
    for (i in 1:n) {
      x[i] <- rpois(1, exp(alpha + beta*x[nb[[i]]]))
    }
  return(x)
}
```

A partir dessa função, simulamos 4 amostras do modelo autoPoisson. Nós consideramos $\alpha = 3$ e os valores $0, -0.05, -0.1$ e -0.2 para β . Os quatro mapas para visualizar essas amostras estão na Figura 8.6. Nós usamos uma mesma escala de cores para todos os quatro mapas. A variância dos dados gerados aumenta à medida que o parâmetro β se afasta de zero. Além disso, observamos um padrão de repulsão espacial muito nítido quando $\beta = -0.2$.

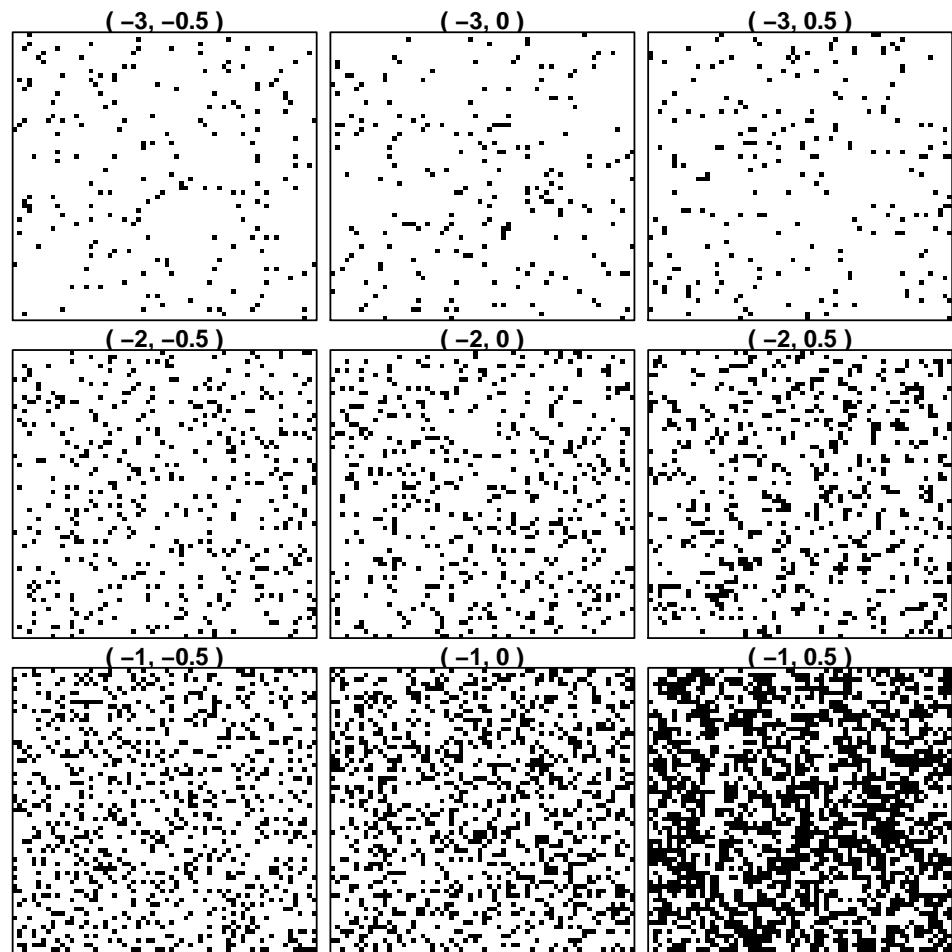


Figura 8.4: Dados simulados do modelo autologistico, com valores de (α, β) indicados entre parênteses.

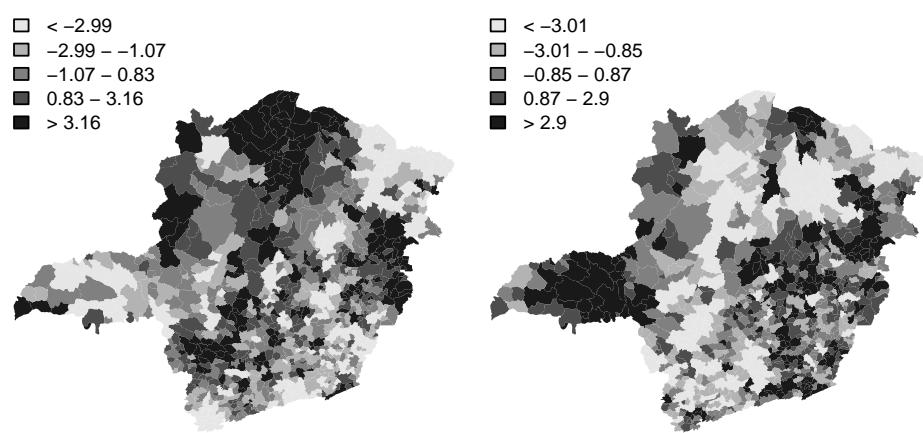


Figura 8.5: Mapas de duas amostras da distribuição ICAR.

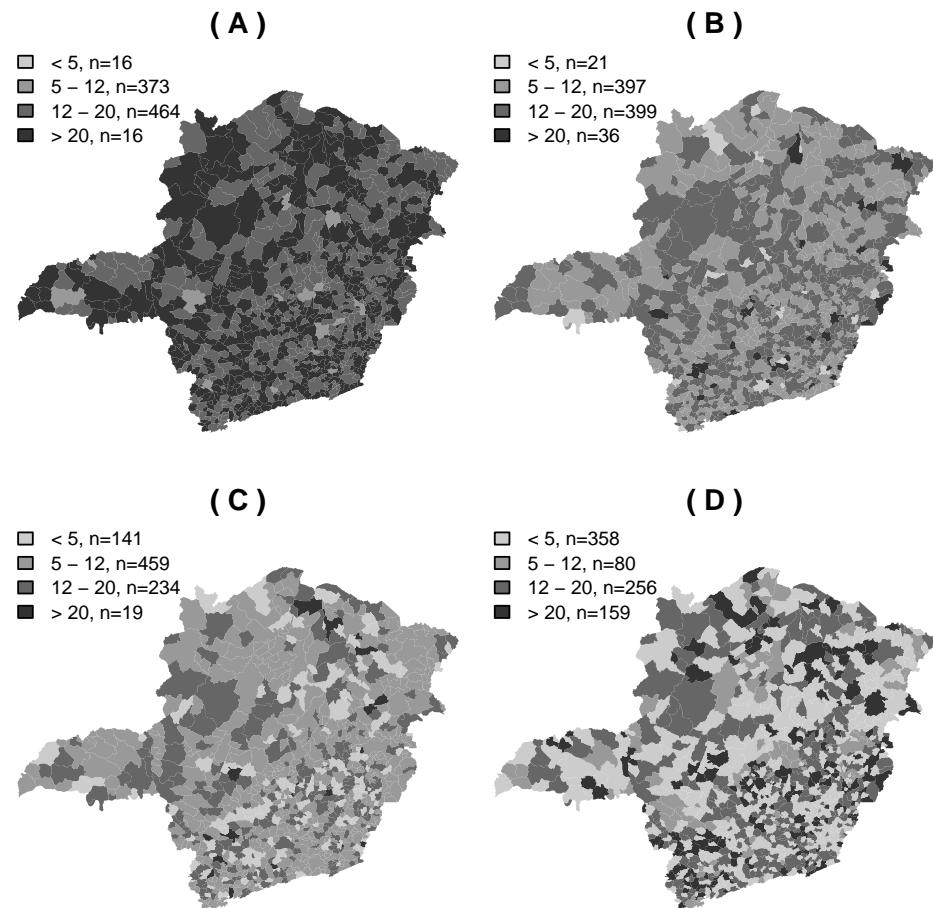


Figura 8.6: Mapas de quatro amostras do modelo autoPoisson com $\alpha = 3$ e $\beta = 0$ (A), $\beta = -0.05$ (B), $\beta = -0.1$ (C) e $\beta = -0.2$ (D).

Capítulo 9

Estimação de Auto-modelos

9.1 Introdução

Os automodelos são semelhantes a um modelo linear generalizado. Porém, nos modelos lineares generalizados as observações são independentes e portanto a densidade conjunta é simplesmente o produto das densidades marginais $f(x_i|\theta)$ para $i = 1, 2, \dots, n$. No caso dos Campos de Markov, não é muito simples estimar os parâmetros diretamente pois a constante de integração não é conhecida.

Por exemplo, num modelo de Ising simples, temos a seguinte densidade conjunta:

$$f(\mathbf{x}; \theta) = \frac{1}{Z(\beta)} \exp \left(-\beta \sum_{i \sim j} x_i x_j \right).$$

A constante de integração $Z(\beta)$ depende do parâmetro β e, em geral, não é conhecida de forma analítica. Portanto, métodos usuais de maximização da versossimilhança não vão funcionar neste problema.

Para grades regulares, Besag (1972) sugeriu usar o sistema de coding e estimar dois modelos, cada um considerando o subconjunto de observações definido por um dos códigos como variável resposta condicionalmente nos valores dos demais sítios que são usados no cálculo de vizinhos similares (ver Figura 8.3). Assim, temos um vetor \mathbf{y} de dados observados com apenas $n/2$ das observações do vetor original \mathbf{x} e um segundo vetor s do mesmo tamanho

de \mathbf{y} e com a soma dos x_j 's na vizinhança de cada observação em \mathbf{x} .

Neste procedimento, podemos usar a teoria usual de verossimilhança já que o subconjunto de dados selecionados por cada um dos códigos do sistema de coding são condicionalmente independentes. Assim, basta usar um programa de regressão logística para estimar os parâmetros do modelo autologístico. Os coeficientes estimados são assintoticamente eficientes e sua matriz de covariância é corretamente estimada.

Algumas pessoas estimam dois modelos, para cada um dos dois subconjuntos definidos pelo sistema de coding, e combinam por uma média aritmética as estimativas de ambos os modelos para fazer inferência. Note entretanto que as estimativas de variância dos dois subconjuntos são diferentes e portanto não saberemos como calcular um desvio-padrão para a média resultante.

Um outro método engenhoso, também sugerido por Besag (1975) é imitar a verossimilhança usual de dados independentes usando um produto das densidades condicionais. Ele sugeriu trabalhar com a função

$$L(\theta|x) = \prod_{i=1}^n f(x_i|\theta, \mathbf{x}_i) \quad (9.1.1)$$

como se fosse a função de verossimilhança verdadeira e usar um modelo linear generalizado usual para estimar os parâmetros. No caso de automodelos, a função (9.1.1) não é a verdadeira verossimilhança do modelo e por isso esse método é chamado de pseudo-verossimilhança (PL).

O método de PL é razoável para estimar os parâmetros, mas subestima a variância das estimativas (Gumpertz *et al.*, 1997) podendo conduzir a conclusões erradas. Uma alternativa seria estimar a variância das estimativas utilizando bootstrap. Porém, no contexto de dados com dependência espacial, este procedimento bootstrap não é correto pois as reamostras não vão preservar o padrão espacial dos dados originais. Um solução é fazer uma reamostragem por blocos (ver Cressie, 1993). Outra solução, que adotaremos aqui, é usar a distribuição condicional de cada observação e usar o algoritmo amostrador de Gibbs para obter as reamostras (ver Gumpertz *et al.*, 1997).

A seguir, vamos mostrar exemplos de estimação dos parâmetros do modelo CAR e do modelo autologístico.

9.2 Modelo CAR

A estimativa dos parâmetros do modelo CAR pode ser feita considerando o fato de que a distribuição conjunta é normal multivariada. Podemos escrever a densidade (ou log-densidade) da distribuição normal multivariada e maximizá-la em relação aos parâmetros. Tomamos a distribuição conjunta dada por (5.1.2) e maximizamos esta função em relação aos parâmetros ρ e σ^2 , considerando média conhecida, ou em relação a esses dois parâmetros mais o parâmetro de média (ou um parâmetro β de efeito de covariáveis).

Em **R** há várias funções de otimização disponíveis que podemos utilizar para a estimativa por máxima verossimilhança. Vamos exemplificar a estimativa dos parâmetros do modelo CAR usando a função `optim()`. Nesta função estão implementados os algoritmos de Nelder Mead, Gradiente Conjugado e alguns do tipo quasi-Newton.

O procedimento de maximização numérica de funções de verossimilhança é usualmente feito considerando o logaritmo da função, para que haja estabilidade numérica. O *default* da função `optim()` é encontrar os valores dos parâmetros que minimizam uma função. Portanto, inicialmente vamos definir uma função para a obtenção do negativo da log-verossimilhança da função de verossimilhança do modelo CAR, ou seja, do negativo do log da densidade da distribuição normal multivariada definida por esse modelo. Em (5.1.2), nós vamos considerar que a média também é desconhecida (e constante) e vamos estimá-la a partir dos dados.

A função `optim()` requer essencialmente os valores iniciais para o processo de otimização e a função objetivo. Esta função deve ser definida de forma que os argumentos em relação aos quais a função será otimizada sejam informados no primeiro argumento. Nós vamos definir a função `negloglikfun()` a seguir para usá-la na função `optim()`. A função `negloglikfun()` tem por argumentos: `pars` - um vetor com os valores dos parâmetros, `y` - o vetor de dados, `w` - a matriz de vizinhança padronizada por linhas e `d` - um vetor com o número de áreas vizinhas de cada área. Como o número de parâmetros do modelo é três, o argumento `pars` deve ter três elementos. Nos cálculos internos da função, nós vamos considerar que o valor da primeira posição é ρ , o valor na segunda posição é σ^2 e o valor da terceira posição é μ , o parâmetro de média.

```
### função para cálculo do
```

```
### negativo do log-verossimilhanca
negloglikfun <- function(pars, y, w, d) {
  ### diferenca entre y e a media
  z <- y-pars[3]
  ### calculo da matriz de precisao
  q <- (diag(length(y)) - pars[1]*w)*d/pars[2]
  ### calculo do negativo do log da densidade
  return(.5*(length(y)*log(2*pi) -
  determinant(q)$mod + z\%*\%q\%*\%z))
}
```

No exemplo, nós vamos usar o mapa de MG, carregado conforme os comandos a seguir:

```
### carrega pacote maptools para ler o mapa de MG
require(maptools)

### carrega mapa dos municipios de MG
mg <- readShapePoly("31mu2500g")

### carrega o pacote spdep para obter a lista de vizinhanca
require(spdep)

### obtém lista de vizinhanca
mg.nb <- poly2nb(mg)

### obtém o numero de vizinhos de cada municipio
mg.d <- sapply(mg.nb, length)

### obtém a matriz de vizinhança padronizada por linha
mg.w <- nb2mat(mg.nb)
```

Usando a função `rcar()` definida no capítulo anterior, nós simulamos um conjunto de dados, com $\mu = 0$ (o parâmetro de média), $\sigma^2 = 1$ e $\rho = 0.7$, conforme os comandos a seguir:

```
### simula um conjunto de dados
```

```
set.seed(1)
y <- rcar(mg.nb, 0.7)
```

No primeiro argumento da função `optim()` informamos os valores iniciais dos parâmetros e no segundo passamos a função objetivo. A seguir devemos escolher os valores dos demais argumentos da função `optim()` e passar os demais argumentos da função objetivo nomeando cada um. Como queremos calcular a variância das estimativas, nós colocamos `hessian=TRUE`. Além disso, passamos os demais argumentos da função `negloglikfun()`.

```
### estima os parametros e a matriz hessiana
opt <- optim(c(0, 0.7, 1), negloglikfun, hessian=TRUE,
              y=y, w=mg.w, n=length(mg.nb), d=mg.d)
```

O valor das estimativas obtidas é acessado fazendo:

```
opt$par
[1] 0.5718131 1.0970262 -0.0052717
```

Notamos que as estimativas de μ (terceiro valor) e de σ^2 (segundo valor) são próximos dos valores verdadeiros. O valor da estimativa de ρ porém é razoavelmente menor que o valor verdadeiro.

A seguir vamos obter o erro-padrão das estimativas, usando a matriz hessiana estimada numericamente pela função `optim()`. Para isso, invertemos a matriz hessiana estimada e obtemos a matriz de covariância das estimativas. O erro-padrão das estimativas é, então, a raiz quadrada dos elementos da diagonal. Obtemos estes valores com os comandos a seguir:

```
matv <- solve(opt$hessian)
se <- sqrt(diag(matv))
se
[1] 0.07831677 0.05448442 0.02295065
```

Se considerarmos a propriedade de normalidade assintótica dos estimadores de máxima verossimilhança, podemos obter os intervalos de 95% de confiança, para os três parâmetros com os seguintes comandos:

```
cbind(li=qnorm(0.025, opt$par, se),
```

```

ls=qnorm(0.975, opt$par, se))
      li      ls
[1,] 0.41831500 0.72531110
[2,] 0.99023873 1.20381375
[3,] -0.05025415 0.03971075

```

Esses intervalos contêm os valores verdadeiros do parâmetros.

9.3 Modelo autologístico

Na inferência para os parâmetros do modelo autologístico, nós vamos usar dois procedimentos de estimação: o sistema de coding e a maximização da pseudo-verossimilhança. Para este último, fazemos a estimação da variância das estimativas via bootstrap com amostras geradas usando o algoritmo amostrador de Gibbs. Iniciamos a estimação via PL por ser o método mais simples.

9.3.1 Usando pseudo-verossimilhança

A estimação via pseudo-verossimilhança é o procedimento mais simples porque neste caso basta calcular o número de vizinhos doentes e organizar um `data.frame` de dados contendo duas colunas, uma com os dados e a outra com a soma na vizinhança. Assim, podemos usar a função `glm()` e estimar o modelo logístico usual considerando as observações como variável resposta e a soma na vizinhança como uma covariável.

Abaixo, nós implementamos uma função para o ajuste do modelo usando PL, na qual o argumento `x` é a matriz de dados:

```

fit.pseudo <- function(x) {
  ### numero de linhas da matriz x
  n <- nrow(x)
  ### s - soma na vizinhanca
  s <- cbind(x[,-1], 0) + cbind(0, x[,-n]) +
    rbind(x[-1,], 0) + rbind(0, x[-n,])
  ### organizando x e s num data.frame
  dat <- data.frame(x=as.vector(x), s=as.vector(s))
  ### usando a funcao glm() para estimar os parametros

```

```
    glm(x~s, binomial, dat)
}
```

Simulamos um conjunto de dados binário usando a função definida no capítulo anterior.

```
### simula um conjunto de dados
n <- 50
set.seed(1)
### fazendo alpha = -1 e beta = 0.1
dat <- rautologistic(n, -1, 0.1)
```

A seguir, usamos a função definida anteriormente para obter as estimativas de α e β por PL.

```
pseudo <- fit.pseudo(dat)
summary(pseudo)$coef
  Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.13810   0.07057 -16.128 < 2e-16
s            0.16040   0.04741    3.383 0.000717
```

A primeira estimativa, (Intercept), corresponde ao parâmetro α , e a segunda estimativa, s, corresponde ao parâmetro β . O valor da estimativa de α está razoavelmente próximo do valor verdadeiro mas o valor da estimativa de β é um pouco maior que o valor verdadeiro.

Como a variância estimada usando PL não é confiável, não devemos nos basear nos Std. Error's da saída acima para calcular intervalos de confiança. Mais a frente, nós estimamos a variância da estimativa de α e β usando bootstrap com as reamostras obtidas via amostrador de Gibbs.

9.3.2 Estimação via sistema de coding

A estimação usando o sistema de coding consiste basicamente em selecionar os dados de um dos subconjuntos definido por um dos códigos do sistema de coding e utilizá-lo como resposta (ver Figura 8.3). As demais observações são consideradas para o cálculo da covariável de soma dos vizinhos. Usando um dos subconjuntos, basta simplesmente organizar os dados num `data.frame` e usar a função `glm()` para obter as estimativas dos parâmetros.

A função a seguir implementa essa opção de estimação, onde x é a matriz de dados e `codtype` define qual subconjunto será utilizado como resposta.

```
### estimativas via coding
fit.coding <- function(x, codtype=1) {
  n <- nrow(x)
  s.nb <- cbind(x[,-1], 0) + cbind(0, x[,-n]) +
    rbind(x[-1,], 0) + rbind(0, x[-n,])
  ## cod1 - indice das linhas e colunas da
  ## matriz cujo codigo e' ponto
  cod1 <- seq(2, n, 2)
  ## cod2 - indice das linhas e colunas da matriz
  ## cujo codigo e' asterisco
  cod2 <- seq(1, n, 2)
  ### define o data.frame de dados
  if (codtype==1)
    x <- data.frame(y=c(x[cod1, cod2], x[cod2,cod1]),
                      s=c(s.nb[cod1, cod2], s.nb[cod2,cod1]))
  else
    x <- data.frame(y=c(x[cod2,cod2], x[cod1,cod1]),
                      s=c(s.nb[cod2,cod2], s.nb[cod1,cod1]))
  ### ajusta o modelo logistico usando a funcao glm()
  return(glm(y~s, binomial, x))
}
```

A seguir nós vamos estimar os parâmetros do modelo autologístico para o conjunto de dados simulados anteriormente. Vamos estimar dois modelos, considerando cada um dos dois subconjuntos definidos pelo sistema de coding como resposta de cada vez.

```
aj1 <- fit.coding(dat)
aj2 <- fit.coding(dat,2)
summary(aj1)$coef
```

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-1.1086520	0.09871656	-11.23066	2.883213e-29
s	0.1558188	0.06709330	2.32242	2.021034e-02

```
summary(aj2)$coef
    Estimate Std. Error     z value     Pr(>|z|)
(Intercept) -1.168898 0.10093446 -11.580765 5.158366e-31
s            0.165810 0.06704635   2.473065 1.339597e-02
```

Notamos que o valor das estimativas é próximo do valor das estimativas obtidas por PL. Aliás, o valor da estimativa obtida por PL está entre o valor das estimativas obtidas com cada um dos subconjuntos de dados, para cada um dos parâmetros α e β . Notamos também que as variâncias estimadas em cada um dos dois sub-modelos é maior que aquela aparecendo na saída do método PL.

Por independência condicional, a verossimilhança utilizada para cada um dos subconjuntos é verdadeira. Portanto, as variâncias estimadas tem as propriedades de modelos lineares generalizados em cada um dos subconjuntos de dados. Porém, cada subconjunto usa apenas metade dos dados, condicionando nos demais, para obter a estimativa. Neste sentido, coding é um método menos eficiente do que desejariam. Não existe uma maneira simples de fazer inferência se combinarmos as estimativas de cada bloco de coding.

9.3.3 Estimação da variância via bootstrap

A estimação da variância via bootstrap foi sugerida por Gumpertz *et al.* (1997). Este procedimento consiste em usar o método de PL para obter as estimativas dos parâmetros. Com o parâmetro estimado como se fosse o verdadeiro, obtemos reamostras usando a distribuição condicional de cada observação dados os valores das observações vizinhas. Para cada reamostra, estima-se os parâmetros via PL. A variância das estimativas é então estimada pelo valor da variância amostral calculada com os valores das estimativas obtidas em cada reamostra.

Podemos resumir esse algoritmo nos seguintes passos:

- estime os parâmetros com os dados observados usando PL
- considere os valores observados como valor inicial de x para o algoritmo amostrador de Gibbs
- para b de 1 até B , faça:

- alternadamente, para cada um dos dois conjuntos de nós do grid (pontos ou asteriscos), faça:
 - * encontre $f(X_i|\mathbf{x}_{-i})$ usando os valores de \mathbf{x}_{-i} atuais e usando as estimativas de PL obtidas a partir dos dados observados
 - * simule x_i^* de $f(X_i|\mathbf{x}_{-i})$
 - * atualize x_i fazendo $x_i = x_i^*$
- estime os parâmetros via PL com a matriz de dados atualizada (reamostra)
- com as B estimativas de cada parâmetro, estime a variância, aplicando a fórmula da variância amostral às B estimativas de cada parâmetro

Para implementar o procedimento de reamostragem via amostrador de Gibbs, nós podemos utilizar o mesmo algoritmo utilizado na implementação da função para simular dados do modelo autologístico, definida no capítulo anterior. A função implementada a seguir, estima os parâmetros do modelo autologístico para cada uma de B reamostras, com B informado no argumento B da função, retornando uma matriz com $B + 1$ linhas e duas colunas. Na primeira linha dessa matriz, são retornados os valores das estimativas por PL para os dados observados. Nas demais linhas, são retornados os valores das estimativas por PL de cada uma das B reamostras.

```
### bootstrap with Gibbs Sampler
fit.boot <- function(x, B) {
  n <- nrow(x)
  cod1 <- seq(2, n, 2)
  cod2 <- seq(1, n, 2)
  res.pseudo <- fit.pseudo(x)
  alpha <- coef(res.pseudo)[1]
  beta <- coef(res.pseudo)[2]
  mat.res <- matrix(0, B+1, 2)
  mat.res[1,] <- c(alpha, beta)
  for (b in 2:(B+1)) {
    ## soma os vizinhos
    s <- cbind(x[,-1], 0) + cbind(0, x[,-n]) +
      rbind(x[-1,], 0) + rbind(0, x[-n,])
    mat.res[b,] <- c(alpha, beta)
  }
}
```

```

## calcula probabilidade
pred <- exp(alpha + beta*s)
prob <- pred/(1+pred)
if (runif(1)>0.5) {
  ## atualiza localizacoes codificadas como pontos
  x[cod1, cod2] <- rbinom(n*n/4, 1, prob[cod1,cod2])
  x[cod2, cod1] <- rbinom(n*n/4, 1, prob[cod2,cod1])
  ## atualiza a soma dos vizinhos
  s <- cbind(x[,-1], 0) + cbind(0, x[,-n]) +
    rbind(x[-1,], 0) + rbind(0, x[-n,])
  ## atualiza probabilidade
  pred <- exp(alpha + beta*s)
  prob <- pred/(1+pred)
  ## atualiza localizacoes codificadas como asteriscos
  x[cod2, cod2] <- rbinom(n*n/4, 1, prob[cod2,cod2])
  x[cod1, cod1] <- rbinom(n*n/4, 1, prob[cod1,cod1])
}
else {
  ## atualiza localizacoes codificadas como asteriscos
  x[cod2, cod2] <- rbinom(n*n/4, 1, prob[cod2,cod2])
  x[cod1, cod1] <- rbinom(n*n/4, 1, prob[cod1,cod1])
  ## atualiza a soma dos vizinhos
  s <- cbind(x[,-1], 0) + cbind(0, x[,-n]) +
    rbind(x[-1,], 0) + rbind(0, x[-n,])
  ## atualiza probabilidade
  pred <- exp(alpha + beta*s)
  prob <- pred/(1+pred)
  ## atualiza localizacoes codificadas como pontos
  x[cod1, cod2] <- rbinom(n*n/4, 1, prob[cod1,cod2])
  x[cod2, cod1] <- rbinom(n*n/4, 1, prob[cod2,cod1])
}
mat.res[b,] <- coef(fit.pseudo(x))
}
return(mat.res)
}

```

A seguir, nós utilizamos a função `boo.res()` para estimar a variância das estimativas dos parâmetros via bootstrap. Usamos 999 reamostras e descartamos as primeiras 99 para o cálculo da variância, tomando as demais de 10 em 10, conforme os comandos dados a seguir:

```
boo.res <- fit.boot(dat, 999)
sd.boot <- apply(boo.res[seq(101,1000,10),], 2, sd)
sd.boot
[1] 0.08597195 0.06647264
```

Os valores obtidos para a variância das estimativas dos parâmetros do modelo são um pouco maiores que aquelas obtidas por PL.

O teste de hipótese para testar a nulidade dos parâmetros pode ser feito considerando que as estimativas dos parâmetros obtida por PL tem distribuição assintótica normal e usando a variância obtida por bootstrap. Portanto podemos calcular o *p*-valor para esse teste de hipóteses com o seguinte comando:

```
round(2*pnorm(abs(boo.res[1,]/sd.boot),lower=F), 4)
[1] 0.0000 0.0158
```

Rejeitamos a hipótese de que $\alpha = 0$ e também que $\beta = 0$, ambos ao nível de 5%.

9.3.4 Pequena avaliação no procedimento de bootstrap

Com as funções implementadas, nós podemos fazer uma pequena comparação do tamanho dos testes realizados. Vamos considerar o teste de hipótese usando as variâncias das estimativas obtidas por PL e as variâncias obtidas por bootstrap.

A seguir, nós simulamos 1000 conjuntos de dados com $\alpha = -1$ e $\beta = 0$. Para cada um, nós obtemos as estimativas dos parâmetros e a respectiva variância usando o método de PL, implementado na função `fit.pseudo()`. Nós fazemos um sumário do modelo e guardamos apenas os valores-*p* do teste de hipótese.

```
s100 <- t(sapply(1:1000, function(i)
  summary(fit.pseudo(rautologistic(n,-1,0)))
  $coef[,4]))
```

A seguir, calculamos a proporção de vezes em que a hipótese de nulidade de cada parâmetro (α e β) foi rejeitada:

```
colMeans(s100<0.05)
(Intercept)      s
1.000        0.168
```

O nível descriptivo do teste de hipótese para β foi cerca de três vezes maior que o nível nominal do teste, pois neste caso a hipótese deveria ser rejeitada 5% das vezes.

Agora, vamos fazer o mesmo procedimento, porém considerando a variância obtida por bootstrap, considerando 200 reamostras. Também vamos guardar apenas o valor-p em cada uma das 1000 simulações.

```
s100b <- t(sapply(1:1000, function(i) {
  r <- fit.boot(rautologistic(n,-1,0), 200)
  2*pnorm(abs(r[1,]/apply(r[-1,], 2, sd)), lower=FALSE)
}))
```

O cálculo da proporção de vezes em que a hipótese foi rejeitada neste caso é dado por:

```
colMeans(s100b<0.05)
[1] 1.000 0.051
```

A proporção de vezes em que a hipótese foi rejeitada é praticamente igual ao nível de significância nominal do teste.

Referências Bibliográficas

- [1] Arnold B.C., and Press S.J. Compatible conditional distributions. *Journal of the American Statistical Association*, 84, 152–156, 1989.
- [2] Arnold B., Castillo, E. and Sarabia, J.M. *Conditional Specification of Statistical Models*. Springer, University of California, 1999.
- [3] Assunção, R. M. and Krainski, E. T.. Neighboorhood dependence in Bayesian spatial models. *Biometrical Journal*, 5, 851–869, 2009.
- [4] Banerjee, S., Carlin, B. P. and Gelfand, A. E.. *Hierarchical Modeling and Analysis for Spatial Data*. CRC Press, Florida, 2004.
- [5] Besag, J. Nearest-Neighbour Systems and the Auto-logistic Model for Binary Data. *Journal of the Royal Statistical Society, Ser. B*, 34, 75–83, 1972.
- [6] Besag, J. Statistical Analysis of Non-lattice Data. *The Statistician*, 24, 179–195, 1975.
- [7] Besag, J. Spatial interaction and the statistical analysis of lattice systems (with discussion). *Journal of the Royal Statistical Society, Series B*, 36, 192–236, 1974.
- [8] Besag, J., York, J. and Mollié, A. Bayesian image restoration, with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics*, 43, 1–20, 1991.
- [9] Bowman, K.O., Hutcheson, K., Odum, E.P., and Schenton, L.R.. Comments on the distributions of indices of diversity. In *Patil, G.P., Pielou*,

- E.C., and Waters, W.E., eds., *Statistical Ecology*, Vol. 3, Pennsylvania State University Press, University Park, PA, 315–366, 1971.
- [10] Brook, D. On the distinction between the conditional probability and the joint probability approaches in the specification of nearest-neighbour systems. *Biometrika*, 51, 481–483, 1964.
 - [11] Brush S.G. History of the Lenz-Ising model. *Reviews of Modern Physics*, 39, 883–893, 2005.
 - [12] Clifford, P. Markov random fields in statistics. In G.R. Grimmett and D.J.A. Welsh eds., *Disorder in Physical Systems*, J.M. Hammersley Festschrift, Oxford University Press, 19–32, 1990.
 - [13] Coleman J.S. *The Adolescent Society: The Social Life of the Teenager and its Impact on Education*. The Free Press of Glencoe, Glencoe, 1961.
 - [14] Coleman J.S., Katz E., and Menzel H. *Medical Innovation: A Diffusion Study*. Bobbs-Merrill Co, Indianapolis, 1966.
 - [15] Cressie, N. A. C. *Statistics for Spatial Data*. Wiley Series in Probability and Statistics (Hardcover), 1993.
 - [16] Grimmett, G.R. A theorem about random fields. *Bulletin of the London Mathematical Society*, 5, 81–84, 1973.
 - [17] Gumpertz, M.L., Graham, J.M., and Ristaino, J.B. Autologistic Model of Spatial Pattern of Phytophthora Epidemic in Bell Pepper: Effects of Soil Variables on Disease Presence *Journal of Agricultural, Biological, and Environmental Statistics*, 2, 131–156, 1995.
 - [18] Kotz S. and Johnson, N.L. *Breakthroughs in Statistics: Volume 2: Methodology and Distribution*. Springer Series in Statistics, 1993.
 - [19] Kotz S. and Johnson, N.L. *Breakthroughs in Statistics: Volume 1: Foundations and Basic Theory*. Springer Series in Statistics, 1993.
 - [20] Leroux, B.G., Lei, X. and Breslow, N. Estimation of disease rates in small areas: A new mixed model for spatial dependence. In *Statistical Models in Epidemiology; the Environment and Clinical Trials*, 179–192, 1999.

- [21] Moussouris J. Gibbs and Markov random systems with constraints. *Journal of Statistical Physics*, 10, 11–33, 1974.
- [22] Preston, C.J. Generalised Gibbs states and Markov random fields. *Advances in Applied Probability*, 5, 242–261, 1973.
- [23] R (2009), R Development Core Team (2009). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL <http://www.R-project.org>.
- [24] R (2010), Roger Bivand, with contributions by Luc Anselin, Renato Assunção, Olaf Berke, Andrew Bernat, Eric Blankmeyer, Marilia Carvalho, Yongwan Chun, Bjarke Christensen, Carsten Dormann, Stéphane Dray, Rein Halbersma, Elias Krainski, Nicholas Lewin-Koh, Hongfei Li, Jielai Ma, Giovanni Millo, Werner Mueller, Hisaji Ono, Pedro Peres-Neto, Gi-anfranco Piras, Markus Reder, Michael Tiefelsdorf and and Danlin Yu. (2010). spdep: Spatial dependence: weighting schemes, statistics and models. R package version 0.4-58. <http://CRAN.R-project.org/package=spdep>
- [25] Rue, H. and Held, L. *Gaussian Markov random fields: theory and applications*. CRC Press, 2005.
- [26] Sherman S. Markov random fields and Gibbs random fields. *Israeli Journal of Mathematics*, 14, 92–103, 1973.
- [27] Spitzer F. Markov random fields and Gibbs ensembles. *American Mathematical Monthly*, 78, 142–154, 1971.
- [28] Wilson, J.R.. *Introduction to Graph Theory*. Longman, University of California, 1997.

Sobre os autores

Renato Martins Assunção é professor titular do Departamento de Estatística da Universidade Federal de Minas Gerais desde 1988. Ele recebeu seu título de Bacharel em matemática pela UFMG (1984), mestre pelo IMPA-Instituto de Matemática Pura e Aplicada (1987) em matemática aplicada (ênfase em estatística), e PhD em estatística pela University of Washington, Seattle, EUA (1994). É Bolsista de Produtividade em Pesquisa do CNPq, Nível 1D. Ele tem trabalhado em problemas de estatística espacial, a modelagem de dados estocásticos que implicam de alguma forma no uso de sua localização geográfica, e na análise de risco atuarial. É autor de mais de 30 artigos publicados em revistas indexadas e possui mais de 290 citações de seus trabalhos, de acordo com o Web of Science. Como pesquisador visitante, trabalhou ou fez conferências nas universidades americanas de Princeton, Harvard, Texas, Connecticut, Utah, Columbia, no instituto SAMSI e nas universidades européias de Lancaster (Inglaterra), Imperial College, Valencia, Ludwig-Maximilians (Alemanha) e Florença (Itália), além de várias outras no país. Na parte de extensão, é o principal responsável por vários contratos de prestação de serviços da UFMG com empresas públicas e privadas para análise de dados, consultoria ou desenvolvimento de software.

Erica é aluna do Mestrado em Estatística da Universidade Federal de Minas Gerais, sob a orientação do primeiro autor do livro. Recebeu o título de Bacharel em Estatística também pela UFMG no segundo semestre de 2009. Já fez alguns trabalhos na área de Processos Estocásticos Espaciais e atualmente trabalha com análise Bayesiana de dados de área. Seu trabalho atual, que terá como fruto sua dissertação de mestrado e um artigo recentemente submetido, está diretamente relacionado a um dos tópicos deste livro, os campos

gaussianos de Markov.

Elias é professor assistente do Departamento de Estatística da Universidade Federal do Paraná desde agosto de 2009. Recebeu seu título de Bacharel em Estatística pela UFPR (2006) e Mestre em Estatística pela UFMG (2008). Ele tem trabalhado com modelos estatísticos com componente estocástica dependente de localização geográfica. Publicou três artigos com aplicações de alguns desses modelos para análise do padrão espacial de doenças em plantas em revistas aplicadas e um artigo junto com o primeiro autor deste livro analisando a influência da estrutura de vizinhança na correlação espacial de áreas vizinhas quando os dados são modelados via modelo CAR ou com a distribuição ICAR como priori.



(a)



(b)



(c)

Figura 9.1: Autores: Renato Assunção, Erica Castilho e Elias Krainski.