UNIVERSIDADE ESTADUAL DE CAMPINAS
INSTITUTO DE MATEMÁTICA, ESTATÍSTICA E COMPUTAÇÃO CIENTÍFICA
DEPARTAMENTO DE ESTATÍSTICA

Estimação Não-Paramétrica para Função de Covariância de Processos Gaussianos Espaciais

José Clelto Barros Gomes

Dissertação de Mestrado orientada pelo

Prof. Dr. Ronaldo Dias

Campinas - SP

2009

ESTIMAÇÃO NÃO-PARAMÉTRICA PARA FUNÇÃO DE COVARIÂNCIA DE PROCESSOS GAUSSIANOS ESPACIAIS

Este exemplar corresponde à redação final da dissertação devidamente corrigida e defendida por José Clelto Barros Gomes e aprovada pela comissão julgadora.

Campinas, 30 de abril de 2009.

Prof. Dr. Ronaldo Dias
Orientador

Banca Examinadora:

1. Prof. Dr. Ronaldo Dias (IMECC-UNICAMP)
2. Prof(a). Dra. Roseli Aparecida Leandro (ESALQ-USP)
3. Prof. Dr. Jesus Enrique Garcia (IMECC-UNICAMP)

Dissertação apresentada ao Instituto de Matemática, Estatística e Computação Científica, UNICAMP, como requisito parcial para obtenção do Título de MESTRE em ESTATÍSTICA.

FICHA CATALOGRÁFICA ELABORADA PELA BIBLIOTECA DO IMECC DA UNICAMP
Bibliotecária: Miriam Cristina Alves - CRB8/5094

Gomes, José Clelto Barros

G585e Estimação não-paramétrica para função de covariância de processos Gaussianos espaciais / José Clelto Barros Gomes -- Campinas, [S.P. :s.n.], 2009.

Orientador : Ronaldo Dias

Dissertação (Mestrado) - Universidade Estadual de Campinas, Instituto de Matemática, Estatística e Computação Científica.

1. Estatística não paramétrica. 2. Processos gaussianos. 3. Spline, Teoria do. 4. Função de covariância. I. Dias, Ronaldo. II. Universidade Estadual de Campinas. Instituto de Matemática, Estatística e Computação Científica. III. Título.

Título em inglês: Nonparametric estimation for covariance function of spatial Gaussian processes

Palavras-chave em inglês (Keywords): 1. Nonparametric statistics. 2. Gaussian processes. 3. Splines theory. 4. Covariance function.

Área de concentração: Estatística não paramétrica.

Titulação: Mestre em Estatística.

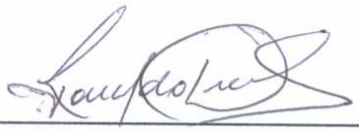
Banca examinadora: Prof. Dr. Ronaldo Dias (IMECC-UNICAMP)
Profa. Dra. Roseli Aparecida Leandro (ESALQ-USP)
Prof. Dr. Jesus Enrique Garcia (IMECC-UNICAMP)

Data da defesa: 30/04/2009

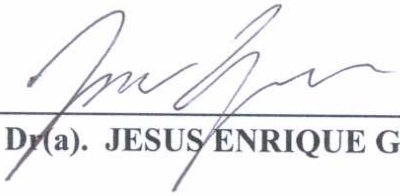
Programa de pós-graduação: Mestrado em Estatística.

Dissertação de Mestrado defendida em 30 de abril de 2009 e aprovada

Pela Banca Examinadora composta pelos Profs. Drs.



Prof(a). Dr(a). RONALDO DIAS



Prof(a). Dr(a). JESUS ENRIQUE GARCIA

Prof(a). Dr(a). ROSELI APARECIDA LEANDRO

Em memória dos meus avós
José Felipe, Raimundo "Doca" e
Maria "Loló".

Agradecimentos

Em primeiro lugar, agradeço a Deus, pela Sua maravilhosa criação e por ter permitido que eu chegasse até aqui.

Aos meus pais, Antônia Barros Gomes e Antônio Glene Vieira Gomes pelo apoio e incentivo durante todos os anos de estudos.

Aos meus irmãos, Antônia Débora Barros Gomes, Francisco Antônio Barros Gomes e Maria Cleia Barros Gomes pelo apoio e incentivo nessa longa trajetória de estudos e ao meu sobrinho Max William pelos momentos de felicidades.

Aos meus professores do Departamento de Estatística da Universidade Federal do Amazonas, em nome dos professores José Raimundo Gomes Pereira, José Cardoso Neto e Celso Rômulo Barbosa Cabral, pelo incentivo desde o início da graduação.

Aos meus amigos desde a graduação James Dean Oliveira dos Santos Junior (Dean), Antônio Alcirley da Silva Balieiro (Toin), Robério Rebouças da Silva (o Robit), Francisco de Oliveira Farias (Quiquin), Lúcia Rolim Santana e Áurea Barbosa Leitão.

Aos meus amigos de mestrado, em especial à Marta Cristina Colozza Bianchi, pelo tempo que passamos juntos estudando e pelas divertidas caronas em seu fusquinha (que capota mas não breca!).

Ao meu amigo e grande matemático Eduardo Xavier Miqueles por seu tempo concedido, me explicando transformada de Fourier das B-splines e pelas conversas aleatórias e incentivos nos estudos.

Ao meu orientador, Prof. Dr. Ronaldo Dias, pela paciência, apoio e direcionamento durante a elaboração deste trabalho.

Aos membros da banca examinadora, Prof(a). Dra. Roseli Aparecida Leandro (ESALQ-USP) e Prof. Dr. Jesus Enrique Garcia (IMECC-UNICAMP), pela leitura, correções e sugestões da dissertação.

Aos professores do Departamento de Estatística do IMECC-UNICAMP, pelos ensinamentos concedidos.

À Fundação de Amparo à Pesquisa do Estado do Amazonas (FAPEAM) pelo apoio financeiro a este projeto.

Não importa o que está acontecendo

nesse momento em sua vida, anime-se.

Existe esperança.

A Tua vontade Senhor, não a minha. Benny Hinn.

Resumo

O desafio na modelagem de processos espaciais está na descrição da estrutura de covariância do fenômeno sob estudo. Um estimador não-paramétrico da função de covariância foi construído de forma a usar combinações lineares de funções B-*splines*. Estas bases são usadas com muita frequência na literatura graças ao seu suporte compacto e a computação tão rápida quanto a habilidade de criar aproximações suaves e apropriadas. Verificou-se que a função de covariância estimada era definida positiva por meio do teorema de Bochner. Para a estimação da função de covariância foi implementado um algoritmo que fornece um procedimento completamente automático baseado no número de funções bases. Então foram realizados estudos numéricos que evidenciaram que assintoticamente o procedimento é consistente, enquanto que para pequenas amostras deve-se considerar as restrições das funções de covariância. As funções de covariâncias usadas na estimação foram as de exponencial potência, gaussiana, cúbica, esférica, quadrática racional, ondular e família de Matérn. Foram estimadas ainda covariâncias encaixadas. Simulações foram realizadas também a fim de verificar o comportamento da distribuição da afinidade. As estimativas apresentaram-se satisfatórias.

Palavras-Chave: Estatística não-paramétrica, Funções *Splines*, Função de Covariância, Processo Gaussiano.

Abstract

The challenge in modeling of spatials processes is in description of the framework of covariance of the phenomenon about study. The estimation of covariance functions was done using a nonparametric linear combinations of basis functions B-splines. These bases are used frequently in literature thanks to its compact support and fast computing as the ability to create smooth and appropriate approaches There was positive definiteness of the estimator proposed by the Bochner's theorem. For the estimation of the covariance functions was implemented an algorithm that provides a fully automated procedure based on the number of basis functions. Then numerical studies were performed that showed that the procedure is consistent assynthotically. While for small samples should consider the restrictions of the covariance functions, so the process of optimization was non-linear optimization with restrictions. The following covariance functions were used in estimating: powered exponential, Gaussian, cubic, spherical, rational quadratic and Matérn family. Nested covariance funtions still were estimated. Simulations were also performed to verify the behavior of affinity and affinity partial, which measures how good is the true function of the estimated function. Estimates showed satisfactory.

Keywords: Nonparametric Statistics, Splines Functions, Covariance Function, Spatial Gaussian Processes.

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

Diversos procedimentos tem sido propostos na definição da estimação da função de covariância espacial, por exemplo, aqueles baseados no método dos momentos ou sobre a utilização de uma função *kernel*, tal como consta em Cressie (1993), Hall e Patil (1994) e Bjϕrnstad e Falck (2001). No entanto, esta última proposta não é válida para a previsão, uma vez que a condição para que seja definida positiva normalmente não é satisfeita.

A função de covariância está inserida no contexto da geoestatística que é uma área da Estatística Espacial. As técnicas de covariância espacial são comumente usadas para descrever modelos genéticos e ecológicos. Em populações genéticas, a covariância espacial denota o modo que composições genéticas variam entre indivíduos distribuídos através do espaço. Diferentes modelos descrevem esta covariância como sendo, por exemplo, um modelo exponencial, um modelo de Matérn ou um modelo Gaussiano. A suavidade de um processo Gaussiano está diretamente relacionada à especificação de sua função de correlação. Estudos apontam que a covariância na composição genética de indivíduos ou

no crescimento de populações pode ser uma função da distância separadas pelas unidades amostrais.

O estudo de processos aleatórios Gaussianos é em grande parte um estudo de funções de covariância ou funções de correlação. Os processos Gaussianos são amplamente usados na literatura geoestatística. Esta área compreende dados pontuais referenciados, onde cada elemento da amostra tem uma locação referenciada por coordenadas geográficas (longitude, latitude), associada às observações de interesse. Um exemplo comum é o nível de certo poluente sobre uma região.

O método mais comum para se estimar a relação entre covariância e distância é um método não-paramétrico, chamado correlograma espacial. Correlogramas são muito úteis na inferência biológica. Há, no entanto, questões em que o aperfeiçoamento pode ser requerido. Por exemplo o fato do correlograma aproximar a função de covariância espacial contínua por uma função discreta. E também o próprio estimador não ser uma função de covariância válida, desde que ele não é definido positivo. Um correlograma é uma função que mostra as correlações entre os pontos amostrais separados pelas distâncias. A correlação geralmente decresce com a distância até alcançar zero.

Consideremos a medida z_i em um indivíduo i na coordenada (x_i, y_i). Agora assuma que há n indivíduos, e que a covariância espacial aos pares é uma função, $\rho(\tau)$, da distância, τ, separando os indivíduos. A distância geográfica, τ_{ij}, entre os indivíduos i e j é a distância Euclidiana:

$$\tau_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}.$$

A covariância amostral entre os dois indivíduos é:

$$\mathcal{C}ov(z_i, z_j) = (z_i - \bar{z})(z_j - \bar{z})$$

em que $\bar{z} = \frac{1}{n}\sum_{l=1}^{n} z_l$ é a média amostral. A correlação amostral entre os indivíduos i e j é estimada usando a equação:

$$\hat{\rho}_{ij} = \hat{\rho}(z_i, z_j) = \frac{(z_i - \bar{z})(z_j - \bar{z})}{1/n\sum_{l=1}^{n}(z_l - \bar{z})^2} \tag{1.1}$$

Entre todos os indivíduos há $n(n-1)/2$ pares de autocorrelações únicos, correspondendo ao triângulo superior (ou inferior) da matriz de autocorrelação amostral, isto é, $\hat{\rho}_{ij}$ para $i = 1, \ldots, n$, $j = i+1, \ldots, n$.

Bjϕrnstad e Falck (2001) apresentaram uma versão modificada da função de covariância não-paramétrica de Hall *et al.* 1994. Eles fizeram as seguintes suposições: os dados vem de um campo estacionário de segunda ordem; a esperança e a variância não mudam através do espaço; o campo é isotrópico, então a covariância depende somente da distância; os dados seguem uma distribuição Gaussiana. Aqui também serão feitas as mesmas suposiçoes. A seguir vamos falar um pouco mais sobre a geoestatística.

1.1 Geoestatística

Os métodos geoestatísticos, ou simplesmente geoestatística, foram desenvolvidos graças aos estudos do engenheiro de minas Georges Matheron na França no final da década de 50 e início dos anos 60. Estes métodos estão fundamentados na Teoria das Variáveis Regionalizadas, que foi formalizada por Matheron, a partir de estudos práticos desenvolvidos por Daniel G. Krige, no cálculo de reservas nas minas de ouro na África do Sul. Atualmente a geoestatística é aplicada em vários campos, desde as Ciências da Terra e Atmosfera, na Agricultura, nas Ciências do Solo e Hidrologia, Estudos Ambientais, processamento

de imagens e mais recentemente na Epidemiologia. A geoestatística tem sua origem na estimação de reservas de minério na indústria de mineração e está preocupada com a variação espacial contínua.

A essência de seu objetivo é que há um fenômeno espacial, $s(x)$, que desejamos estudar através de uma região espacial contínua e que pode ser pensada como uma realização de um processo estocástico contínuo, $S(\cdot) = \{S(x) : x \in G\}$ onde $G \subset R^2$ é uma região contínua que estamos estudando. Contudo, em geral $S(x)$ não é observável e nós somente medimos Y_i nas locações $x_i, i = 1, \ldots, n$, onde Y_i pode ser visto como uma versão de ruído de $S(x)$. E é assumido que o planejamento amostral para x_i ou é determinístico ou é estocástico, mas independente de $S(x)$. De uma forma geral a estatística espacial contém três grandes áreas: geoestatística, dados de área e processos pontuais. O desafio na modelagem de processos espaciais e espaço-temporais está na descrição da estrutura de covariância do fenômeno sob estudo.

As observações sobre uma ou mais variáveis são tomadas em posições múltiplas, identificáveis em alguma posição do domínio espacial. As localizações dessas posições são observadas, anexadas e classificadas como observações. A análise leva em conta a localização espacial dessas posições. As observações ou a localização espacial são modeladas como variáveis aleatórias, e as inferências são feitas sobre esses modelos e/ou sobre variáveis adicionais não observadas.

Dados geoestatísticos - pontos de observações de quantidades variando continuamente sobre uma região, por exemplos: acúmulo anual de chuva ácida nos Estados Unidos, riqueza de minério de ferro acumulada por um grupo comercial.

O termo geoestatística pode inicialmente ser confundido entre a totalidade de apli-

cações da estatística nas geociências com as técnicas específicas desenvolvidas na Escola de Minas de Paris segundo a orientação de Matheron e motivadas originalmente pelos problemas em mineração. A generalidade de sua formação e a oportunidade de sua aplicação em problemas com dados distribuídos espacialmente abriu a possibilidade de sua utilização em diversos domínios das ciências da natureza. No Brasil, seguindo de certa forma o mesmo percurso dos países desenvolvidos a geoestatística evoluiu do ambiente de mineração para se inserir de uma forma geral como ferramenta de trabalho e pesquisa do profissional de Ciências da Natureza, Estatística e Computação.

Segundo Isaaks e Srivastava (1989) três métodos para descrição de continuidade espacial foram sugeridos: a função de correlação, a função de covariância, e o variograma. Como ferramenta descritiva, qualquer uma das três são úteis. Mas para a proposta de estimação não são equivalentes. A teoria clássica de estimação é mais adequada para a função de covariância.

1.2 Objetivo

O principal objetivo desta dissertação foi construir um estimador não-paramétrico para a função de covariância espacial por meio de combinações lineares de funções B-*splines*, em que foram obtidas as condições de validação da função de covariância. Além disso, foi realizado um estudo para verificar o comportamento da distribuição da afinidade entre a função de covariância estimada e a função de covariância verdadeira.

1.3 Organização da dissertação

O Capítulo 2 refere-se sobre *splines* que podem ser representados por polinômios por partes. Então se apresenta em seguida os conhecidos B-*splines* que são funções bases que tornam os cálculos das funções *splines* mais fáceis. Os B-*splines* formam uma base de espaços *splines*. O uso de B-*splines* em regressão não paramétrica é imenso devido a importante propriedade computacional de ter suporte compacto, ou seja, por ser não-nulo num intervalo pequeno e zero fora desse intervalo.

O Capítulo 3 aborda os processos aleatórios, bem como os processos estacionários e os processos gaussianos. Apresenta as definições e os teoremas necessários para a validação do estimador da função de covariância. Mostra como o estimador foi construído e sob quais condições ele é válido.

O Capítulo 4 traz a forma como os números de bases dos B-*splines* foram estimados. Faz um estudo, por meio de exemplos, sobre o comportamento da distribuição da afinidade entre algumas funções de covariâncias estimadas e funções de covariâncias verdadeiras.

Quanto ao Capítulo 5, apresentamos os conceitos das funções de covariâncias a serem estimadas, como: exponencial, gaussiana, circular, cúbica, esférica, quadrática racional ou de Cauchy, quadrática racional generalizada, ondular e família de Matérn. Em seguida, as simulações foram feitas para estimar as funções de covariâncias. E aqui também se aplicou o estimador para o caso de funções de covariâncias encaixadas.

O Capítulo 6 se dedica as conclusões e sugestões para trabalhos futuros.

Todos os programas e figuras desta dissertação foram feitos no programa *R-Gui* versão 2.7.1.

Capítulo 2

Aproximação por Funções Splines

Uma função *spline* em $[0,1]$ com n nós, $0 \leq t_1 < t_2 < \ldots < t_n \leq 1$, é definida como um polinômio em cada um dos intervalos $[0,t_1), \ldots, (t_j, t_{j+1}), \ldots (t_n, 1]$. As partes dos polinômios são geralmente unidas no sentido de garantir continuidade, possuam um número específico de derivadas contínuas. Um *spline* univariado é ainda pensado como polinômios por partes polinomiais. Mais ainda, certas funções que satisfazem alguma equação diferencial são também chamadas funções *splines*. Há diferentes generalizações de funções *splines* com dimensões maiores, por exemplo a esfera. Todas elas chamadas *splines*, mas nem todas são representadas como polinômios por partes.

Definição 2.1. *Uma função $s(t)$ é chamada de spline com grau r e nós em $\{t_j\}_{j=1}^{k}$ se $-\infty = t_0 < t_1 < t_2 < \ldots < t_k < t_{k+1} = \infty$ e*

- *para cada $j = 0, \ldots, k$, $s(t)$ coincide em $[t_j, t_{j+1}]$ com polinômio de grau não maior do que r;*
- *$s(t), s'(t), \ldots, s^{r-1}(t)$ são funções contínuas em $\mathbb{R}$.*

O conjunto de tais funções, $\mathcal{S}_m(t_1, \ldots, t_k)$, é um espaço linear de dimensão $r + k + 1$, cujo os elementos são funções *splines* e é chamado de espaço *spline*. Note que a ordem do polinômio é igual a $r + 1$ e k é o número de nós internos. Ainda temos que um *spline* de ordem m com nós em $t_1, \ldots, t_k$ é qualquer função da forma:

$$s(x) = \sum_{j=0}^{m-1} \theta_j x^j + \sum_{j=1}^{k} \delta_j (x - t_j)_+^{m-1}, \tag{2.1}$$

em que os coeficientes $\theta_0, \ldots, \theta_{m-1}, \delta_1, \ldots, \delta_k$ são números reais e ξ_+^{m-1} é definido abaixo. Então $\mathcal{S}_m(t_1, \ldots, t_k)$ é um espaço vetorial desde que as funções $1, x, \ldots, x^{m-1}, (x-t_1)_+^{m-1}$, $\ldots, (x-t_k)_+^{m-1}$ são linearmente independentes. Conclui-se que qualquer função *spline* pode ser escrita como uma combinação linear de $r + k + 1$ funções bases (Schumaker, 1981).

Definição 2.2. *Dado um ponto $x \in [t_j, t_{j+k}], j = 0, \ldots, k$ a função*

$$(x - t)_+^r = \begin{cases} (x-t)^r, & se\ x \geq t \\ 0 \qquad\quad , & se\ x < t \end{cases} \tag{2.2}$$

é chamada função poder truncada de grau r com nó t. Assim, um *spline* satisfaz o seguinte:

- s é um polinômio por partes de ordem m em qualquer subintervalo $[t_j, t_{j+1})$;
- s tem $m - 2$ derivadas contínuas;
- s tem a $(m - 1)$- ésima derivada e é uma função escala com saltos em $t_1, \ldots, t_k$.

Seria interessante se tivéssemos funções bases que tornassem os cálculos das funções *splines* mais fáceis. Os conhecidos B-*splines* formam uma base de espaços *spline*. Além disso, os B-*splines* possuem uma importante propriedade computacional de ter suporte compacto, por ser não-nulo num intervalo pequeno e zero fora desse intervalo.

2.1 Bases B-*spline*

Antes de definirmos B-*spline* vamos definir diferença dividida. Diferenças divididas podem ser definida de diversas maneiras, todas equivalentes, (Schumaker, 1981; Hoffmann e Hämmerlin, 1991). Para uma função $f(t)$ e um conjunto de pontos distintos $\{t_0, t_1, \ldots, t_m\}$ temos:

- diferença dividida de ordem 1 nos pontos $\{t_0, t_1\}$

$$[t_0, t_1]f = \frac{f(t_1) - f(t_0)}{t_1 - t_0};$$

- diferença dividida de ordem 2 nos pontos $\{t_0, t_1, t_2\}$

$$[t_0, t_1, t_2]f = \frac{[t_1, t_2]f - [t_0, t_1]f}{t_2 - t_0};$$

- de um modo geral, pode-se usar a notação $D^k f(t_i) = [t_i, t_{i+1}, \ldots, t_{i+k}]f$, $i = 0, 1, \ldots, m$, para designar a diferença dividida de ordem $k\,(k \geq 1)$ entre os $(k+1)$ pontos $t_i, t_{i+1}, \ldots, t_{i+k}$, sendo

$$D^k f(t_i) = \frac{D^{k-1} f(t_{i+1}) - D^{k-1} f(t_i)}{t_{i+k} - t_i}. \tag{2.3}$$

Agora a definição de B-*spline*. Seja $B_{j,m}(x)$ a B-spline de Curry e Schoenberg com nós $t_j < t_{j+1} < \ldots < t_{j+m}$ $(j \in \mathbb{Z},\ m = 1, 2, \ldots)$, isto é, $B_{j,m}(x)$ é dada pela fórmula,

$$B_{j,m}(x) = m[t_j, t_{j+1}, \ldots, t_{j+m}]q_m(\bullet, x), \tag{2.4}$$

em que

$$q_m(t, x) = (t - x)_+^{m-1} = \begin{cases} (t-x)^{m-1}, & \text{se } t \geq x \\ 0 \qquad\quad , & \text{se } t < x. \end{cases} \tag{2.5}$$

De Boor (2001) construiu um algoritmo para calcular B-*splines* de qualquer grau a partir de B-*spline* com ordem inferior, de maneira recursiva. Pois uma B-*spline* de grau zero é uma constante num intervalo entre dois nós. Aqui usamos somente nós equidistantes, mas o algoritmo pode ser usado para qualquer posicionamento dos nós. Assim, a j-ésima B-*spline* de ordem m para uma sequência de nós não-decrescentes $x = \{x_j\}$ pode ser calculada por,

$$B_{j,m}(x) = \frac{x - t_j}{t_{j+m-1} - t_j} B_{j,m-1}(x) + \frac{t_{j+m} - x}{t_{j+m} - t_{j+1}} B_{j+1,m-1}(x), \tag{2.6}$$

em que

$$B_{j,1}(x) = \begin{cases} 1, & \text{se } t_j \leq x < t_{j+1} \\ 0, & \text{caso contrário.} \end{cases} \tag{2.7}$$

As Figuras 2.1 - 2.4 mostram sequências de B-*splines* até ordem quatro com nós igualmente espaçados no intervalo $(0, 1)$. A Figura 2.2 à esquerda mostra uma B-*spline* de grau 1. Ela consiste de duas peças lineares. Uma peça de x_1 a x_2 e outra de x_2 a x_3. Os nós são x_1, x_2 e x_3. À esquerda de x_1 e à direita de x_3 esta B-*spline* é zero. Ao lado direito da Figura 2.2 são mostradas mais cinco B-*splines* de grau 1. Note que é possível construir um conjunto tão grande quanto se queira de B-*splines*, pela introdução de mais nós.

A Figura 2.3 à esquerda mostra uma B-*spline* de grau 2. Ela consiste de três peças quadráticas unidas em dois nós internos. A B-*spline* é baseada em quatro nós adjacentes. Ao lado direito da Figura 2.3 são mostradas mais seis B-*splines* de grau 2.

A Figura 2.4 à esquerda mostra uma B-*spline* de grau 3. Ela consiste de quatro quadráticas unidas em três nós internos. A B-*spline* é baseada em cinco nós adjacentes. Ao lado direito da Figura 2.4 são mostradas mais sete B-*splines* de grau 3. As B-*splines* sobrepõem umas nas outras, sendo os extremos com menos sobreposição.

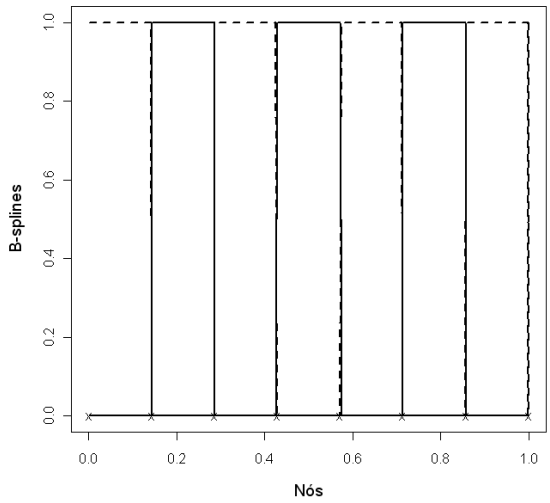


Figura 2.1: Ilustração de B-*spline*, de ordem 1.

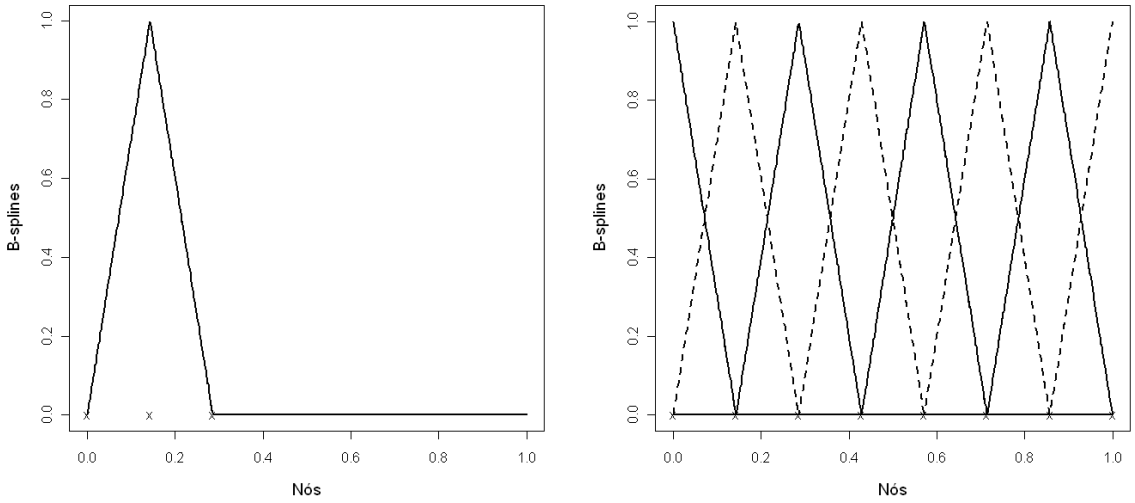


Figura 2.2: Ilustrações de uma B-*spline* isolada e outras B-*splines* sobrepostas, de ordem 2.

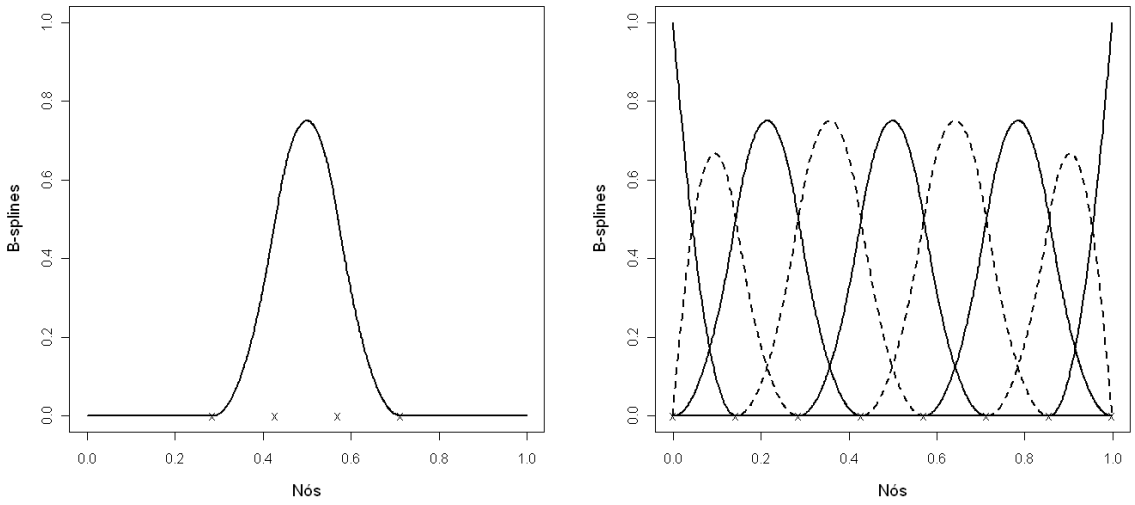


Figura 2.3: Ilustrações de uma B-*spline* isolada e outras B-*splines* sobrepostas, de ordem 3.

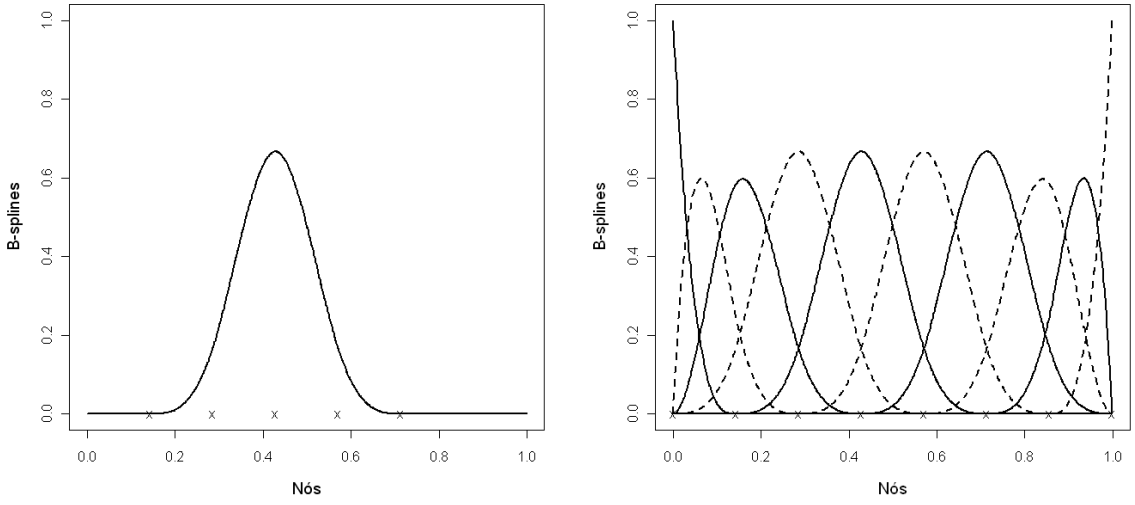


Figura 2.4: Ilustrações de uma B-*spline* isolada e outras B-*splines* sobrepostas, de ordem 4.

As figuras ilustrativas mostram as propriedades gerais de um B-*spline* de grau r, (Eilers e Marx, 1996):

- consiste de $r+1$ pedaços polinomiais, cada de grau r;
- as partes polinomiais se unem em r nós internos;
- na junção dos pontos, as derivadas até a ordem $r-1$ são contínuas;
- a B-*spline* é positiva no domínio gerado por $r+2$ nós;
- exceto nos extremos, ela se sobrepõe com $2r$ peças polinomiais de seus vizinhos;
- em um dado t, $r+1$ B-*splines* são não-nulas.

O conjunto de B-*splines* $B_{j,m,t}$ forma uma base que gera um espaço vetorial, $S_{m,t}$. Qualquer função *spline* polinomial de ordem m pode ser formulada como uma combinação linear das funções bases B-*spline*, a seguir:

$$s(x) = \sum_{j=1}^{n} a_j B_{j,m,t}(x). \tag{2.8}$$

Os coeficientes a_j podem ser selecionados de forma a impor condições específicas sobre a função *spline*. Por exemplo, impondo que a função *spline* concorda com um determinado conjunto de dados $(x,\ g(x),\ j = 1, \ldots, n)$ chegamos às condições de interpolação de *spline*,

$$\sum_{i=1}^{n} a_i B_{i,m,t}(x_j) = g(x_j),\ j = 1, \ldots, n. \tag{2.9}$$

Este é um sistema linear de equações nos coeficientes a_i's. O sistema é invertível desde que,

$$t_j \leq x_j < t_{j+m}.$$

Sistema linear semelhante pode ser configurado para outras condições, como a aproximação por mínimos quadrados. Modelos *splines* da forma de (2.8) são modelos não-paramétricos.

Um modelo de regressão não-paramétrica assume em geral propriedades qualitativas da função adjacente que são satisfeitas se o modelo é selecionado de uma coleção de funções tal como o subespaço $S_{m,t}$. Isso permite maior flexibilidade na possível forma da função sem a necessidade dos pressupostos atrelados aos modelos paramétricos. Modelos não-paramétricos dependem mais fortemente dos dados e são não viesados pela informação a priori. As bases B-*splines* são usadas com muita frequência para dados não periódicos no contexto de dados funcionais. Graças ao seu suporte compacto e a computação rápida tanto quanto a habilidade de criar aproximações suaves e apropriadas. Para tratar dados ou funções periódicas as séries de Fourier podem ser usadas com maior propriedade. Uma B-*spline* consiste de partes polinomiais polinomiais unidas de maneira especial.

As B-*splines* cúbicas são escolhas comuns para funções base, pois estas funções possuem desempenho computacional superior e tem a característica de que cada B_j tem suporte compacto. Na prática, isto nos diz que a matriz resultante, com entradas $B_{i,j} = B_j(x_i)$, para $j = 1, \ldots, K$ e $i = 1, \ldots, n$, é tridiagonal. Hastie e Tibshirani, (1998) acrescenta que além do desempenho computacional as técnicas do modelo linear padrão podem ser aplicadas. Mas a suavidade da estimativa não pode ser facilmente variada como função de um parâmetro de suavização simples.

Ramsay (1988) diz que a aplicação de polinômios do tipo $f(t) = \sum_{i=1}^{K} a_i t^{i-1}$ na matemática aplicada se deve ao fato deles serem lineares nos parâmetros a_i, a serem estimados e as funções t^{i-1} que são combinadas linearmente são facilmente manipuladas algebricamente e numericamente, em especial com relação a diferenciação e integração.

Infelizmente, a principal dificuldade no trabalho de regressão por *splines* é selecionar o número e as posições de uma sequência de pontos chamados nós, em que as partes dos polinômios cúbicos são ligados para reforçar continuidade e derivadas contínuas de baixa ordem (veja detalhes em Schumaker, 1972).

Para exemplificar a ação de K na curva estimada, vamos considerar um exemplo por simulação com $y(x) = \exp(-x)\text{sen}(\pi x/2)\cos(\pi x) + \varepsilon$, com $\varepsilon \sim N(0; 0,025^2)$. As estimativas das curvas foram obtidas pelo método de mínimos quadrados com três diferentes números de funções bases, que são as B-*splines*.

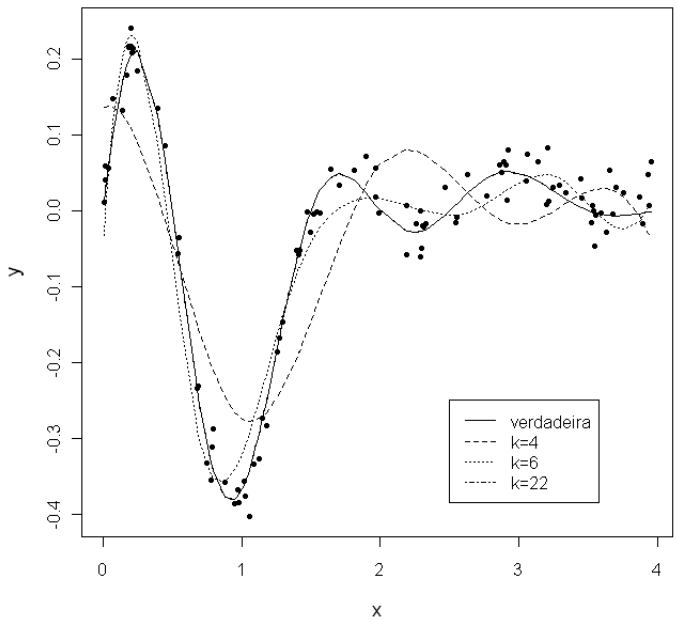


Figura 2.5: Ajuste de mínimos quadrados *splines* para diferentes valores de K.

A Figura 2.5 mostra o efeito de variação das funções bases na estimativa da curva verdadeira. Observe que valores pequenos de K deixa a estimativa muito suave, podendo

ocorrer uma sobresuavização. Para valores grande de K pode causar uma subsuavização.

Também é possível generalizar o conceito de *splines* de mínimos quadrados para dimensões maiores.

2.2 Produto Tensorial de *Splines*

A construção de produto tensorial é útil na generalização dos métodos de aproximação para funções com várias variáveis.

A maneira de solucionar a suavização sobre uma janela finita é via produto tensorial de *splines* que consiste de métodos sistemáticos de famílias de funções suaves unidimensionais para gerar superfícies em espaços de altas dimensões.

Dadas funções unidimensionais $\delta : T \longrightarrow \mathbb{R}$ e $\epsilon : U \longrightarrow \mathbb{R}$, o produto tensorial de δ e ϵ é a função $\delta \otimes \epsilon : T \times U \longrightarrow \mathbb{R}$ definida por $(\delta \otimes \epsilon)(t, u) = \delta(t)\epsilon(u)$. Para um conjunto de funções linearmente independente $\{\delta_{j_1} : j_1 = 1, 2, \ldots, q_1\}$ definida em T, queremos que seja suave e uma base $\epsilon_{j_2} : j_2 = 1, 2, \ldots, q_2$ de funções suaves em U. O produto tensorial desses dois espaços de funções é o conjunto de todas as combinações lineares de produtos tensoriais de combinações lineares destas funções base, dada por

$$G = \left\{ \sum_r c_r \left(\sum_{j_1=1}^{q_1} a_{r_{j_1}} \delta_{j_1} \right) \otimes \left(\sum_{j_2=1}^{q_2} b_{r_{j_2}} \epsilon_{j_2} \right) \right\} \tag{2.10}$$

que pode ser representada pelo conjunto de todas as combinações lineares de produtos tensoriais de funções bases

$$G = \left\{ \sum_{j_1=1}^{q_1} \sum_{j_2=1}^{q_2} \delta_{j_1} \otimes \epsilon_{j_2} \right\}. \tag{2.11}$$

splines cúbicos de produto tensorial são exemplos óbvios desta construção. Suponha $\{\tau_1, \tau_2, \ldots, \tau_{m_1}\}$ e $\{U_1, U_2, \ldots, U_{m_2}\}$ são sequências regularmente espaçadas, tal que $T = [\tau_1, \tau_{m_1}]$ e $U = [U_1, U_{m_2}]$. Então os *splines* cúbicos em cada uma dessas sequências de nós, restrita aos intervalos T e U, respectivamente, são espaços finito dimensional de dimensão $q_1 = m_1 + 2$ e $q_2 = m_2 + 2$ (veja De Boor (2001) para mais detalhes).

O caso unidimensional pode ser estendido para múltiplas dimensões por meio da construção do produto tensorial de *spline*. Um subespaço de *spline* é definido para cada dimensão. Portanto uma função do produto tensorial na n-ésima dimensão pode ser expressada como:

$$f(x_1, \ldots, x_r) = \sum_{j_1=0}^{n_1} \ldots \sum_{j_r=0}^{n_r} a_{j_1,\ldots,j_r} B_{j_1,m_1,t_1}(x_1) \ldots B_{j_r,m_r,t_r}(x_r). \tag{2.12}$$

O produto tensorial de *splines* fornece uma generalização simples e direta das funções *splines* unidimensional. A aproximação de funções multivariadas por produto tensorial tem um forte vício direcional ao longo das linhas paralelas à direção do eixo. Características funcionais difíceis que ocorrem ao longo de diferentes direções exigem uma malha fina em todos os espaços funcionais, a fim de alcançar uma aproximação mais precisa. Isto aumenta o custo computacional do problema. Recentemente, avanços em pesquisa matemática sobre a teoria e métodos computacionais de *splines* multivariados mostraram grande potencial a várias aplicações sem a condição da construção de produto tensorial.

Para o espaço bi-dimensional, o produto tensorial da B-*spline* equivalente de (2.4) pode ser escrito como:

$$f(x, y) = \sum_{i=0}^{n_x} \sum_{j=0}^{n_y} a_{ij} B_{i,m_x,t_x}(x) B_{j,m_y,t_y}(y). \tag{2.13}$$

Em que $B_{j,m,t}$ é a j-ésima B-*spline* de ordem m para a sequência de nós (t), n_x e n_y o número de nós na direção X e Y, respectivamente. Como no caso unidimensional, os

coeficientes podem ser obtidos pela aplicação de interpolação ou outras condições mais gerais.

Capítulo 3

A construção do Modelo

3.1 Processos Aleatórios

Um processo aleatório ou estocástico é um conjunto de variáveis aleatórias que tem algumas locações espaciais e cuja dependência uma da outra é especificada por algum mecanismo de probabilidade. Para uma descrição completa de seu mecanismo de geração probabilística podemos calcular diversos parâmetros que caracteriza um processo aleatório.

Uma definição formal de processos aleatórios pode ser vista como:

Definição 3.1. *Dados um espaço de probabilidade,* $(\Omega, \mathcal{A}, P)$*, e um conjunto de índices,* G*. Um processo aleatório é então uma função real ou finita* $Y(s, \omega)$ *que, para todo* $s \in G$ *fixado é uma função mensurável de* $\omega \in \Omega$*.*

Os sinônimos como campos aleatórios e processos estocásticos são comumente usados

por alguns autores. Sendo campos aleatórios usados para indicar que a dimensão é maior que um. Enquanto que processo estocástico é usado para dizer que é um campo aleatório unidimensional. Para detalhes sobre espaço de probabilidade veja James (2004).

3.1.1 Processos Estacionários

Neste trabalho usamos os conceitos de processos estacionários. Supomos que $\{y(s) : s \in G\}$ é uma realização de um processo aleatório,

$$\{Y(s) : s \in G\}, \tag{3.1}$$

em que G é um subconjunto de um espaço Euclidiano p-dimensional. Para enfatizar a aleatoriedade, algumas vezes o processo (3.1) é escrito como $\{y(s,\omega) : s \in G; \omega \in \Omega\}$, em que $(\Omega, \mathcal{A}, P)$ é um espaço de probabilidade. A realização $\{y(s) : s \in G\}$ corresponde a um valor particular de ω, digamos $\omega = \omega_0$.

O processo aleatório (3.1) é geralmente definido pela distribuição finita dimensional,

$$\mathbf{F}_{s_1,\ldots,s_k}(y_1,\ldots,y_k) \equiv P(Y(s_1) \leq y_1,\ldots,Y(s_k) \leq y_k),\ k \geq 1. \tag{3.2}$$

Para modelos espaciais com processos aleatórios estacionários, muito frequente em geoestatística, a função de covariância que será vista no próximo capítulo, e o variograma fornecem exatamente a mesma informação em formas diferentes de representação. O variograma possui a mesma forma da função de covariância, exceto que ele é invertido. Enquanto a covariância inicia do máximo de σ^2 quando $\tau = 0$ e decresce para zero, o variograma inicia em 0 e cresce ao máximo que é σ^2. A função de covariância, eventualmente alcança 0 enquanto o variograma eventualmente alcança o valor máximo, conhecido como patamar. Este valor, o patamar, do variograma é também a variância do processo

aleatório. A função de covariância é a covariância entre variáveis aleatórias separadas por uma distância, τ. Na maioria das vezes os processos aleatórios, que são usados na prática em geoestatística, os pares de variáveis aleatórias são independentes uns dos outros.

3.1.2 Estacionariedade Estrita

Um processo é dito estacionário estrito quando as distribuições finito dimensional são invariantes sob translação arbitrária dos pontos pelo vetor τ,

$$P(Y(s_1) \leq y_1, \ldots, Y(s_k) \leq y_k) = P(Y(s_1 + \tau) \leq y_1, \ldots, Y(s_k + \tau) \leq y_k), \tag{3.3}$$

$\forall\ s_1, \ldots, s_k \in G$ e $k \geq 1$.

3.1.3 Estacionariedade de Segunda Ordem

Quando um processo aleatório é estacionário, seus momentos são invariantes sob translações. Considerando somente os dois primeiros momentos, temos para os pontos s e $s+\tau$ de $\mathbb{R}^p$,

$$E[Y(s)] = \mu, \text{ para todo } s \in G. \tag{3.4}$$

$$E[Y(s) - \mu][Y(s + \tau) - \mu] = \mathcal{C}(\tau). \tag{3.5}$$

A média é constante e a função de covariância depende somente de τ. Por definição, uma variável aleatória satisfazendo as condições acima é estacionária de segunda ordem, ou ainda, fracamente estacionária. E mais, a estacionariedade da covariância implica na estacionariedade da variância,

$$Var[Y(s)] = E[Y(s) - \mu]^2 = \mathcal{C}(0). \tag{3.6}$$

Definição 3.2. *Um processo aleatório $Y(\cdot)$ que satisfaz (3.4) e (3.6) é definido ser estacionário de segunda ordem (or fracamente estacionário). Além disso, se $\mathcal{C}(s_1 - s_2)$ é uma função somente de $\|s_1 - s_2\|$, então $\mathcal{C}$ é chamado isotrópico.*

3.1.4 Processos Gaussianos

Um Processo Gaussiano é uma coleção infinita não enumerável de variáveis aleatórias com a propriedade de qualquer subconjunto finito destas variáveis tem distribuição normal multivariada de dimensão p, vetor de média $\boldsymbol{\mu}$ e matriz de covariância $\boldsymbol{\Sigma}$. Um Processo Gaussiano pode ser definido como segue:

Definição 3.3. *Seja $S(\cdot)$ tomando valores $S(\mathbf{x})$ para $\mathbf{x} \in G \subset \mathbb{R}^p$, geralmente com $p = 1, 2$ ou 3, tem distribuição de um Processo Gaussiano com função média $m(\cdot)$ e função de covariância $\mathcal{C}(,\cdot,)$ denotado por $S(\cdot) \sim PG(m(\cdot), \mathcal{C}(,\cdot,))$ se para qualquer $\mathbf{x}_1, \ldots, \mathbf{x}_n \in G$ e qualquer $n = 1, 2, \ldots$, a distribuição conjunta de $S(\mathbf{x}_1), \ldots, S(\mathbf{x}_k)$ é normal multivariada com parâmetros dados por $E(S(\mathbf{x}_j)) = m(\mathbf{x}_j)$ e $\mathcal{C}ov(S(\mathbf{x}_i), S(\mathbf{x}_j)) = \mathcal{C}(\mathbf{x}_i, \mathbf{x}_j)$.*

Distribuição normal p-dimensional ou Gaussiana. Esta distribuição é determinada pela densidade de probabilidade, como segue:

Teorema 3.4. *Se $\boldsymbol{\Sigma}$ é definida positiva ou equivalentemente não singular, então se $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $\mathbf{X}$ tem função densidade dada por,*

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}[det(\boldsymbol{\Sigma})]^{p/2}} \times exp\left\{\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right\}. \tag{3.7}$$

Para a construção do processo Gaussiano adotamos um gride no quadrado $[0, 5] \times [0, 5]$ e baseadas em 64 localizações foram geradas amostras de tamanho $25, 100, 500$ e 1000 da

distribuição normal multivariada em (3.7).

3.2 Covariância e Correlação

Uma função aleatória Y é chamada estacionária de segunda ordem se para cada variável $Y(s)$ ela possui média μ, finita, independente de s. A covariância centrada em relação à média é definida como:

$$\mathcal{C}ov(Y,Z) = \sigma_{YZ} = E[(Y - E(Y))(Z - E(Z))],$$

se estas esperanças existem e são finitas. Desenvolvendo o produto, obtemos uma expressão alternativa para a covariância,

$$\mathcal{C}ov(Y,Z) = E[YZ - ZE(Y) - YE(Z) + E(Y)E(Z)] = E(YZ) - E(Y)E(Z).$$

Uma covariância é uma função par, e pela desigualdade de Schwarz é limitada pelo seu valor na origem

$$\mathcal{C}(\tau) = \mathcal{C}(-\tau), \quad |\mathcal{C}(\tau)| \leq \mathcal{C}(0).$$

Se $\mathcal{C}ov(Y,Z) = 0$, dizemos que as variáveis Y e Z são não-correlacionadas.

Se Y e Z são independentes e integráveis, então são não-correlacionadas, pois neste caso $E(YZ) = E(Y)E(Z)$, mas por outro lado, $E(YZ) = E(Y)E(Z)$ não implica em independência, ou seja, covariância nula não necessariamente implica independência. Algumas propriedades da covariância são:

1. $\mathcal{C}ov(Y,Y) = Var(Y)$

2. $\mathcal{C}ov(aY, bZ) = ab\mathcal{C}ov(Y, Z)$.

Se $Y_1, Y_2, \ldots, Y_n$ são variáveis aleatórias integráveis, então:

$$\begin{aligned} Var(Y_1 + Y_2 + \ldots + Y_n) &= \sum_{i=1}^{n}\sum_{j=1}^{n}\mathcal{C}ov(Y_i, Y_j) \\ &= \sum_{i=1}^{n} Var(Y_i) + 2\sum_{i=1}^{n}\sum_{j=1}^{n-1}\mathcal{C}ov(Y_i, Y_j). \end{aligned}$$

Esta fórmula é muito importante na geoestatística, pois na maioria das vezes, são manipuladas combinações lineares do tipo $Z = \sum_{i=1}^{n}\lambda_i Y_i$ e para deduzí-la é necessário saber o seguinte resultado algébrico:

$$Z^2 = \left(\sum_{i=1}^{n}\lambda_i Y_i\right)^2 = \sum_{i=1}^{n}\sum_{j=1}^{n}\lambda_i\lambda_j Y_i Y_j.$$

Sejam Y e Z variáveis aleatórias com variâncias finitas e positivas. O coeficiente de correlação entre Y e Z, ρ é igual a:

$$\rho(Y, Z) = \frac{\mathcal{C}ov(Y, Z)}{\sigma_Y\sigma_Z} = \left[E\left(\frac{Y - E(Y)}{\sigma_Y}\right)\left(\frac{Z - E(Z)}{\sigma_Z}\right)\right].$$

3.3 Apresentação do Estimador

Geralmente possuímos um processo estocástico, $\{Y(\mathbf{x}), \mathbf{x} \in G\}$, em que G é um subconjunto fixado de um espaço Euclidiano p-dimensional, observado em um conjunto finito de locações $\mathbf{x}_1, \ldots, \mathbf{x}_n$. Desse modo temos uma realização parcial do processo estocástico e, baseado nas n observações, poderíamos querer, por exemplo, fazer inferência sobre o

processo espacial $Y(\cdot)$ e predição do processo em novas locações de interesse. E para isso é preciso caracterizar o processo espacial, como saber a média e a covariância do processo.

Na literatura geoestatística é comum assumirmos que o processo de interesse segue um processo Gaussiano. Normalmente, assumimos que

$$Y(\mathbf{x}) = \mu(\mathbf{x}) + S(\mathbf{x}), \tag{3.8}$$

em que $\mu(\mathbf{x})$ representa a média do processo espacial, possivelmente dependendo do conjunto de covariáveis, $f_j(\cdot),\ j = 1, \ldots, q$ e $S(\cdot)$ é um processo Gaussiano com estrutura de covariância, $\mathcal{C}ov(Y(\mathbf{x}), Y(\mathbf{x}')) = \sigma^2\rho(\|\mathbf{x} - \mathbf{x}'\|, \phi^*)$. Observe que $S(\cdot)$ é um processo estacionário com variância σ^2 e função de correlação, $\rho(\|.\|, \phi^*)$, que depende da distância Euclidiana entre as locações e um vetor de parâmetro ϕ^*. O cálculo das distâncias são indispensáveis para a análise espacial (Banerjee, 2006).

Nosso interesse está no modelo,

$$Y(\mathbf{x}) = \mu(\mathbf{x}) + S(\mathbf{x}) \tag{3.9}$$

em que $\mu(\mathbf{x})$ é a função média do processo e $S(\mathbf{x})$ o processo Gaussiano com média zero e estrutura de covariância dada por,

$$\mathcal{C}ov(Y(\mathbf{x}), Y(\mathbf{x}')) = \sigma(\|\mathbf{x} - \mathbf{x}'\|).$$

em que σ agora significa uma função que depende apenas da distância Euclidiana. Neste caso, $S(\cdot)$ é um processo estacionário totalmente caracterizado pela função de covariância $\sigma(\|\mathbf{x} - \mathbf{x}'\|)$. Para que as funções σ deem origem a uma estrutura de covariância válida é necessário garantir que a σ seja definida positiva. A função σ pode ser aproximada por uma combinação linear da forma,

$$\sigma(\theta, u) = \sum_{j=1}^{K} \theta_j B_j(u),$$

em que $B_j, j = 1, \ldots, K$ são B-splines.

Para estimar σ, poderíamos usar um estimador do tipo

$$\tilde{\sigma}(u) = \sum_{j=1}^{K} \tilde{\theta}_j B_j(u),$$

em que $\tilde{\theta}_j$ são encontrados minimizando a distância quadrática. Para garantir que $\tilde{\sigma}$ seja definida positiva vamos usar a transformada de Fourier,

$$\tilde{\sigma}^*(a) = \int_{-\infty}^{\infty} \tilde{\sigma}(t) e^{-iat} dt = 2 \int_0^{\infty} \tilde{\sigma}(t) \cos(at) dt.$$

Portanto, a estimativa de $\hat{\sigma}$ é dada por

$$\hat{\sigma}(u) = \frac{1}{2\pi} \int_{-a^*}^{a^*} \tilde{\sigma}^*(u) du,$$

em que $a^* = \inf\{a > 0;\ \tilde{\sigma}^*(a) \geq 0\}$. Agora $\hat{\sigma}$ é definida positiva .

3.4 Validação da Função de Covariância

A escolha de um modelo apropriado pode ser uma tarefa difícil, pois devemos tomar cuidado se este é um modelo de covariância válido. A função $\mathcal{C}(s, s')$ é uma função de covariância válida se ela satisfaz a seguinte definição:

Definição 3.5. *$\mathcal{C}(s, s')$ é definida positiva se para qualquer inteiro positivo n, $s_j \in G$ e $c_j \in \mathbb{R}$ para $j = 1, \ldots, n$,*

$$\sum_{i,j} c_i c_j \mathcal{C}(s_i, s_j) > 0 \tag{3.10}$$

a expressão acima é igual a 0 se, e somente se, $c_i = 0$ para todo i.

O Teorema de Bochner (Yaglom, 1987), dá a condição suficiente para uma função ser de tipo positivo. Uma função é de tipo positivo se ela for a transformada de Fourier de uma medida positiva mensurável. A seguir, apresentaremos o teorema que sustentará a condição de que a função de covariância seja definida positiva.

Teorema 3.6. *Teorema de Bochner. Uma função complexa $\mathcal{C}(\omega)$, $\omega \in \mathbb{R}$, é definida positiva se, e somente se, ela pode ser representada como a integral de Fourier-Stieltjes da forma,*

$$\mathcal{C}(\omega) = \int_{-\infty}^{\infty} e^{-i\omega x} d\mathbf{F}(x), \tag{3.11}$$

em que $\mathbf{F}(x)$ é uma função monótona não-crescente.

A classe de funções características de distribuições de probabilidades coincidem com a classe de funções definidas positivas contínuas, tomando o valor 1 em $\omega = 0$.

Agora seja $\sigma(\tau), \tau \in \mathbb{R}$, a função de covariância de um processo Gaussiano. Pelo Teorema de Bochner teríamos que, $\tilde{\sigma}^*(a)$

$$\begin{aligned} \tilde{\sigma}^*(a) &= \int_{-\infty}^{\infty} e^{-iat} \tilde{\sigma}(t) dt \\ &= \int_{-\infty}^{\infty} e^{-iat} \sum_{j=1}^{K} \tilde{\beta}_j B_j(t) dt \\ &= \sum_{j=1}^{K} \tilde{\beta}_j \int_{-\infty}^{\infty} e^{-iat} B_j(t) dt \\ &= \sum_{j=1}^{K} \tilde{\beta}_j \mathcal{F}(B_j). \end{aligned} \tag{3.12}$$

fosse positivo. Note que para a transformada de Fourier de $\tilde{\sigma}(t)$, $\tilde{\sigma}^*(a)$, ser positiva basta que a transformada de Fourier da B-*spline* e os coeficientes, $\tilde{\beta}_j$, sejam não negativos. Embora a transformada de Fourier da B-*spline* seja positiva, que veremos na Seção 3.5 deste capítulo, mas isso ainda não garante que os coeficientes, β_j, sejam positivos.

Nota-se que poderia haver inúmeras soluções para os coeficientes de tal forma que a tranformada de Fourier seja positiva. Uma solução seria se os coeficientes fossem zero, mas para outras soluções não é tão imediata. Assim, neste trabalho propomos uma solução que é válida para as seguintes condições apresentadas na Tabela 3.1.

Tabela 3.1: Condições em que o sistema em (3.12) precisa de restrições nos coeficientes dos β_j.

Função de Covariância	Amostras Grandes	Amostras Pequenas
$\mathcal{C}(\tau) \geq 0$	sem restrição	com restrição
$\mathcal{C}(\tau) < 0$	sem restrição	não resolvido

Vamos mostrar por meio de simulações, no Capítulo 5, que assintoticamente não se precisa das restrições nos coeficientes β_j. Quando o tamanho amostral é pequeno a estimativa da função de covariância não ficou tão boa quanto para o tamanho amostral grande, que não precisa das restrições nos coeficientes.

Para o estimador de $\mathcal{C}(\tau) \geq 0$ funcionar em pequenas amostras forçamos sua positividade, resolvendo o sistema de mínimos quadrados de tal forma que os coeficientes fossem positivos. Assim, quando aplicado na transformada de Fourier conseguimos que a função de covariância fosse definida positiva, pelo Teorema de Bochner. Quando $\mathcal{C}(\tau)$ assume valores negativos, que é o caso da função de covariância ondular, não foi verificado sob quais condições seria definida positiva.

Nota-se que a função *lm* do programa *R-Gui*, que resolve o sistema de mínimos quadrados, não garante a positividade dos coeficientes. A função *lm* serve quando $\mathcal{C}(\tau)$ assume valores negativos. Por isso, buscou-se encontrar outra função que retornasse os coeficientes positivos.

Agora vamos apresentar o estimador da função de covariância usado para amostras pequenas no caso $\mathcal{C}(\tau) \geq 0$, em mais detalhes. Podemos escrever

$$S(x) = \sum_{j=1}^{K} \hat{\beta} B_j(x)$$

onde $B_1, \ldots, B_K$ são bases B-*splines*. Assim, precisamos somente encontrar os coeficientes $\hat{\beta} = (\hat{\beta}_1, \ldots, \hat{\beta}_K)^T$. Pela expansão de S nas bases podemos reescrever a minimização como segue:

$$\begin{aligned} \text{Minimize:} \quad & (S - B\beta)^T (S - B\beta) \\ \text{sujeito à} \quad & \beta \geq 0 \end{aligned} \tag{3.13}$$

onde $B_{ij} = B_j(x_i)$ uma matriz $n \times K$.

O sistema em (3.13) foi solucionado, sob o critério de mínimos quadrados ordinários não-negativos, usando o algoritmo de Lawson-Hanson (veja Mullen e Stokkum, 2007). O pacote *nnls*, que está implementado no programa *R-Gui*, resolve este sistema pela função $nnls(B, S)$; onde B indica uma matriz numérica com n linhas e K colunas e S um vetor numérico de comprimento n. Note que o sistema (3.13) também pode ser usado para grandes amostras quando $\mathcal{C}(\tau) \geq 0$. Com isso todas as estimativas das funções de covariância foram obtidas usando o mesmo algoritmo. Os programas para a estimação das funções de covariância deste trabalho se encontram no Apêndice B.

Na próxima seção obteremos os cálculos para a transformada de Fourier da B-*spline*.

3.5 Transformada de Fourier da B-*spline*

Seja $u(x)$ uma função (Lebesgue-mensurável) de $x \in \mathbb{R}$, a norma L^2 de u é um número real infinito ou não negativo,

$$\|u\| = \left[\int_{-\infty}^{\infty} |u(x)|^2 dx\right]^{1/2}. \tag{3.14}$$

O símbolo L^2 denota o conjunto para todas as funções no qual esta integral é finita:

$$L_2 = \{u : \|u\| < +\infty\}. \tag{3.15}$$

Para qualquer $u \in L^2$, a transformada de Fourier de u é a função $\hat{u}(\omega)$ definida por

$$\hat{u}(\omega) = (\mathcal{F}u)(\omega) = \int_{-\infty}^{\infty} e^{-i\omega x} u(x) dx, \tag{3.16}$$

onde $i = \sqrt{-1}$.

A seguir, mostraremos um resultado de mudança de variável, que será útil mais adiante no cálculo da transformada de Fourier da B-*spline*. Seja uma função $h(x) = g(x-\xi) \overset{\mathcal{F}}{\longmapsto} ?$. A ideia é obter a transformada de Fourier de $h(x)$. Assim,

$$\begin{aligned}
\hat{h}(\omega) &= \int_{-\infty}^{\infty} h(x) e^{-ix\omega} dx \\
&= \int_{\mathbb{R}} g(x-\xi) e^{-ix\omega} dx \quad (\text{façamos, } u = x-\xi,\ du = dx) \\
&= \int_{\mathbb{R}} g(u) e^{-i(u+\xi)} du \\
&= e^{-i\xi\omega} \int_{\mathbb{R}} g(u) e^{-iu\omega} du \\
&= e^{-i\xi\omega} \hat{g}(\omega).
\end{aligned} \tag{3.17}$$

3.5.1 Transformada de Fourier da B-*spline* com $j = 0$

Seja $B_{j,k}(x)$ a B-*splines* de Curry e Schoenberg com nós $t_j < t_{j+1} < \ldots < t_{j+k}$ ($j \in \mathbb{Z},\ k = 1, 2, \ldots$), isto é, $B_{j,k}(x)$ é dada pela fórmula

$$B_{j,k}(x) = kq_k(\bullet, x)[t_j, t_{j+1}, \ldots, t_{j+k}]$$

Sem perda de generalidade trabalhamos com o índice j igual a zero. Assim,

$$B_{0,k}(x) = kq_k(\bullet, x)[t_0, t_1, \ldots, t_k] \tag{3.18}$$

em que,

$$q_k(t, x) = (t - x)_+^{k-1} = \begin{cases} (t-x)^{k-1}, & \text{se } t \geq x \\ 0 \qquad\quad , & \text{se } t < x. \end{cases}$$

A $k-$ésima diferença dividida dada por

$$f[t_0, t_1, \ldots, t_k] = \sum_{\mu=0}^{k} \frac{f(t_\mu)}{\phi'(t_\mu)}$$

em que $\phi(t) = (t - t_0) \ldots (t - t_k)$. Assim, podemos escrever a B-*spline*

$$B_{0,k}(x) = k \sum_{\mu=0}^{k} \frac{q_k(t_\mu, x)}{\phi'(t_\mu)}.$$

Logo, $B_{0,k}(x) \xrightarrow{\mathcal{F}} \hat{B}_{0,k}(\omega)$ e,

$$\begin{aligned} \hat{B}_{0,k}(\omega) &= \mathcal{F}(B_{0,k})(\omega) \\ &= \int_{-\infty}^{\infty} e^{-i\omega x} B_{0,k}(x) dx \\ &= \int_{-\infty}^{\infty} e^{-i\omega x} k \sum_{\mu=0}^{k} \frac{q_k(t_\mu, x)}{\phi'(t_\mu)} dx \\ &= k \sum_{\mu=0}^{k} \frac{\hat{q}_k(t_\mu, \omega)}{\phi'(t_\mu)}. \end{aligned} \tag{3.19}$$

Basta calcular a transformada de Fourier de q_k. Agora definimos uma função auxiliar $g = g(x)$,

$$g(x) = x_+^{k-1} = \begin{cases} x^{k-1}, & \text{se } x \geq 0 \\ 0 \quad , & \text{se } x < 0. \end{cases}$$

Note que $g(t-x) = q_k(t,x)$. Vamos obter a transformada de Fourier da g,

$$\hat{g}(\omega) = \int_0^\infty e^{-i\omega x} x^{k-1} dx = \frac{(k-1)!}{(i\omega)^k}$$

como $h(x) = g(t-x)$ tem transformada de Fourier igual a $\hat{h}(\omega) = e^{-i\omega t}\hat{g}(-\omega)$ então a transformada de Fourier da q_k (na variável x) é dada por

$$\hat{q}_k(t,\omega) = \hat{h}(\omega) = e^{-i\omega t}\hat{g}(-\omega),$$

ou seja,

$$\begin{aligned} \hat{q}_k(t,\omega) &= e^{-i\omega t}\frac{(k-1)!}{(-i\omega)^k} \\ &= e^{-i\omega t}\frac{(k-1)!}{(-1)^k(i\omega)^k} \\ &= e^{-i\omega t}\frac{(k-1)!}{(-1)^k(i\omega)^k} \end{aligned} \tag{3.20}$$

finalmente a transformada de Fourier da B-*splines*

$$\begin{aligned} \hat{B}_{0,k}(\omega) &= k\sum_{\mu=0}^{k} e^{-i\omega t_\mu}\frac{(k-1)!}{(-1)^k(i\omega)^k}/\phi'(t_\mu) \\ &= \frac{k!}{(-1)^k(i\omega)^k}\sum_{\mu=0}^{k}\frac{e^{-i\omega t_\mu}}{\phi'(t_\mu)}. \end{aligned} \tag{3.21}$$

Neste trabalho foram usados nós equidistantes, $t_\mu = \mu h$. Para estes nós, $\phi'(t_\mu) = (-1)^{k-\mu} h^k \mu!(k-\mu)!$,

$$\begin{aligned}
\hat{B}_{0,k}(\omega) &= \frac{k!}{(-1)^k (i\omega)^k} \sum_{\mu=0}^{k} \frac{e^{-i\omega(\mu h)}}{(-1)^{k-\mu} h^k \mu!(k-\mu)!} \\
&= \frac{1}{(-1)^k (i\omega h)^k} \sum_{\mu=0}^{k} \frac{k! e^{-i\omega\mu h}}{(-1)^{k-\mu} \mu!(k-\mu)!} \\
&= \frac{1}{(-1)^k (i\omega h)^k} \sum_{\mu=0}^{k} \binom{k}{\mu} e^{-i\omega h\mu} (-1)^{k-\mu} \\
&= \frac{1}{(-1)^k (i\omega h)^k} \left(e^{-i\omega h} - 1\right)^k \\
&= \frac{1}{(-1)^k} \left(\frac{e^{-i\omega h} - 1}{i\omega h}\right)^k \\
&= \left[\frac{(-1)(-(1 - e^{-i\omega h}))}{i\omega h}\right]^k \\
&= \left(\frac{1 - e^{-i\omega h}}{i\omega h}\right)^k
\end{aligned} \tag{3.22}$$

note que, $1 - e^{-i\omega h} = e^{-\frac{i\omega h}{2}} \left(\frac{e^{\frac{i\omega h}{2}} - e^{-\frac{i\omega h}{2}}}{2i}\right) 2i = e^{-\frac{i\omega h}{2}} \text{sen}\left(\frac{\omega h}{2}\right) 2i$. Substituindo em (3.22),

$$\hat{B}_{0,k}(\omega) = \left[\frac{e^{-\frac{i\omega h}{2}} \text{sen}(\omega h/2) 2i}{i\omega h}\right]^k = e^{-\frac{ik\omega h}{2}} \left[\frac{\text{sen}(\omega h/2)}{\omega h/2}\right]^k.$$

Note que a transformada de Fourier da B-*spline* é positiva, uma vez que $k = 4$ e $\omega \in (0, \infty)$.

Em outro caso, quando B-*spline* central, $t_\mu = -\frac{k}{2} + \mu$, $(\mu = 0, 1, \ldots, k)$. Para estes

nós, $\phi'(t_\mu) = (-1)^{k-\mu}\mu!(k-\mu)!$,

$$\begin{aligned}
\hat{B}_{0,k}(\omega) &= \frac{k!}{(-1)^k(i\omega)^k}\sum_{\mu=0}^{k}\frac{e^{-i\omega(-\frac{k}{2}+\mu)}}{(-1)^{k-\mu}\mu!(k-\mu)!} \\
&= \frac{e^{i\omega k/2}}{(-1)^k(i\omega)^k}\sum_{\mu=0}^{k}\frac{k!e^{-i\omega\mu}}{(-1)^{k-\mu}\mu!(k-\mu)!} \\
&= \frac{e^{i\omega k/2}}{(-1)^k(i\omega)^k}\sum_{\mu=0}^{k}\binom{k}{\mu}e^{-i\omega}(-1)^{k-\mu} \\
&= \frac{e^{i\omega k/2}}{(-1)^k(i\omega)^k}\left(e^{-i\omega}-1\right)^k
\end{aligned} \tag{3.23}$$

note que, $e^{-i\omega} - 1 = -e^{-\frac{i\omega}{2}}\left(\dfrac{e^{\frac{i\omega}{2}} - e^{-\frac{i\omega}{2}}}{2i}\right)2i = -e^{-\frac{i\omega}{2}}\operatorname{sen}\left(\dfrac{\omega}{2}\right)2i.$

Substituindo em (3.23),

$$\begin{aligned}
\hat{B}_{0,k}(\omega) &= -\frac{e^{\frac{i\omega k}{2}}}{(-1)^{k+1}(i\omega)^k}e^{-\frac{i\omega k}{2}}\left[\operatorname{sen}\left(\frac{\omega}{2}\right)\right]^k(2i)^k \\
&= \frac{1}{(-1)^{k+1}}\left(\frac{\operatorname{sen}(\omega/2)}{\omega/2}\right)^k
\end{aligned} \tag{3.24}$$

3.5.2 Transformada de Fourier para B-*spline* no caso geral

Novamente, como na subseção anterior, seja $B_{j,k}(x)$ a B-*spline* de Curry e Schoenberg com nós $t_j < t_{j+1} < \ldots < t_{j+k}$ $(j \in \mathbb{Z},\ k = 1, 2, \ldots)$, isto é, $B_{j,k}(x)$ é dada pela fórmula, para o caso em que j assume todos os valores

$$\begin{aligned}
B_{j,k}(x) &= kq_k(\bullet, x)[t_j, t_{j+1}, \ldots, t_{j+k}] \\
&= k\sum_{\mu=0}^{k}\frac{q_k(t_{j+\mu}, x)}{\phi'(t_{j+\mu})}.
\end{aligned} \tag{3.25}$$

Logo,

$$\hat{B}_{j,k}(\omega) = k \sum_{\mu=0}^{k} \frac{\hat{q}_k(t_{j+\mu}, \omega)}{\phi'(t_{j+\mu})}$$

como

$$\hat{q}_k(t_{j+\mu}, \omega) = \frac{e^{-i\omega t_{j+\mu}}(k-1)!}{(-1)^k (i\omega)^k}$$

então

$$\hat{B}_{j,k}(\omega) = k \sum_{\mu=0}^{k} \frac{e^{-i\omega t_{j+\mu}}(k-1)!}{(-1)^k (i\omega)^k} \frac{1}{\phi'(t_{j+\mu})}$$

Para nós equidistantes $t_m = mh$ então $t_{j+\mu} = (j+\mu)h$ $\phi'(t_m) = (-1)^{k-m} h^k m!(k-m)!$ então $\phi'(t_{j+\mu}) = (-1)^{k-j-\mu} h^k (j+\mu)!(k-j-\mu)!$

$$
\begin{aligned}
\hat{B}_{j,k}(\omega) &= \frac{1}{(-1)^k(i\omega)^k}\sum_{\mu=0}^{k}\frac{e^{-i\omega(j+\mu)h}k!}{(-1)^{k-j-\mu}h^k(j+\mu)!(k-j-\mu)!} \\
&= \frac{1}{(-1)^k(i\omega)^k}\sum_{\mu=0}^{k}\frac{e^{-i\omega jh}e^{-i\mu h}k!}{(-1)^{k-j-\mu}h^k(j+\mu)!(k-j-\mu)!} \\
&= \frac{1}{(-1)^k(i\omega h)^k}e^{-i\omega jh}\sum_{\mu=0}^{k}\binom{k}{j+\mu}e^{-i\omega h\mu}(-1)^{k-j-\mu}, \quad m=j+\mu \\
&= \frac{1}{(-1)^k(i\omega h)^k}e^{-i\omega jh}\sum_{m=j}^{j+k}\binom{k}{m}e^{-i\omega h(m-j)}(-1)^{k-m} \\
&= \frac{1}{(-1)^k(i\omega h)^k}e^{-i\omega jh}\sum_{m=j}^{j+k}\binom{k}{m}e^{-i\omega hm}e^{i\omega jh}(-1)^{k-m} \\
&= \frac{1}{(-1)^k(i\omega h)^k}\sum_{m=j}^{j+k}\binom{k}{m}e^{-i\omega hm}(-1)^{k-m} \qquad \text{(agora vamos completar a soma)} \\
&= \frac{1}{(-1)^k(i\omega h)^k}\left\{\sum_{m=0}^{j-1}\binom{k}{m}e^{-i\omega hm}(-1)^{k-m}-\sum_{m=0}^{j-1}\binom{k}{m}e^{-i\omega hm}(-1)^{k-m}+\right. \\
&\quad \left.+\sum_{m=j}^{j+k}\binom{k}{m}e^{-i\omega hm}(-1)^{k-m}\right\} \\
&= \frac{1}{(-1)^k(i\omega h)^k}\left\{\sum_{m=0}^{j+k}\binom{k}{m}e^{-i\omega hm}(-1)^{k-m}-\sum_{m=0}^{j-1}\binom{k}{m}e^{-i\omega hm}(-1)^{k-m}\right\}.
\end{aligned}
\tag{3.26}
$$

Veja que em $\sum_{m=0}^{j-1}(.)$, se $j=0$ temos o caso $\hat{B}_{0,k}$. Ainda, como $\sum_{m=0}^{j-1}(.)$ é zero para $j=0$ então vamos multiplicá-lo por um artifício

$$
\beta(j)=\begin{cases} 1, & \text{se } j\neq 0 \\ 0, & \text{se } j=0. \end{cases}
$$

Assim,

$$
\hat{B}_{j,k}(\omega) = \frac{1}{(-1)^k(i\omega h)^k}\left\{\sum_{m=0}^{j+k}\binom{k}{m}e^{-i\omega hm}(-1)^{k-m}-\beta(j)\sum_{m=0}^{j-1}\binom{k}{m}e^{-i\omega hm}(-1)^{k-m}\right\}.
$$

Sabe-se que $\sum_{m=0}^{C} \binom{k}{m} e^{-i\omega hm}(-1)^{k-m} = (e^{-i\omega h}-1)^{j+k}$. Logo,

$$\hat{B}_{j,k}(\omega) = \frac{1}{(-1)^k(i\omega h)^k}\left\{(e^{-i\omega h}-1)^{j+k} - \beta(j)(e^{-i\omega h}-1)^{j-1}\right\}.$$

Portanto,

$$\begin{aligned}\hat{B}_{j,k}(\omega) = \frac{1}{(-1)^k(i\omega h)^k}\Big\{&\left[(-1)(2i)e^{-i\omega h/2}\text{sen}(\omega h/2)\right]^{j+k} \\ &-\beta(j)\left[(-1)(2i)e^{-i\omega h/2}\text{sen}(\omega h/2)\right]^{j-1}\Big\}.\end{aligned}$$

Chamemos, $\psi(\omega) = (-1)(2i)e^{-i\omega h/2}\text{sen}(\omega h/2)$ então

$$\begin{aligned}\hat{B}_{j,k}(\omega) &= \frac{1}{(-1)^k(i\omega h)^k}(-1)^k(2i)^k e^{-i\omega hk/2}\text{sen}^k(\omega h/2)\psi^j(\omega) - \frac{(-1)^k}{(i\omega h)^k}\psi^{j-1}(\omega)\beta(j) \\ &= e^{-i\omega hk/2}\left(\frac{\text{sen}(\omega h/2)}{\omega h/2}\right)^k \psi^j(\omega) + \frac{(-1)^{k+1}}{(i\omega h)^k}\psi^{j-1}(\omega)\beta(j) \\ &= \hat{B}_{0,k}(\omega)\psi^j(\omega) + \frac{(-1)^{k+1}}{(i\omega h)^k}\psi^{j-1}(\omega)\beta(j).\end{aligned} \tag{3.27}$$

Capítulo 4

Estimação do Número de Funções Bases

Como foi dito no Capítulo 2, uma dificuldade ao trabalhar com B-*splines* é selecionar o número e as posições dos nós (*knots*, em inglês). Em modelos de regressão não-paramétrico, geralmente o que se faz é fixar o número de funções bases a priori. O número de funções bases K funciona também como um parâmetro de suavização, pois quando K aumenta menos suave será a estimativa da função. Assim é preciso desenvolver um critério para encontrar um K ótimo.

Dias (1999) propôs um critério de parada adaptativo para regressão não-paramétrica via *splines* híbridos. A ideia é fornecer um procedimento que estime o número de funções bases iterativamente. O algoritmo aumenta o número de funções bases por um até satisfazer um critério de parada que é baseado numa distância relacionada a distância de Hellinger entre duas estimativas consecutivas da afinidade parcial. Isto é, encontrar a dimensão do espaço de *splines* cúbicos, onde estamos procurando pela aproximação da solução de mínimos quadrados.

Para isto, define-se a seguinte transformação. Dada qualquer função em $\mathcal{W}_2^2[a,b]$, o conjunto de todas as funções de quadrado integrável num certo intervalo $[a,b]$. Consideremos $g \in \mathcal{W}_2^2[a,b]$ e a transformação

$$t_g = \frac{g^2}{\int g^2}, \tag{4.1}$$

de forma que $t_g \geq 0$ e $\int t_g = 1$, satisfazendo as condições de uma função densidade. Para quaisquer funções $f, g \in \mathcal{W}_2^2[a,b]$, definimos uma pseudo distância aproximadamente relacionada ao quadrado da distância de Hellinger,

$$H^2(f,g) = \int \left(\sqrt{t_f} - \sqrt{t_g}\right)^2 = 2(1 - \xi(f,g)), \tag{4.2}$$

onde

$$\xi(f,g) = \int \sqrt{t_f t_g} = \int \sqrt{\frac{f^2 g^2}{\int f^2 \int g^2}} = \int \frac{|fg|}{\sqrt{\int f^2 \int g^2}} \tag{4.3}$$

é a afinidade entre f e g. Não é difícil ver que $0 \leq \xi(f,g) \leq 1, \forall f, g \in \mathcal{W}_2^2[a,b]$ e que $H^2(f,g)$ é mínimo quando $\xi(f,g) = 1$.

Tomamos $f(\cdot) = \pi_K$ e $g(\cdot) = \pi_{K+1}$. Aumentando o número de funções bases por um, o procedimento parará quando $\pi_K \approx \pi_{K+1}$, no sentido da afinidade parcial,

$$\int \frac{|\pi_K \pi_{K+1}|}{\sqrt{\int \pi_K^2 \int \pi_{K+1}^2}} \approx 1. \tag{4.4}$$

O algoritmo abaixo é uma tentativa de implementar esta ideia.

Algoritmo

1. *Seja K_0 o número inicial de funções bases.*

2. *Calculamos π_{K_0}.*

3. *Incrementamos o número de funções bases por um e repetimos o passo* (2) *para obter* π_{K_0+1}.

4. *Dado um número real* $\delta > 0$, *o procedimento para se* $\xi(\pi_{K_0}, \pi_{K_0+1}) < 1 - \delta$ *ou quando* K *alcançar* n *(tamanho amostral). O número delta pode ser determinado empiricamente de acordo com uma distância particular* $\xi(\cdot, \cdot)$.

Ressaltamos que para a distância de afinidade a integral será aproximada por uma soma multiplicada pelo comprimento dos intervalos igualmente espaçados entre as observações.

Seria útil obter a distribuição da afinidade entre a função de covariância verdadeira e a estimativa produzida pelo método de B-*splines*. Estudos anteriores por Dias (1999) mostraram que é uma densidade unimodal empírica com suporte em $[0, 1]$, simétrica a esquerda, sugerindo um modelo beta. Embora, aqui muitos dos histogramas apresentaram um comportamento muito similar a de uma distribuição normal. Por isso, resolvemos estimar a distribuição da afinidade usando tanto a distribuição beta quando a distribuição normal.

A densidade de uma distribuição beta e de uma distribuição normal foram calculadas para a afinidade. A estimativa de densidade por *kernel* também foi calculada. Detalhes sobre o estimador de densidade por *kernel* podem ser encontrados em Silverman (1986). Lembremos que a função densidade de uma variável aleatória $X \sim Beta(a, b)$ é dada por:

$$f(x) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1}, \tag{4.5}$$

para $0 < x < 1$ e $a, b > 0$.

Muitos experimentos numéricos mostraram que o máximo da afinidade e a estabilização da afinidade parcial coincidem. Isto significa que aumentando K arbitrariamente não somente aumenta o custo computacional como também não fornece um melhor ajuste da curva.

Simulações foram realizadas a fim de exemplificar o comportamento da afinidade e da afinidade parcial. As figuras a seguir mostram a distribuição empírica da afinidade usando uma estimativa não-paramétrica da densidade pelo método *kernel* e dois paramétricos usando o modelo beta e o modelo normal, cujos os parâmetros foram estimados usando o método dos momentos.

O teste de Kolmogorov-Smirnov é comumente utilizado para verificar se duas amostras aleatórias independentes possuem a mesma distribuição de probabilidade. Sua estatística do teste está baseada na maior distância vertical entre as funções de distribuição empíricas das amostras. Enquanto o teste de Shapiro-Wilk testa a hipótese de que a amostra vem de uma população normalmente distribuída (veja detalhes em Conover, 1999). Vamos usar estes dois testes para testar se a distribuição da afinidade segue uma distribuição normal, pois em várias estimativas da densidade da afinidade se nota um comportamento semelhante a uma distribuição normal. Em todos os casos consideramos o nível de significância de 5%.

As Figuras 4.2, 4.5, 4.8, 4.11, 4.14, 4.17, 4.20 e 4.23 mostram a função de covariância espacial verdadeira com sua função de covariância estimada.

As Figuras 4.3, 4.6, 4.9, 4.12, 4.15, 4.18, 4.21 e 4.24 mostram os gráficos dos valores da afinidade e afinidade parcial.

As Figuras 4.4, 4.7, 4.10, 4.13, 4.16, 4.19, 4.22 e 4.25 mostram os histogramas da

afinidade com as estimativas das densidades pelo modelo normal, beta e *kernel*. A hipótese de que a distribuição da afinidade seja normal foi rejeitada, utilizando o teste de Kolmogorov-Smirnov para: modelo Cauchy ($\sigma^2 = 1$; $\phi = 3$; $\kappa = 0,6$), modelo exponencial potência ($\sigma^2 = 1$; $\phi = 2$; $\kappa = 1,5$), modelo gaussiano ($\sigma^2 = 1$; $\phi = 2,5$), modelo gaussiano ($\sigma^2 = 1$; $\phi = 2,2$). Enquanto, a hipótese de que a distribuição da afinidade seja normal não foi rejeitada, utilizando o teste de Kolmogorov-Smirnov para: modelo exponencial potência ($\sigma^2 = 1$; $\phi = 0,5$; $\kappa = 1,5$), modelo exponencial potência ($\sigma^2 = 1$; $\phi = 0,4$; $\kappa = 1,5$), modelo expoêncial potência ($\sigma^2 = 1$; $\phi = 0,6$; $\kappa = 1$), modelo gaussiano ($\sigma^2 = 1$; $\phi = 0,8$), modelo exponencial potência ($\sigma^2 = 1$; $\phi = 0,8$; $\kappa = 1,5$).

Agora para o teste de Shapiro-Wilk a hipótese de que os dados seguem uma distribuição normal foi rejeitada para: modelo exponencial potência ($\sigma^2 = 1$; $\phi = 0,6$; $\kappa = 1$), modelo exponencial potência ($\sigma^2 = 1$; $\phi = 0.8$; $\kappa = 1,5$), modelo gaussiano ($\sigma^2 = 1$; $\phi = 2,2$), modelo gaussiano ($\sigma^2 = 1$; $\phi = 2,5$), modelo exponencial potência ($\sigma^2 = 1$; $\phi = 2$; $\kappa = 1,5$), modelo Cauchy ($\sigma^2 = 1$; $\phi = 3$; $\kappa = 0,6$) modelo gaussiano ($\sigma^2 = 1$; $\phi = 0,8$). Enquanto, a hipótese de que os dados seguem uma distribuição normal não foi rejeitada para: modelo exponencial potência ($\sigma^2 = 1$; $\phi = 0,4$; $\kappa = 1,5$), modelo exponencial potência ($\sigma^2 = 1$; $\phi = 0,5$; $\kappa = 1,5$).

Ainda usando o teste de Kolmogorov-Smirnov de duas amostras, testou-se a hipótese que os dados da afinidade seguem uma distribuição beta. Em todos os casos houve evidência para não rejeitar essa hipótese, ao nível de significância de 5%. A estimativa de densidade da afinidade obtida utilizando os parâmetros estimados de uma beta parece ser bastante razoável para a afinidade, já que sabemos que $0 \leq \xi(f, g) \leq 1$.

As Tabelas 4.2 a 4.9 apresentam os parâmetros estimados de uma distribuição beta e uma distribuição normal usando os valores de afinidade obtidos.

O algoritmo retorna o número de bases ótima. Mas agora surge uma questão: como escolher o valor de δ? Queremos que a afinidade entre a função de covariância verdadeira e a função de covariância estimada esteja próxima de 1. Assim, supondo que os dados da afinidade seguem uma distribuição beta, com seus parâmetros estimados pelo método dos momentos, calculamos um ponto de corte, d, que deixa a probabilidade de pegarmos um valor da afinidade a $0,01$ na cauda superior da distribuição dos dados.

Verificou-se o comportamento de d utilizando o procedimento *bootstrap* de reamostragem com reposição. Foram geradas 2000 amostras *bootstrap*, em seguida calculou-se d para cada amostra, a partir da distribuição beta com seus parâmetros estimados pelo método dos momentos. Abaixo, temos o *box plot* dos 2000 pontos de cortes obtidos. No procedimento *bootstrap*, amostras artificiais são obtidas a partir das amostras verdadeiras através de reamostragem com reposição.O procedimento *bootstrap* de reamostragem constitui de método computacionalmente intensivo que pode ser utilizado para testar a hipótese de existência de alguma relação estocástica específica entre dois conjuntos de variáveis aleatórias. Mais sobre o procedimento *bootstrap* veja Efron e Tibshirani (1993).

Tabela 4.1: Estatísticas resumos dos valores dos pontos de corte na reamostragem *bootstrap* de tamanho 2000.

	Média	Desvio Padrão	Viés
Afinidade	$0,9999874$	$7,96 \times 10^{-15}$	$-5,84 \times 10^{-9}$

Verifica-se na Tabela 4.1 que os desvio padrão para o ponto de corte está próximo de zero, fazendo com que o valor do ponto de corte fique concentrado em torno da média. Nota-se que o viés da média também está próximo de zero.

Observa-se na Figura 4.1 que os dados de ponto de corte estão bem comportado com

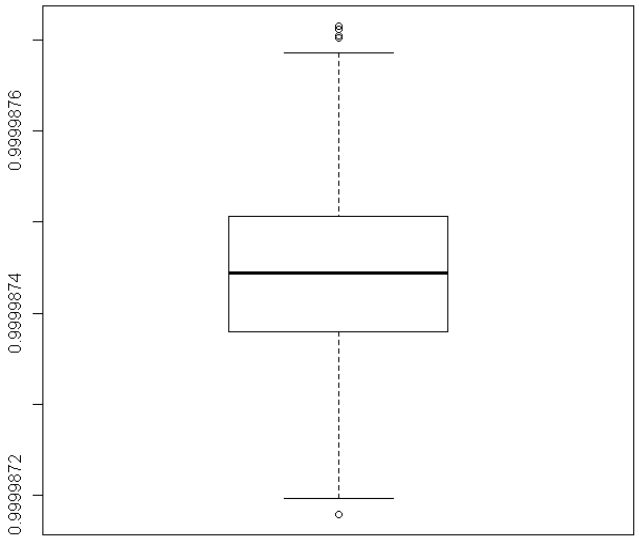


Figura 4.1: *Box plot* dos pontos de corte com probabilidade $0,01$ na extrema direita da distribuição da afinidade.

certa simetria em torno da média, mas com alguns valores atípicos. Contudo, essa seria a forma de escolher o valor de δ, obtendo o ponto de corte, d, a uma certa probabilidade na cauda superior da distribuição da afinidade, em seguida calculamos seu valor como $\delta = 1 - d$. Os programas para o estudo da afinidade e obtenção do ponto de corte se encontram no Apêndice A.

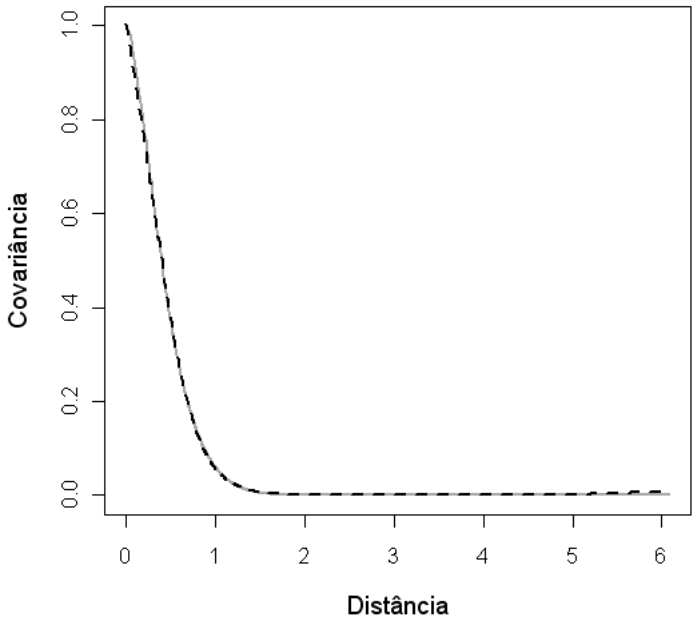
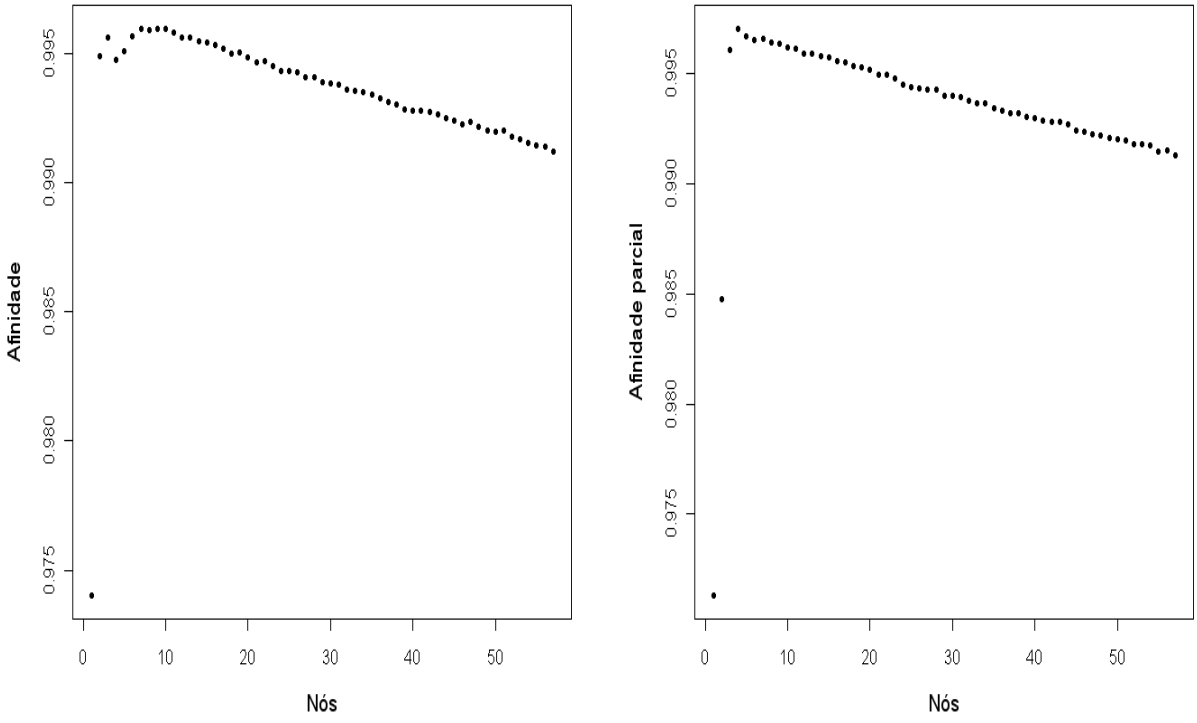


Figura 4.2: Função de covariância exponencial potência com parâmetros: $\sigma^2 = 1$, $\phi = 0,5$ e $\kappa = 1,5$. A linha sólida cinza é a função verdadeira e a linha pontilhada a função estimada.



(a) (b)

Figura 4.3: Mil replicações com $n = 100$ e 56 bases: (a) afinidade e (b) afinidade parcial.

Tabela 4.2: Parâmetros das distribuições beta e normal estimados a partir dos valores da afinidade obtidos: modelo exponencial potência ($\sigma^2 = 1$, $\phi = 0,5$, $\kappa = 1,5$).

	Beta		Normal	
	a	b	Média	Desvio Padrão
Afinidade	93154	611,2024	0,9934815642	0,0002626704

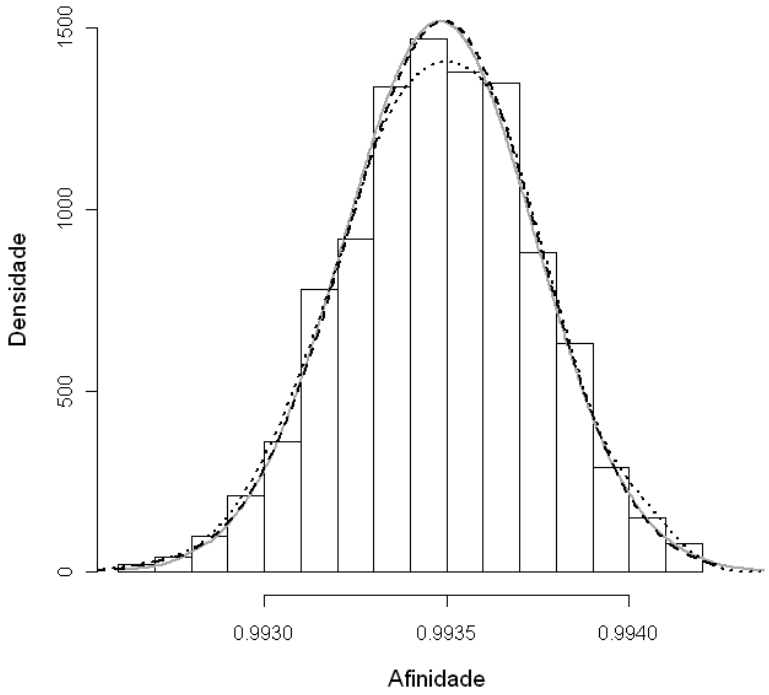


Figura 4.4: Estimativas da densidade da afinidade baseadas em mil replicações, $n = 100$: linha sólida cinza - modelo normal, linha tracejada - modelo beta e a linha pontilhada - método *kernel*.

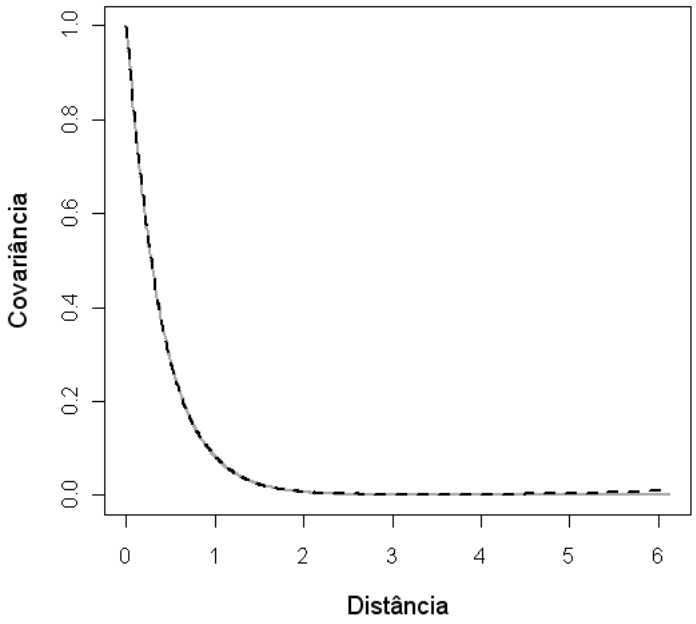
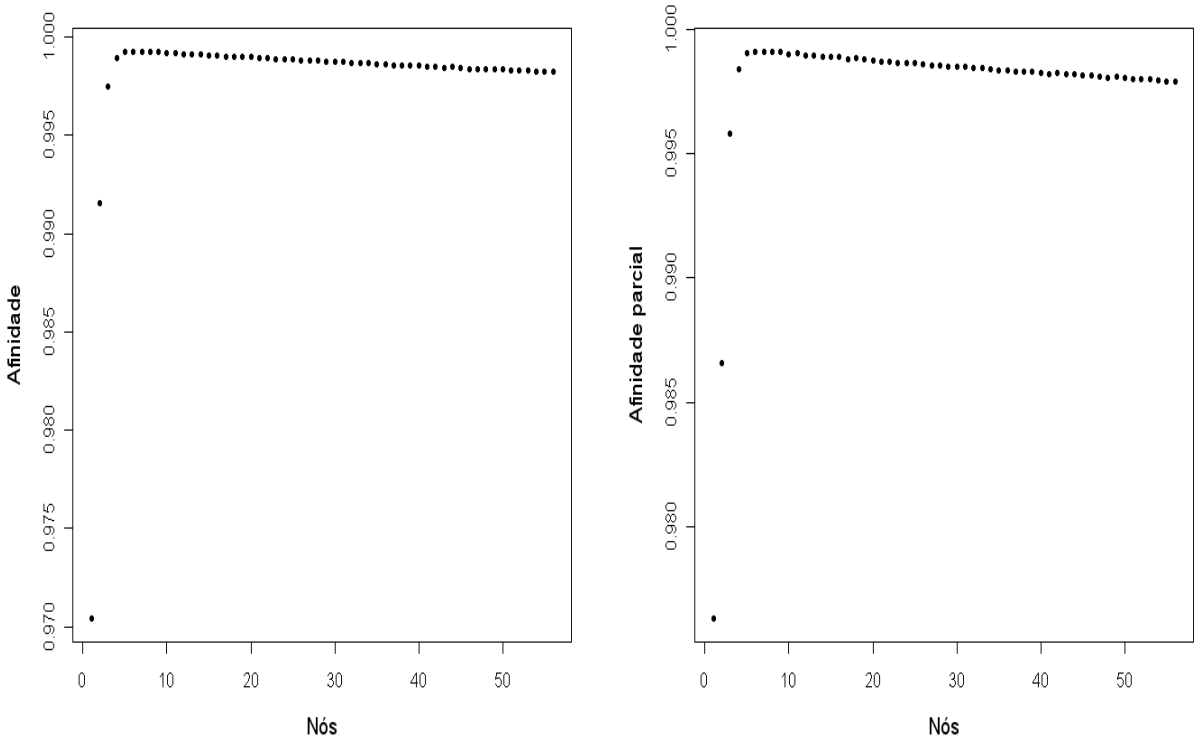


Figura 4.5: Função de covariância exponencial potência com parâmetros: $\sigma^2 = 1$, $\phi = 0,4$ e $\kappa = 1,5$. A linha sólida cinza é a função verdadeira e a linha pontilhada a função estimada.



(a) (b)

Figura 4.6: Mil replicações com $n = 500$ e 56 bases: (a) afinidade e (b) afinidade parcial.

Tabela 4.3: Parâmetros das distribuições beta e normal estimados a partir dos valores da afinidade obtidos: modelo exponencial potência ($\sigma^2 = 1$, $\phi = 0,4$,$\kappa = 1,5$).

	Beta		Normal	
	a	b	Média	Desvio Padrão
Afinidade	513647,6	989,362	9,980776e-01	6,102968e-05

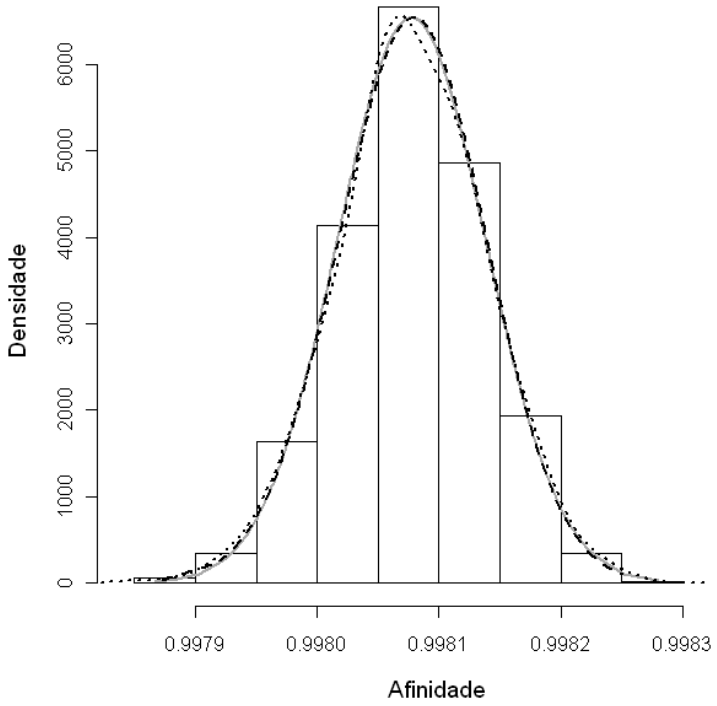


Figura 4.7: Estimativas da densidade da afinidade baseadas em mil replicações, $n = 500$: linha sólida cinza - modelo normal, linha tracejada - modelo beta e a linha pontilhada - método *kernel*.

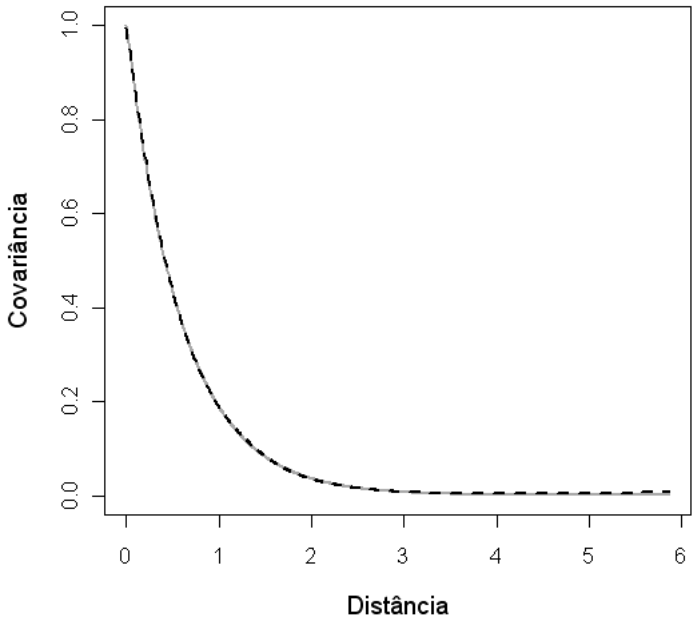
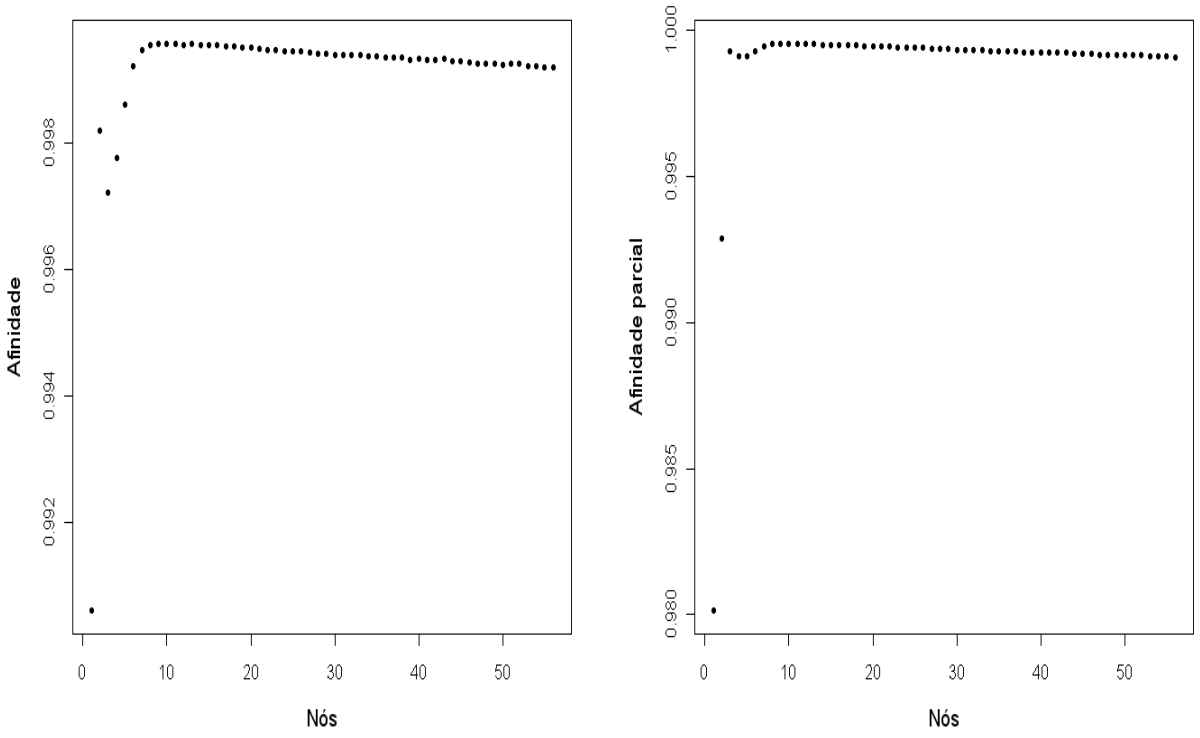


Figura 4.8: Função de covariância exponencial potência com parâmetros: $\sigma^2 = 1$, $\phi = 0,6$ e $\kappa = 1$. A linha sólida cinza é a função verdadeira e a linha pontilhada a função estimada.



(a) (b)

Figura 4.9: Mil replicações com $n = 1000$ e 56 bases: (a) afinidade e (b) afinidade parcial.

Tabela 4.4: Parâmetros das distribuições beta e normal estimados a partir dos valores da afinidade obtidos: modelo exponencial potência ($\sigma^2 = 1$, $\phi = 0,6$, $\kappa = 1$).

	Beta		Normal	
	a	b	Média	Desvio Padrão
Afinidade	975061,6	852,893	9,991261e-01	2,989704e-05

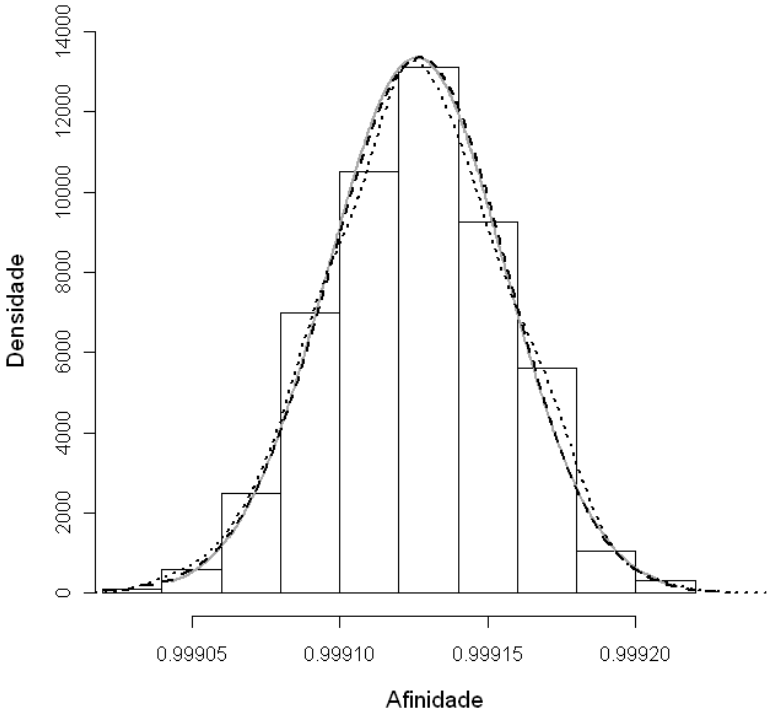


Figura 4.10: Estimativas da densidade da afinidade baseadas em mil replicações, $n = 1000$: linha sólida cinza - modelo normal, linha tracejada - beta e a linha pontilhada - método *kernel*.

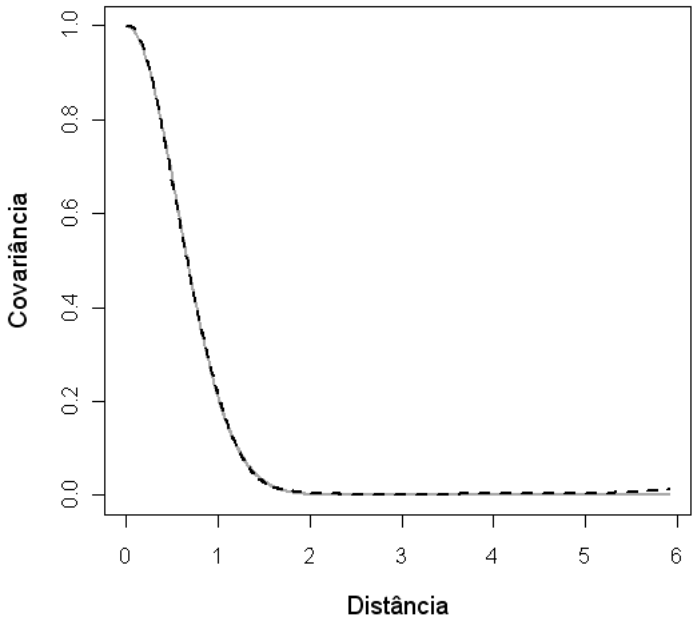
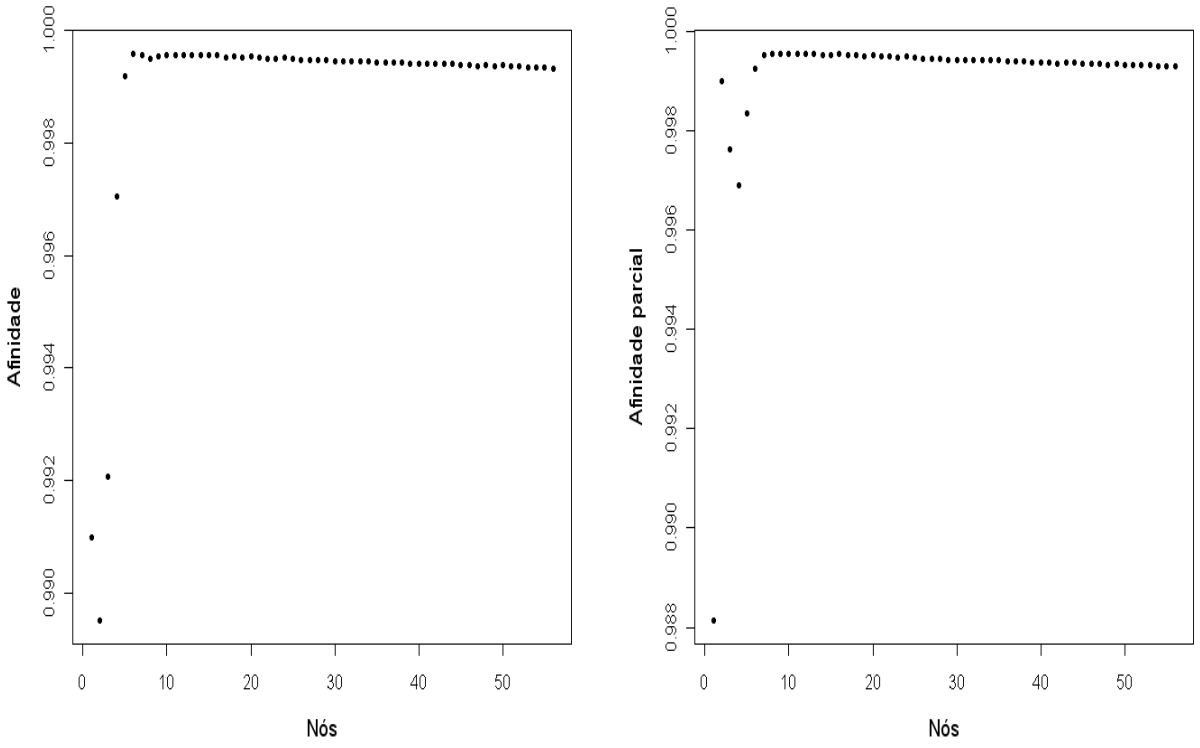


Figura 4.11: Função de covariância gaussiana com parâmetros: $\sigma^2 = 1$, $\phi = 0,8$. A linha sólida cinza é a função verdadeira e a linha pontilhada a função estimada.



(a) (b)

Figura 4.12: Mil replicações com $n = 1000$ e 56 bases: (a) afinidade e (b) afinidade parcial.

Tabela 4.5: Parâmetros das distribuições beta e normal estimados a partir dos valores da afinidade obtidos: modelo gaussiano ($\sigma^2 = 1$, $\phi = 0,8$).

	Beta		Normal	
	a	b	Média	Desvio Padrão
Afinidade	1054304	1101,300	9,989565e-01	3,141155e-05

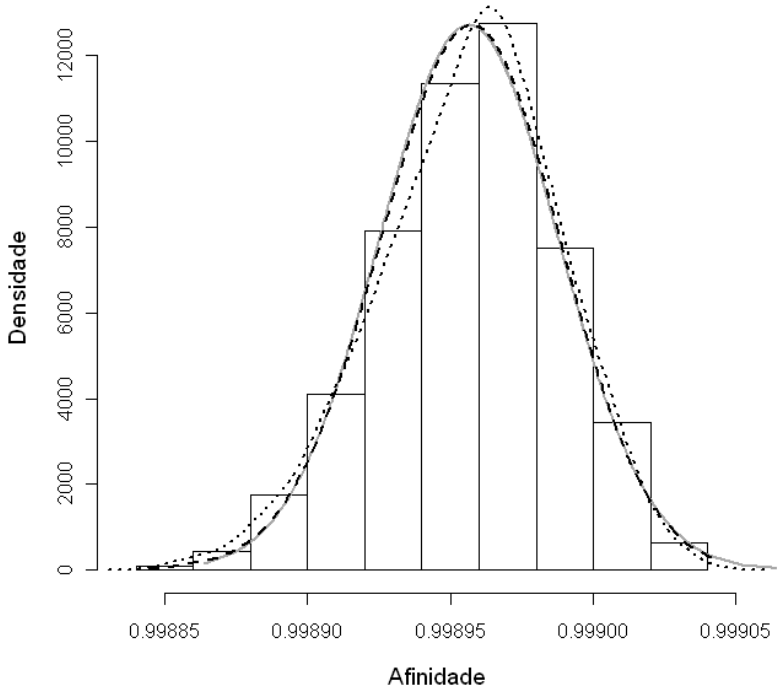


Figura 4.13: Estimativas da densidade da afinidade baseadas em mil replicações, $n = 1000$: linha sólida cinza - modelo normal, linha tracejada - modelo beta e a linha pontilhada - método *kernel*.

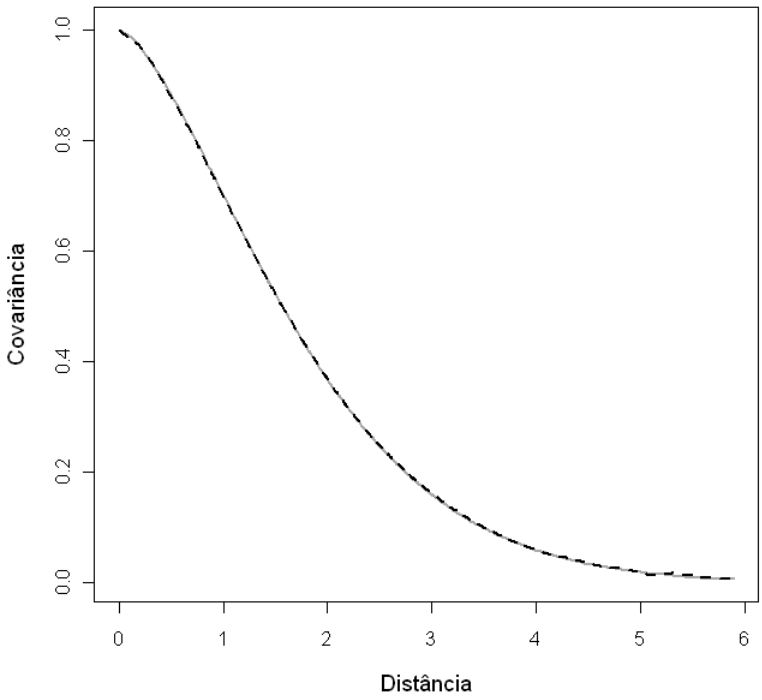
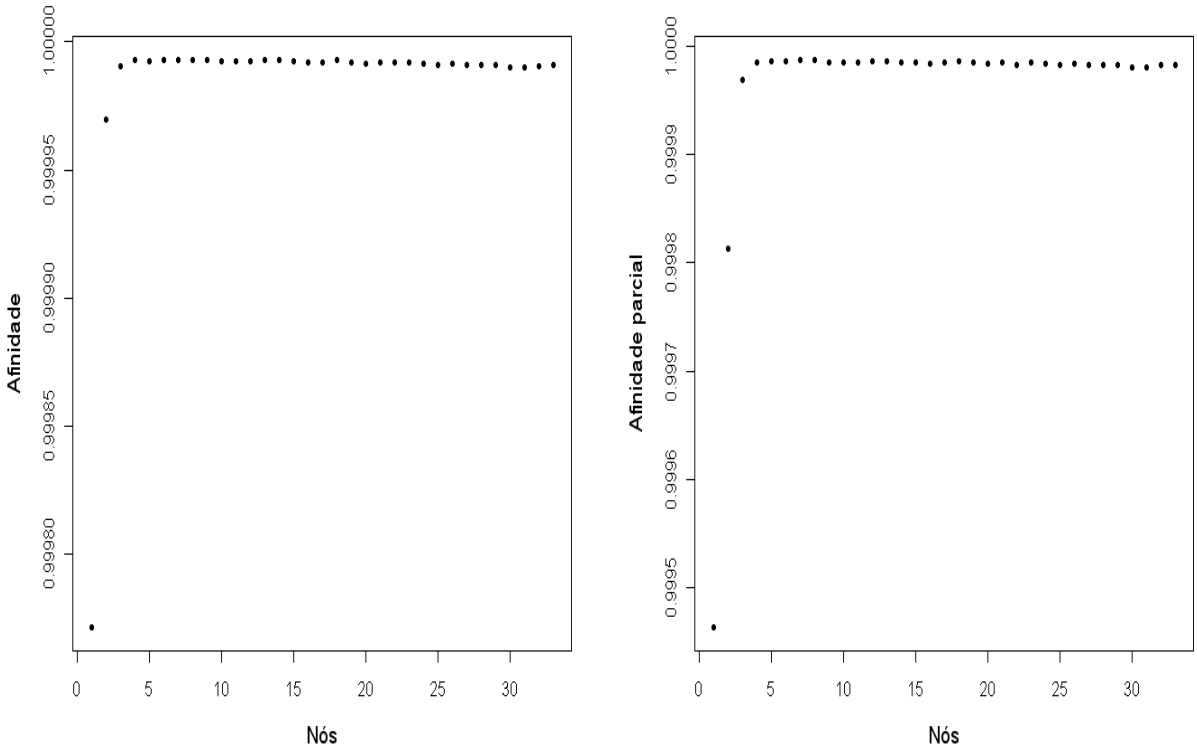


Figura 4.14: Função de covariância exponencial potência com parâmetros: $\sigma^2 = 1$, $\phi = 2$ e $\kappa = 1,5$. A linha sólida cinza é a função verdadeira e a linha pontilhada a função estimada.



(a) (b)

Figura 4.15: Quinhentas replicações com $n = 500$ e 33 bases: (a) afinidade e (b) afinidade parcial.

Tabela 4.6: Parâmetros das distribuições beta e normal estimados a partir dos valores da afinidade obtidos: modelo exponencial potência ($\sigma^2 = 1$, $\phi = 2$, $\kappa = 1,5$).

	Beta		Normal	
	a	b	Média	Desvio Padrão
Afinidade	8793947	136,0529	9,999845e-01	1,325028e-06

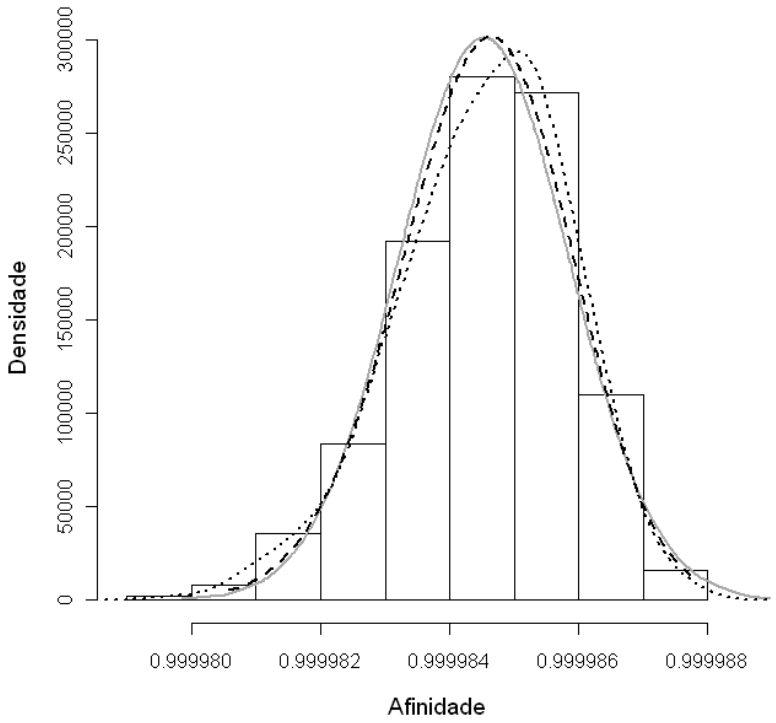


Figura 4.16: Estimativas da densidade da afinidade baseadas em 500 replicações, $n = 500$: linha sólida cinza - modelo normal, linha tracejada - modelo beta e a linha pontilhada - método *kernel*.

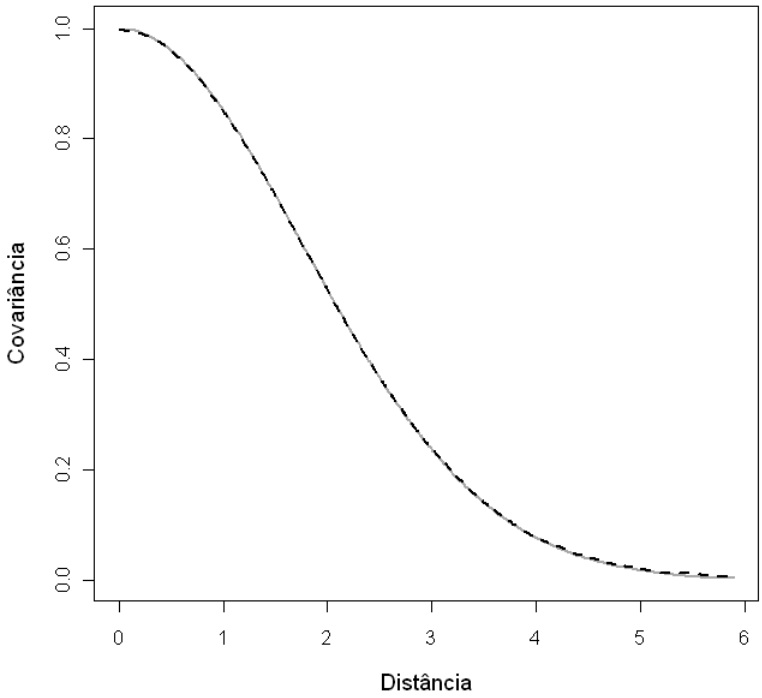
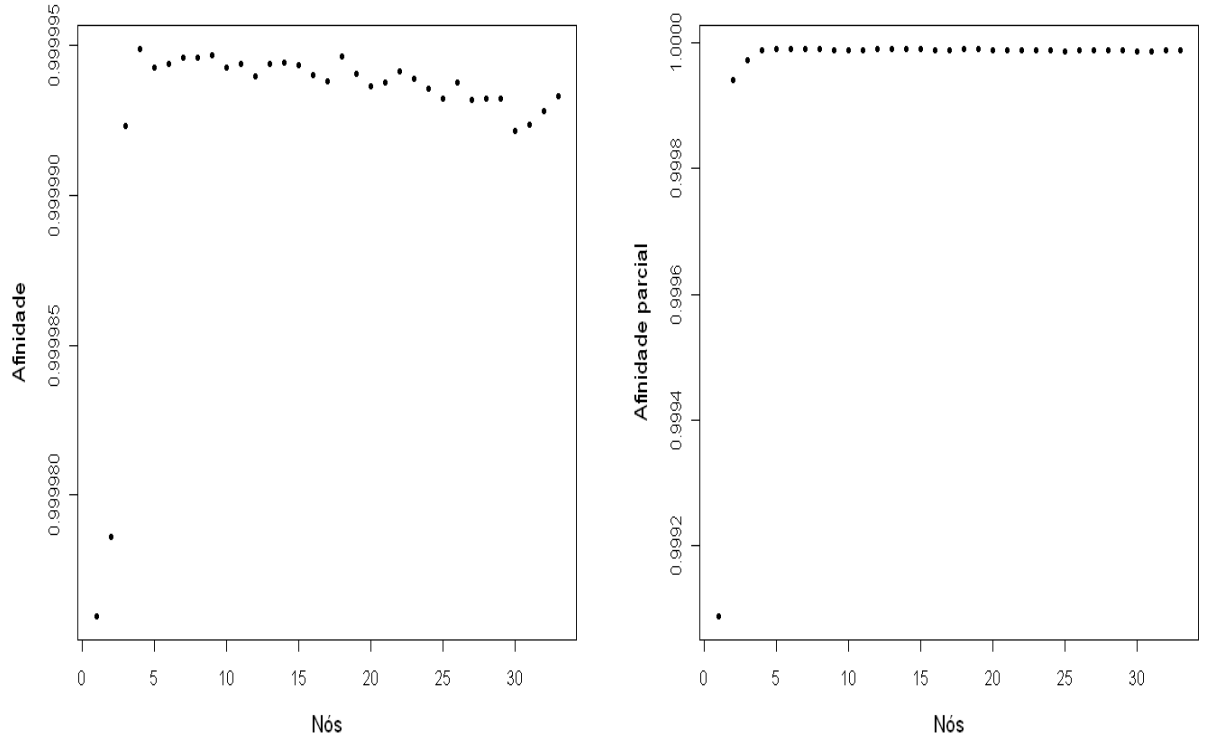


Figura 4.17: Função de covariância gaussiana com parâmetros: $\sigma^2 = 1$, $\phi = 2,5$. A linha sólida cinza é a função verdadeira e a linha pontilhada a função estimada.



(a) (b)

Figura 4.18: Quinhentas replicações com $n = 500$ e 33 bases: (a) afinidade e (b) afinidade parcial.

Tabela 4.7: Parâmetros das distribuições beta e normal estimados a partir dos valores da afinidade obtidos: modelo gaussiano ($\sigma^2 = 1$, $\phi = 2,5$).

	Beta		Normal	
	a	b	Média	Desvio Padrão
Afinidade	5366990	38,63598	9,999928e-01	1,156980e-06

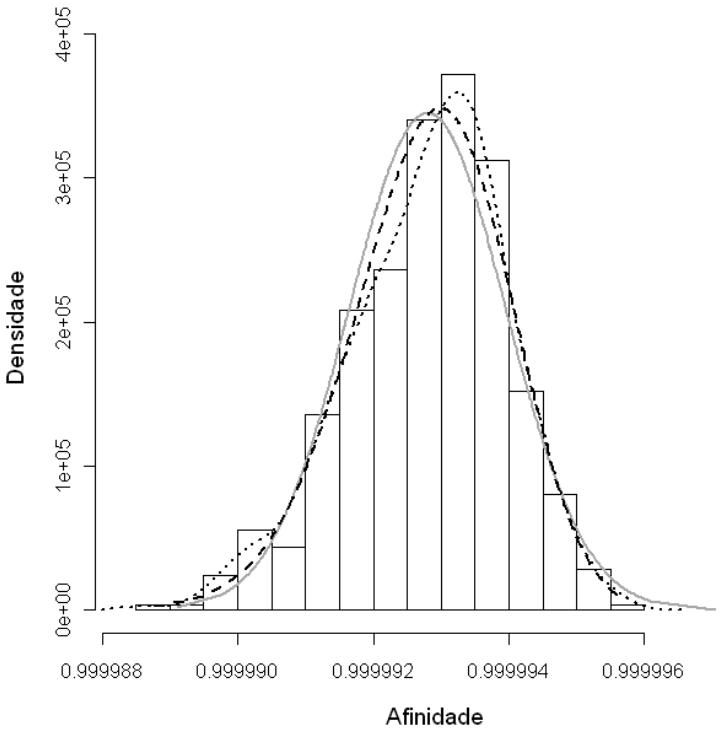


Figura 4.19: Estimativas da densidade da afinidade baseadas em 500 replicações, $n = 500$: linha sólida cinza - modelo normal, linha tracejada - modelo beta e a linha pontilhada - método *kernel*.

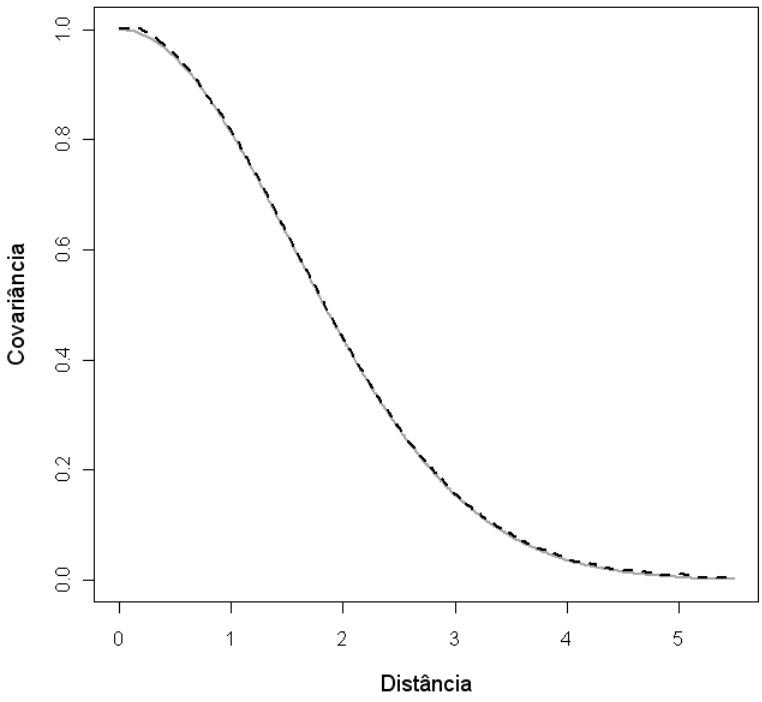
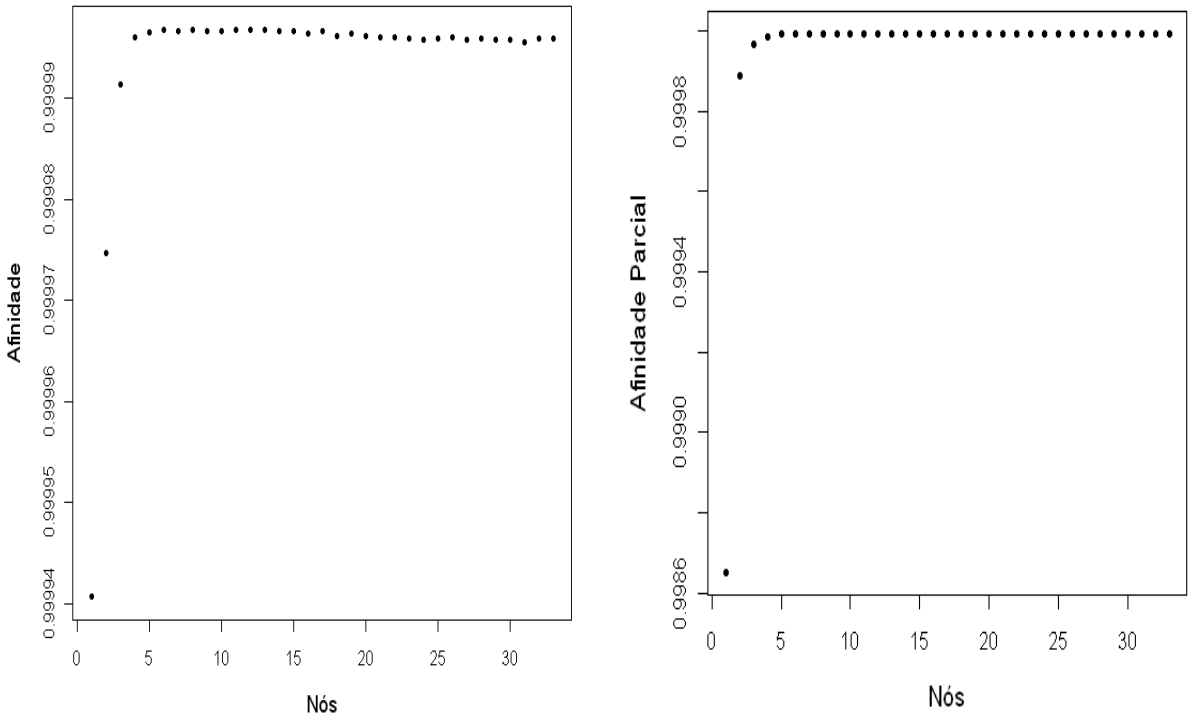


Figura 4.20: Função de covariância gaussiana com parâmetros: $\sigma^2 = 1$, $\phi = 2,2$. A linha sólida cinza é a função verdadeira e a linha pontilhada a função estimada.



(a) (b)

Figura 4.21: Quinhentas replicações com $n = 1000$ e 32 bases: (a) afinidade e (b) afinidade parcial.

Tabela 4.8: Parâmetros das distribuições beta e normal estimados a partir dos valores da afinidade obtidos: modelo gaussiana ($\sigma^2 = 1$, $\phi = 2, 2$).

	Beta		Normal	
	a	b	Média	Desvio Padrão
Afinidade	13356260	83,58862	9,999937e-01	6,838328e-07

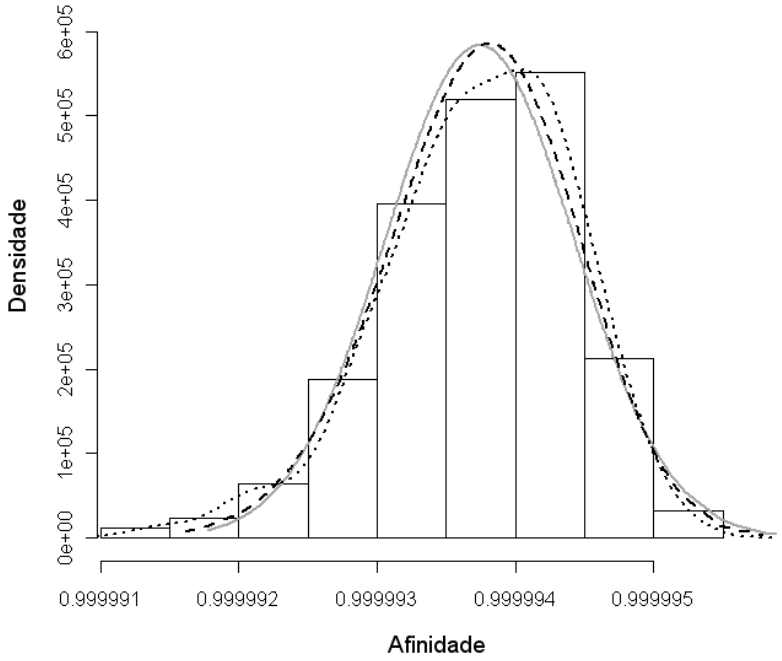


Figura 4.22: Estimativas da densidade da afinidade baseadas em 500 replicações, $n = 1000$: linha sólida cinza - modelo normal, linha tracejada - modelo beta e a linha pontilhada - método *kernel*.

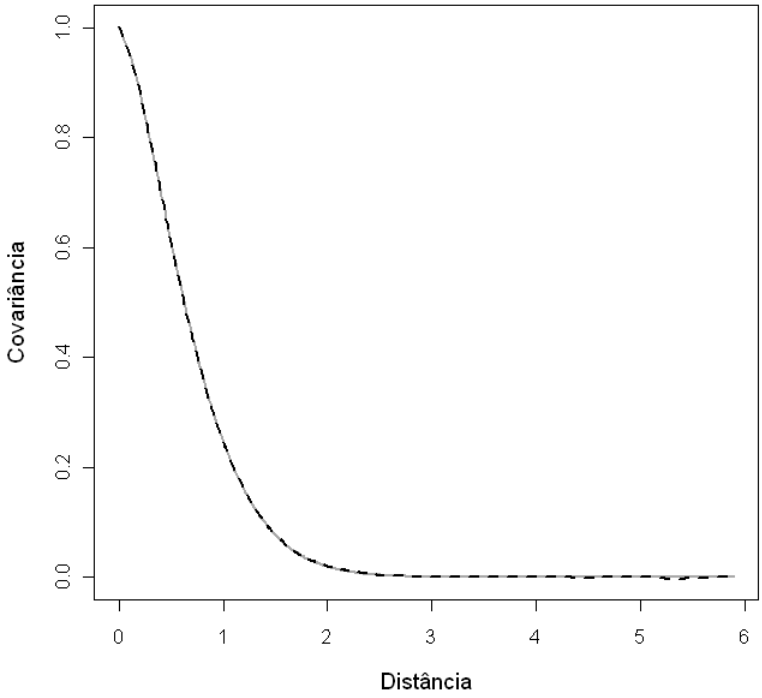
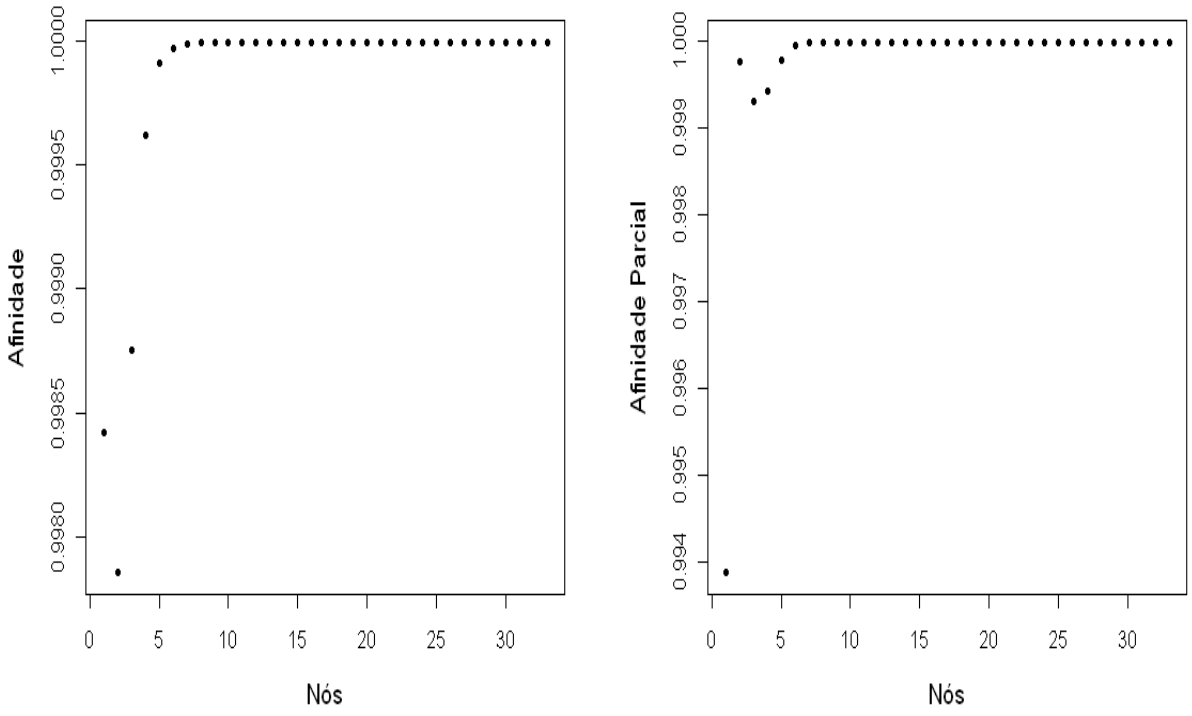


Figura 4.23: Função de covariância exponencial potência com parâmetros: $\sigma^2 = 1$, $\phi = 2,5$ e $\kappa = 1,5$. A linha sólida cinza é a função verdadeira e a linha pontilhada a função estimada.



(a) (b)

Figura 4.24: Quinhentas replicações com $n = 1000$ e 56 bases: (a) afinidade e (b) afinidade parcial.

Tabela 4.9: Parâmetros das distribuições beta e normal estimados a partir dos valores da afinidade obtidos: modelo exponencial potência ($\sigma^2 = 1$, $\phi = 2,5$, $\kappa = 1,5$).

	Beta		Normal	
	a	b	Média	Desvio Padrão
Afinidade	232506994	39935,07	9,998283e-01	8,584097e-07

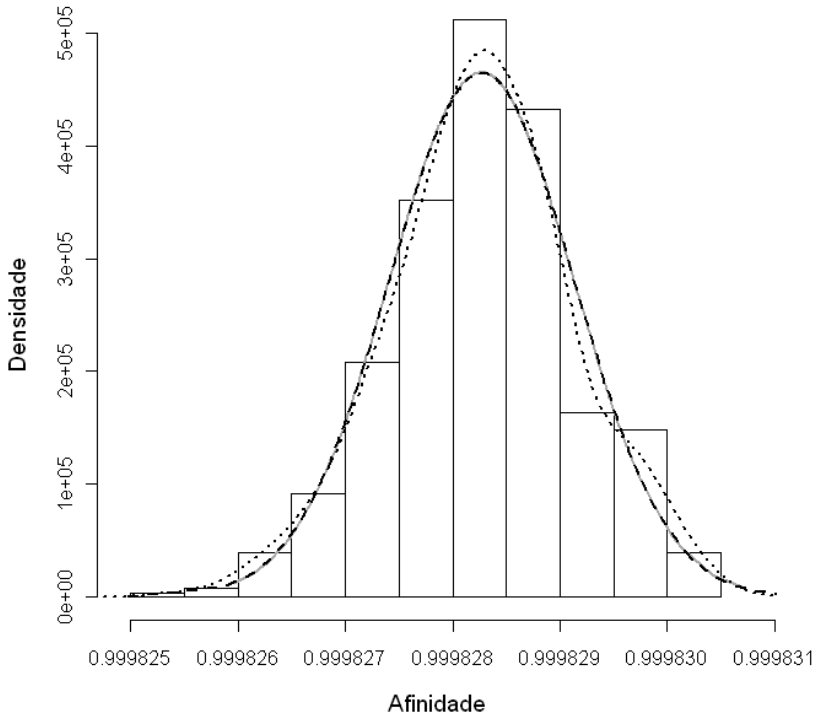


Figura 4.25: Estimativas da densidade da afinidade baseadas em 500 replicações, $n = 1000$: linha sólida cinza - modelo normal, linha tracejada - modelo beta e a linha pontilhada - método *kernel*.

Capítulo 5

Funções de Covariância Espacial

A função de covariância pode ser escrita como um produto do parâmetro de variância vezes uma função de correlação definida positiva $\rho(\tau)$, como

$$\mathcal{C}(\tau) = \sigma^2 \times \rho(\tau). \tag{5.1}$$

Seja ϕ o parâmetro da função de covariância denominado parâmetro de amplitude. Algumas das funções de covariância possuem um parâmetro extra κ, denominado parâmetro de suavidade. Esses parâmetros existem puramente para incluir no modelo alguma flexibilidade na forma total da função de correlação. Assim, um modelo para a função de covariância pode ter um ou dois parâmetros.

No jargão da geoestatística, uma combinação linear de modelos de covariância é dito formar um modelo de estruturas encaixadas, onde cada uma das estruturas encaixadas corresponde a um termo da combinação linear.

5.1 Exponencial Potência

Esta família é definida por

$$\mathcal{C}(\tau) = \sigma^2 \exp\left\{-\left(\frac{\tau}{\phi}\right)^{\kappa}\right\}, \tag{5.2}$$

com $\phi > 0$ e $0 < \kappa \leq 2$.

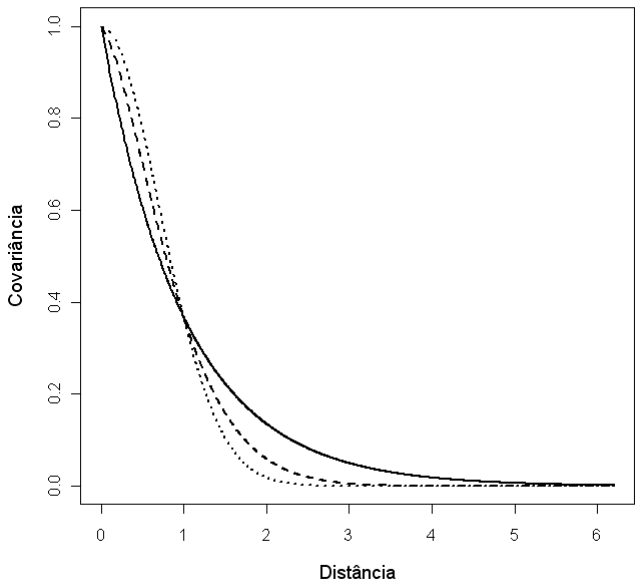


Figura 5.1: Três exemplos da função de covariância exponencial potência com $\phi = 1$ e $\kappa = 1$ (linha sólida), $\kappa = 1,5$ (linha tracejada) e $\kappa = 2$ (linha pontilhada).

Quando $\kappa = 1$ esta função é denominada apenas de função de covariância exponencial com parâmetro de escala $\phi > 0$ e é definida por,

$$\mathcal{C}(\tau) = \sigma^2 \exp\left(-\frac{\tau}{\phi}\right), \tag{5.3}$$

é uma covariância em $\mathbb{R}^n$ para qualquer n.

No caso $\kappa = 2$ temos a função de covariância Gaussiana. Essa família frequentemente dá um ajuste qualitativamente razoável com respeito a estrutura de covariância de dados espaciais, mas suas predições tendem a ser não robusta para pequenas partidas do modelo assumido, nem menos por que a matriz de covariância de qualquer coleção de valores $S(x_i)$ gerados deste modelo é muito mal-condicionada. Além disso, em uma rápida pincelada no comportamento de $\kappa < 2$ para $\kappa = 2$ observa-se que a família exponencial não é tão flexível quanto pareceria ser em um primeiro contato, veja Ribeiro e Diggle (2000).

5.2 Gaussiana

O modelo Gaussiano com parâmetro de escala $\phi > 0$ é definido por,

$$\mathcal{C}(\tau) = \sigma^2 \exp\left\{-\left(\frac{\tau}{\phi}\right)^2\right\}. \tag{5.4}$$

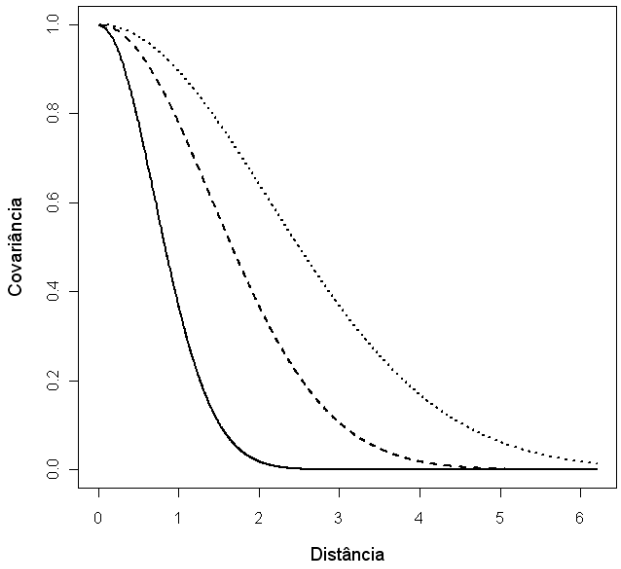


Figura 5.2: Três exemplos da função de covariância gaussiana com $\phi = 1$ (linha sólida), $\phi = 2$ (linha tracejada) e $\phi = 3$ (linha pontilhada).

5.3 Circular

Seja $\theta = \min(\tau/\phi, 1)$ e

$$\gamma(\tau) = 2\left(\theta\sqrt{1-\theta^2} + \text{sen}^{-1}\sqrt{\theta}\right)/\pi. \tag{5.5}$$

Então, o modelo circular é definido por,

$$\mathcal{C}(\tau) = \begin{cases} \sigma^2(1-\gamma(\tau)), & \text{se } \tau < \phi \\ 0, & \text{se } \tau \geq \phi \end{cases} \tag{5.6}$$

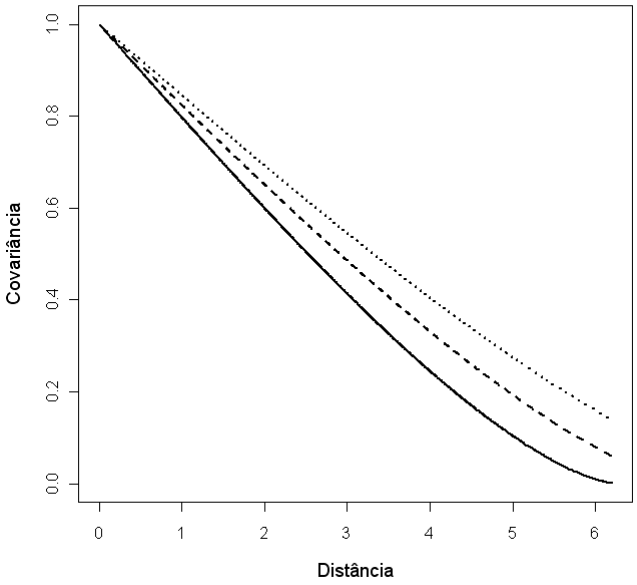


Figura 5.3: Três exemplos da função de covariância circular com $\phi = \max(\tau) + 0,05$ (linha sólida), $\phi = \max(\tau) + 1$ (linha tracejada), e $\phi = \max(\tau) + 2$ (linha pontilhada).

5.4 Cúbica

A função de covariância cúbica é definida por (Wackernagel, 1995, p.218)

$$\mathcal{C}(\tau) = \begin{cases} \sigma^2 \left(1 - 7\left(\frac{\tau}{\phi}\right)^2 + \frac{35}{4}\left(\frac{\tau}{\phi}\right)^3 - \frac{7}{2}\left(\frac{\tau}{\phi}\right)^5 + \frac{3}{4}\left(\frac{\tau}{\phi}\right)^7\right), & \text{se } \tau \leq \phi \\ 0, & \text{se } \tau > \phi \end{cases} \tag{5.7}$$

Este modelo é usado para variáveis diferenciáveis, tal como campos de pressão ou em meteorologia (Chilès e Delfiner, 1999).

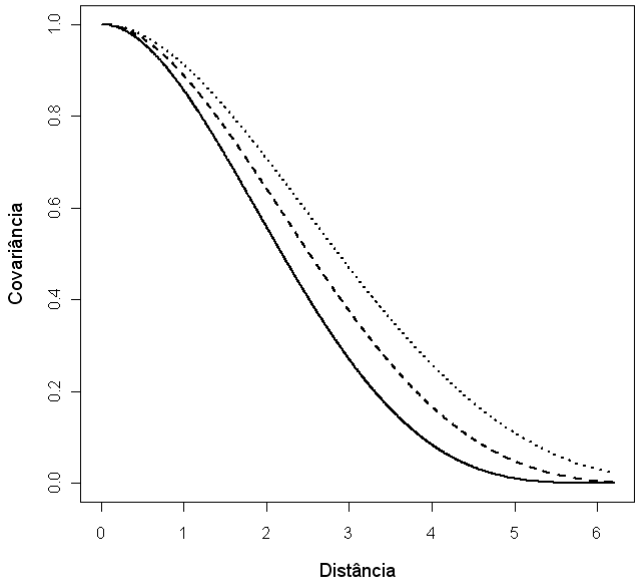


Figura 5.4: Três exemplos da função de covariância cúbica com $\phi = \max(\tau) + 0,05$ (linha sólida), $\phi = \max(\tau) + 1$ (linha tracejada), e $\phi = \max(\tau) + 2$ (linha pontilhada).

5.5 Esférica

Esta família de funções de covariância uniparamétrica é definida por,

$$\mathcal{C}(\tau) = \begin{cases} \sigma^2 \left(1 - \frac{3}{2}\frac{\tau}{\phi} + \frac{1}{2}\left(\frac{\tau}{\phi}\right)^3\right), & \text{se } 0 \leq \tau \leq \phi \\ 0, & \text{se } \tau \geq \phi \end{cases} \tag{5.8}$$

Seu nome é devido ao fato da correlação possuir uma interpretação geométrica como o volume de interseção de duas esferas tridimensional, cujo o centro é uma distância τ à parte. Em algumas propostas é conveniente que a correlação tenha amplitude finita. Pois a família depende somente do parâmetro de escala ϕ, que não dá nenhuma flexibilidade na forma. Sua motivação na área física é de relevância duvidosa para problemas em espaços bidimensional.

5.6 Quadrática Racional ou de Cauchy

A função de covariância quadrática racional é dada por

$$\mathcal{C}(\tau) = \sigma^2 \left(1 + \frac{\tau^2}{\phi^2}\right)^{-\kappa}, \ \tau > 0, \phi > 0, \kappa > 0 \tag{5.9}$$

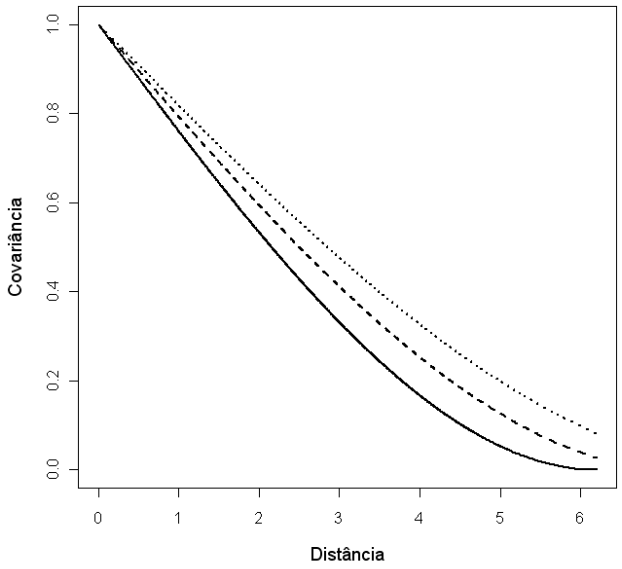


Figura 5.5: Três exemplos da função de covariância esférica com $\phi = \max(\tau)$ (linha sólida), $\phi = \max(\tau) + 1$ (linha tracejada) e $\phi = \max(\tau) + 2$ (linha pontilhada).

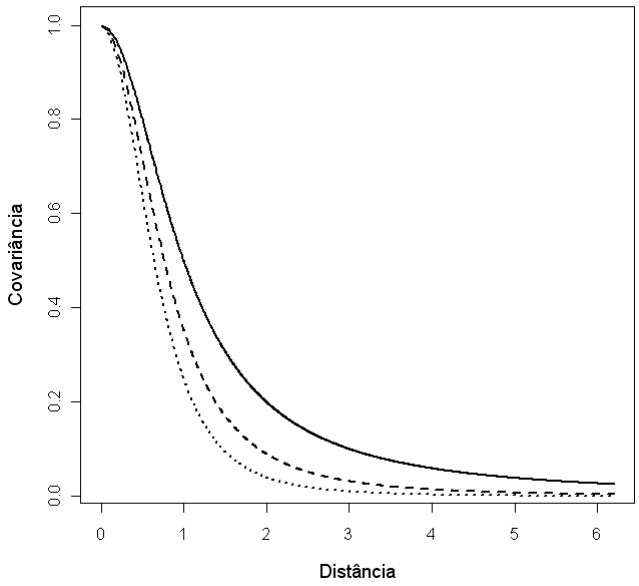


Figura 5.6: Três exemplos da função de covariância quadrática racional com $\phi = 1$ e $\kappa = 1$ (linha sólida), $\kappa = 1,5$ (linha tracejada) e $\kappa = 2$ (linha pontilhada).

5.7 Quadrática Racional Generalizada

Os campos de gravidade e magnético são governados pelas leis da física. Se a geometria, densidade e magnetismo das fontes são conhecidos, os campos correspondentes podem ser determinado. No entando, conhecendo estas características, um modelo estatístico dos principais parâmetros permite a determinação, se não dos campos, pelo menos do seu espectro. A covariância quadrática racional generalizada é definida por,

$$\mathcal{C}(\tau) = \sigma^2 \left(1 + \left(\frac{\tau}{\phi}\right)^{\kappa_2}\right)^{-\frac{\kappa_1}{\kappa_2}}, \ \tau > 0, \phi > 0, \ \kappa_1 > 0, 0 < \kappa_2 \leq 2 \tag{5.10}$$

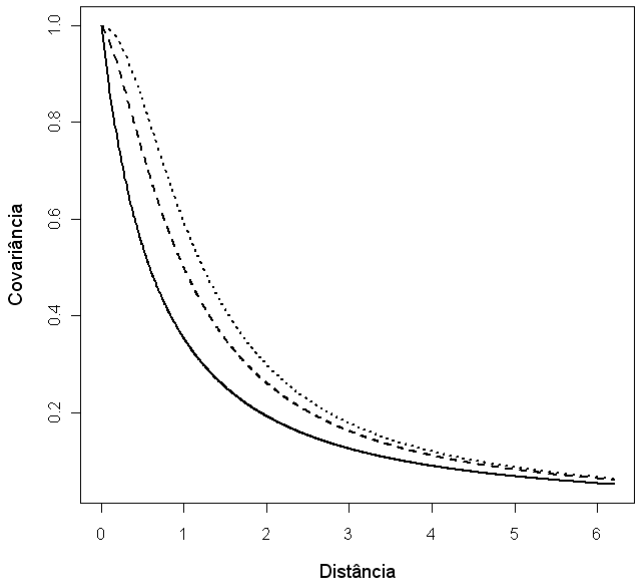


Figura 5.7: Três exemplos da função de covariância quadrática racional generalizada com $\phi = 1$, $\kappa_1 = 1,5$ fixo, para $\kappa_2 = 1$(linha sólida), $\kappa_2 = 1,5$ (linha tracejada) e $\kappa_2 = 2$ (linha pontilhada).

5.8 Ondular

A função de covariância ondular é dada por

$$\mathcal{C}(\tau) = \sigma^2 \frac{\text{sen}(\phi\tau)}{\phi\tau}, \ \phi > 0. \tag{5.11}$$

Apresenta correlações negativas devido a periodicidade do processo (Figura 5.8).

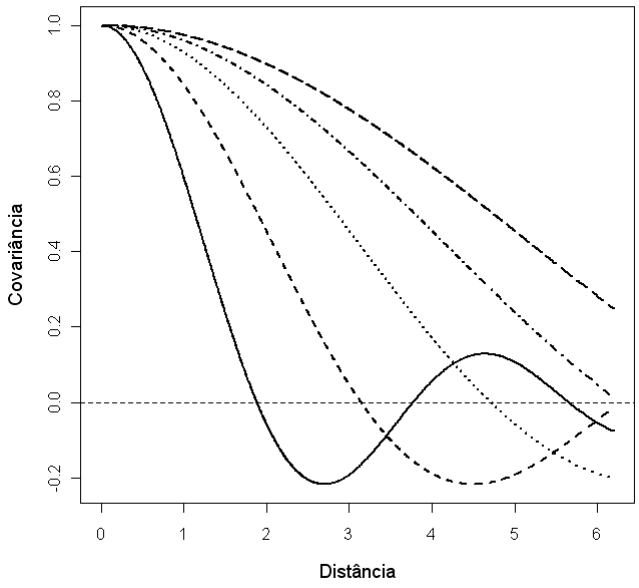


Figura 5.8: Três exemplos da função de covariância ondular: com $\phi = 0,6$ (linha sólida), $\phi = 1$ (linha tracejada), $\phi = 1,5$ (linha pontilhada), $\phi = 2$ (linha ponto-traço), $\phi = 2,5$ (linha traço longo).

Observa-se na Figura 5.9, que poderíamos ter situações aonde a função de covariância ondular somente assumisse valores positivos. Então poderíamos estimar normalmente para tamanho amostral pequeno.

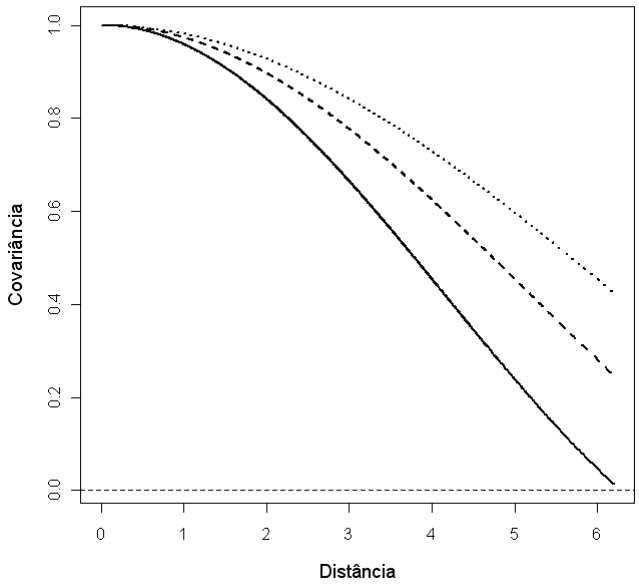


Figura 5.9: Três exemplos da função de covariância ondular: com $\phi = 2$ (linha sólida), $\phi = 2,5$ (linha tracejada), $\phi = 3$ (linha pontilhada).

5.9 Família de Matérn

Esta família é denominada Matérn depois que Bertil Matérn a introduziu em sua tese de doutorado em 1960. Ela é definida por

$$\mathcal{C}(\tau;\phi,\kappa) = \frac{\sigma^2}{2^{\kappa-1}\Gamma(\kappa)}\left(\frac{\tau}{\phi}\right)^{\kappa} K_{\kappa}\left(\frac{\tau}{\phi}\right),\ \tau > 0 \text{ e } \kappa > 0 \tag{5.12}$$

em que ϕ é uma parâmetro de escala, κ é um parâmetro de forma, $\Gamma(\cdot)$ é a função Gama usual e $K_{\kappa}(\cdot)$ denota a função de Bessel modificada do terceiro tipo de ordem κ.

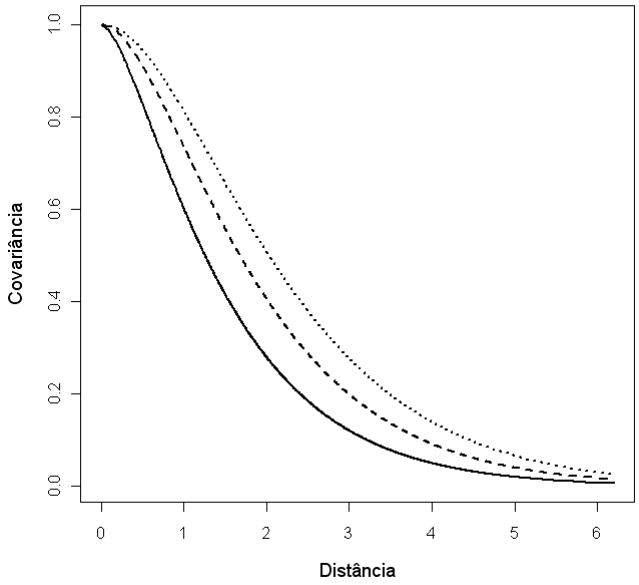


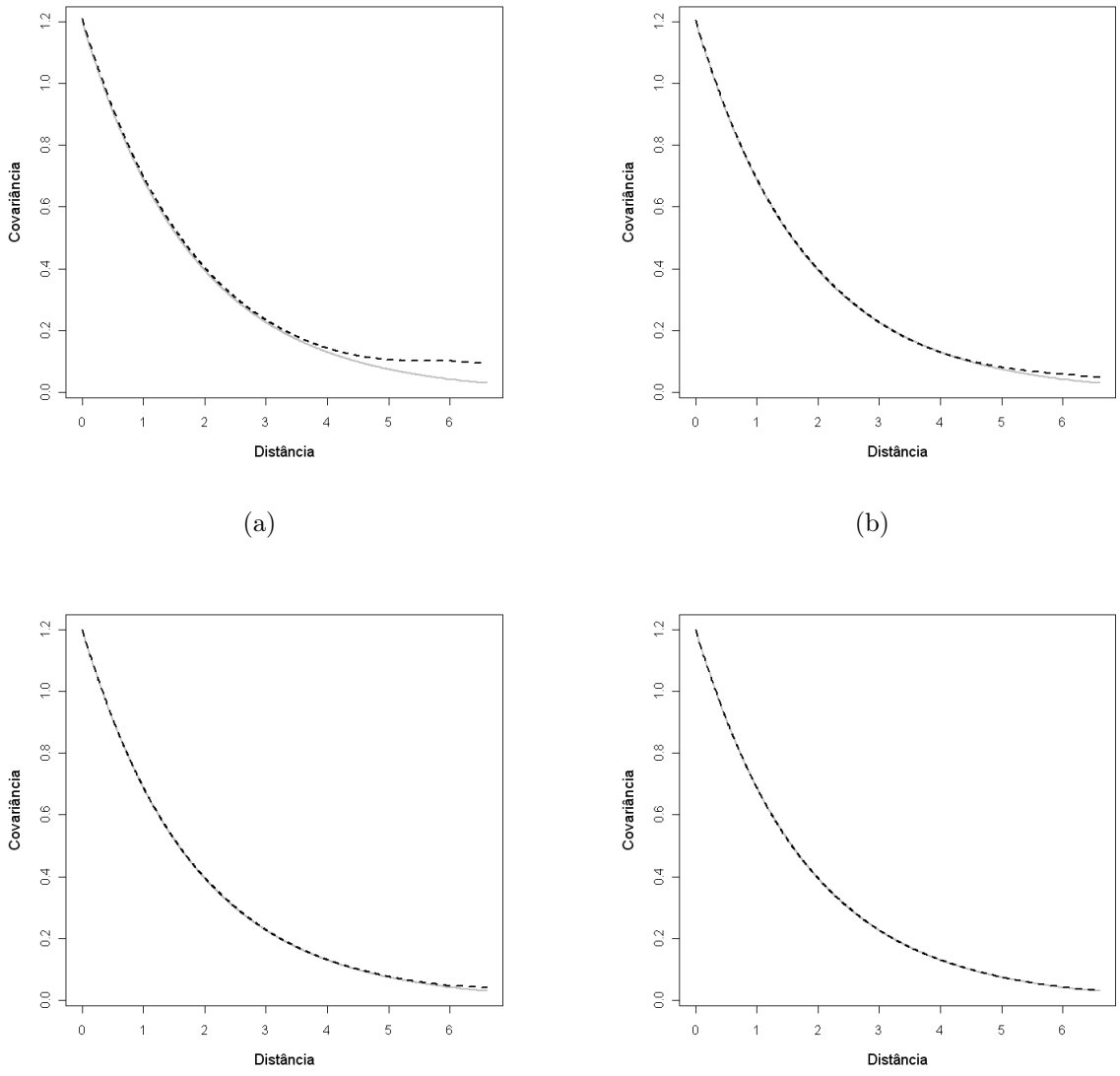
Figura 5.10: Três exemplos da função de covariância de Matérn com $\phi = 1$ e $\kappa = 1$ (linha sólida), $\kappa = 1,5$ (linha tracejada) e $\kappa = 2$ (linha pontilhada).

O parâmetro ϕ controla a taxa de decaimento para zero da covariância quando a distância τ aumenta. Valores grandes de ϕ indica que as localizações que são relativamente distantes umas das outras são moderadamente correlacionados positivamente. O

parâmetro κ indica a ordem do modelo de Matérn, e determina a suavidade analítica do sinal, $S(\mathbf{x})$. O parâmetro κ controla o comportamento da função de covariância para observações que são separadas por distâncias pequenas. A classe de Matérn abrange a função de covariância exponencial quando $\kappa = 0,5$ e a função de covariância Gaussiana quando $\kappa \to \infty$ (Hoeting *et al.*, 2006). Esta função de covariância pode ainda ser adaptada para incluir a possibilidade de medidas de erro, chamada *nugget* em muitos contextos espaciais, que não será visto aqui.

5.10 Simulação das Funções de Covariância

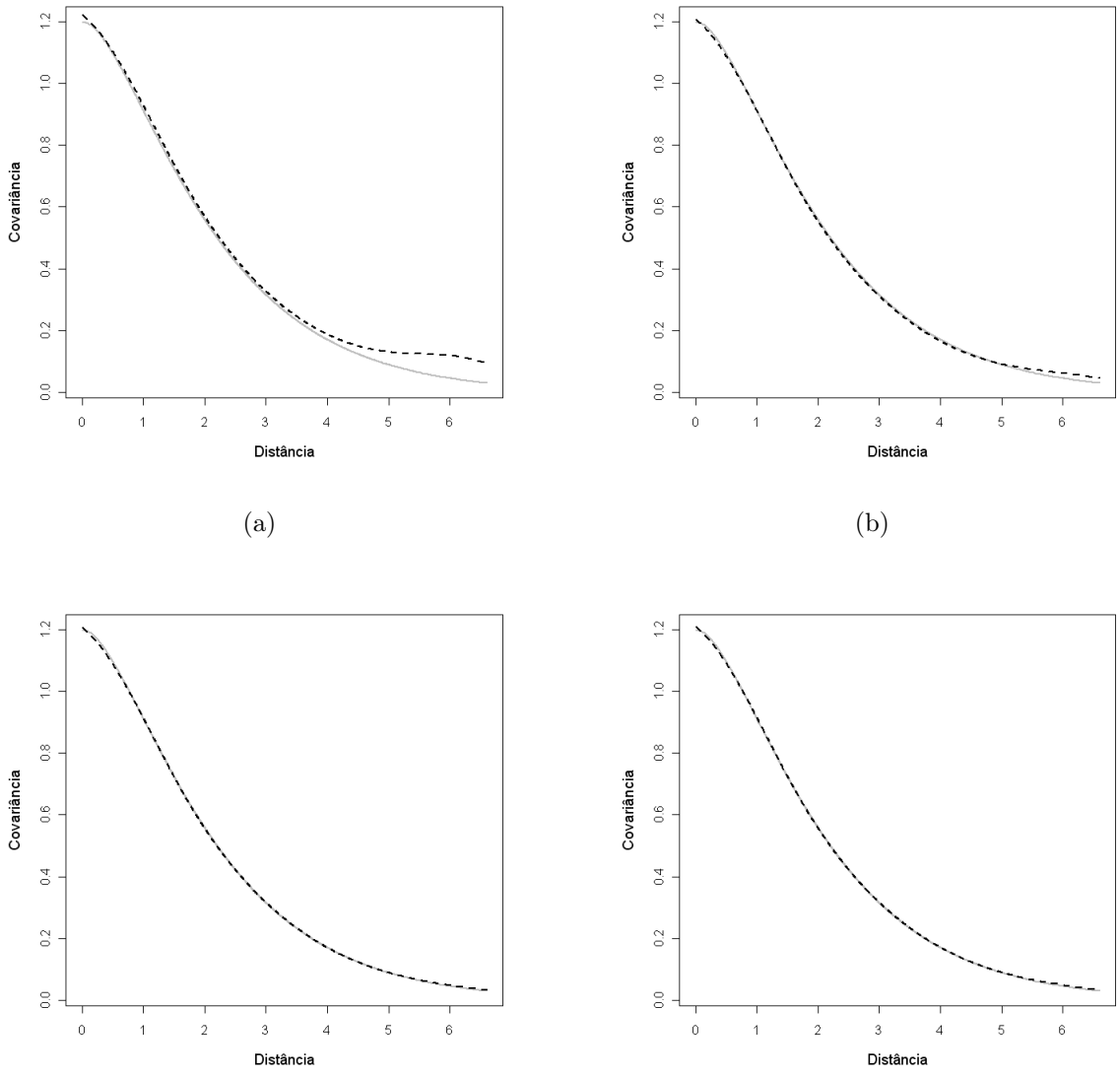
Agora mostraremos as funções de covariâncias estimadas em comparação com a função de covariância verdadeira. Fixou-se o parâmetro da função de covariância, $\sigma^2 = 1, 2$, para todas as simulações e variamos o tamanho amostral em $n = 25, 100, 500, 1000$, respectivamente em cada figura desta seção. Os valores dos parâmetros ϕ e κ foram tomados de acordo com valores comuns encontrados na literatura. Se beneficiando de uma análise visual nota-se nos gráficos abaixo que as estimativas foram melhorando à medida que se aumentou o tamanho amostral. Observa-se que nos tamanhos amostrais de $n = 25$ e $n = 100$ ocorre problema nas caudas, onde o decaimento deixa de ser exponencial e passa a subir para em seguida decair. Enquanto, para $n = 500$ e $n = 1000$ a função tem decaimento exponencial na totalidade, mostrando-se estimativas muito boas da funções de covariâncias. Para a construção do processo Gaussiano foi adotado um gride no quadrado $[0, 5] \times [0, 5]$. Baseado em 64 localizações foram geradas amostras de tamanho $25, 100, 500$ e 1000 da distribuição normal multivariada, veja Capítulo 3, com vetor de médias de $\mathbf{1}$'s e matriz de covariância $\Sigma_{64\times64}$. As simulações foram realizadas usando as funções de covariância espacial implementadas no pacote *geoR* do programa *R-Gui*, veja Ribeiro e Diggle (2001).



(a) (b)

(c) (d)

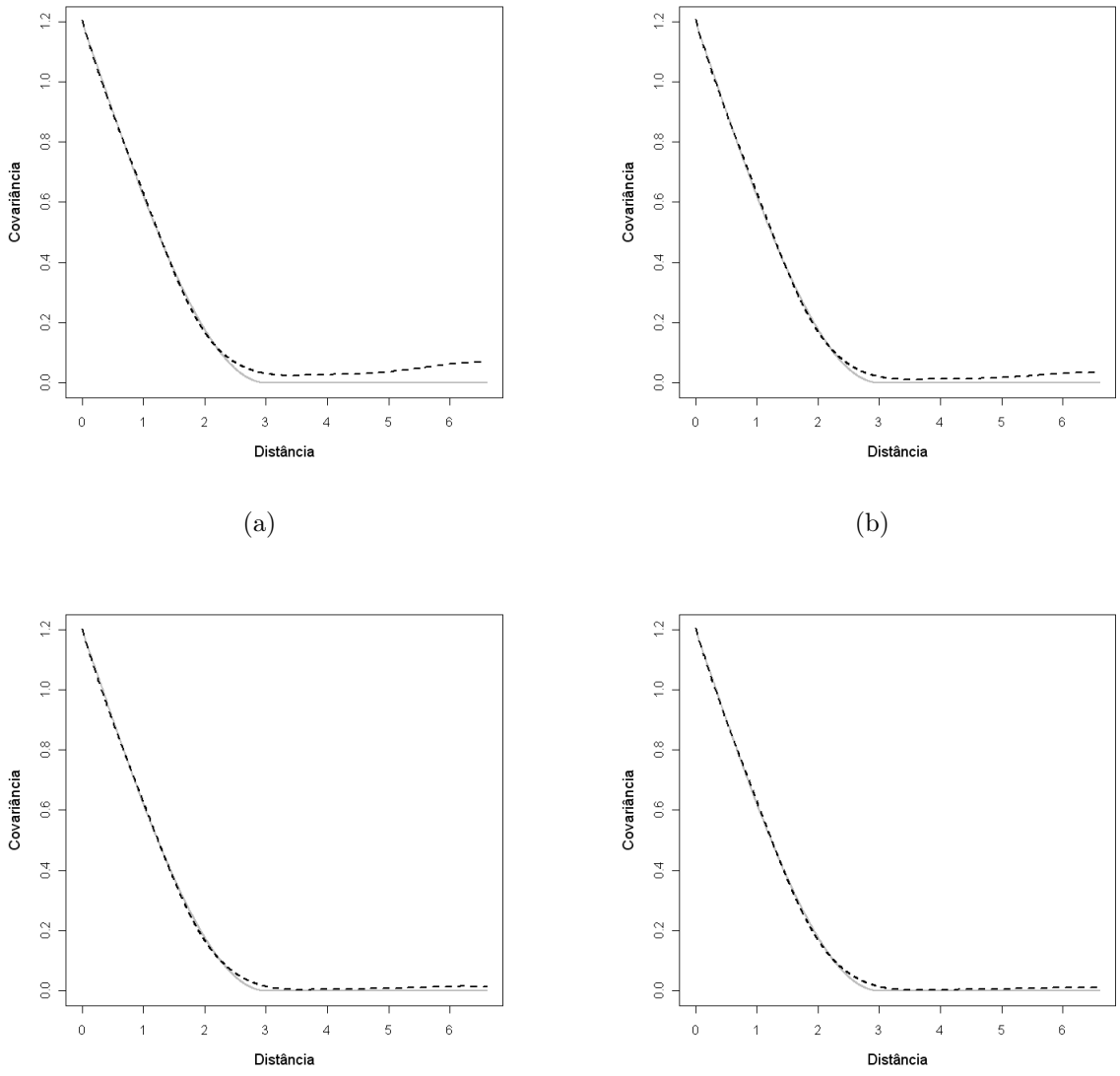
Figura 5.11: Função de covariância exponencial com $\phi = 1,8$: a linha cheia em cinza representa a função verdadeira e a linha tracejada representa a função estimada para: (a) $n = 25$; 3 bases (b) $n = 100$; 2 bases (c) $n = 500$; 2 bases (d) $n = 1000$; 2 bases.



(a) (b)

(c) (d)

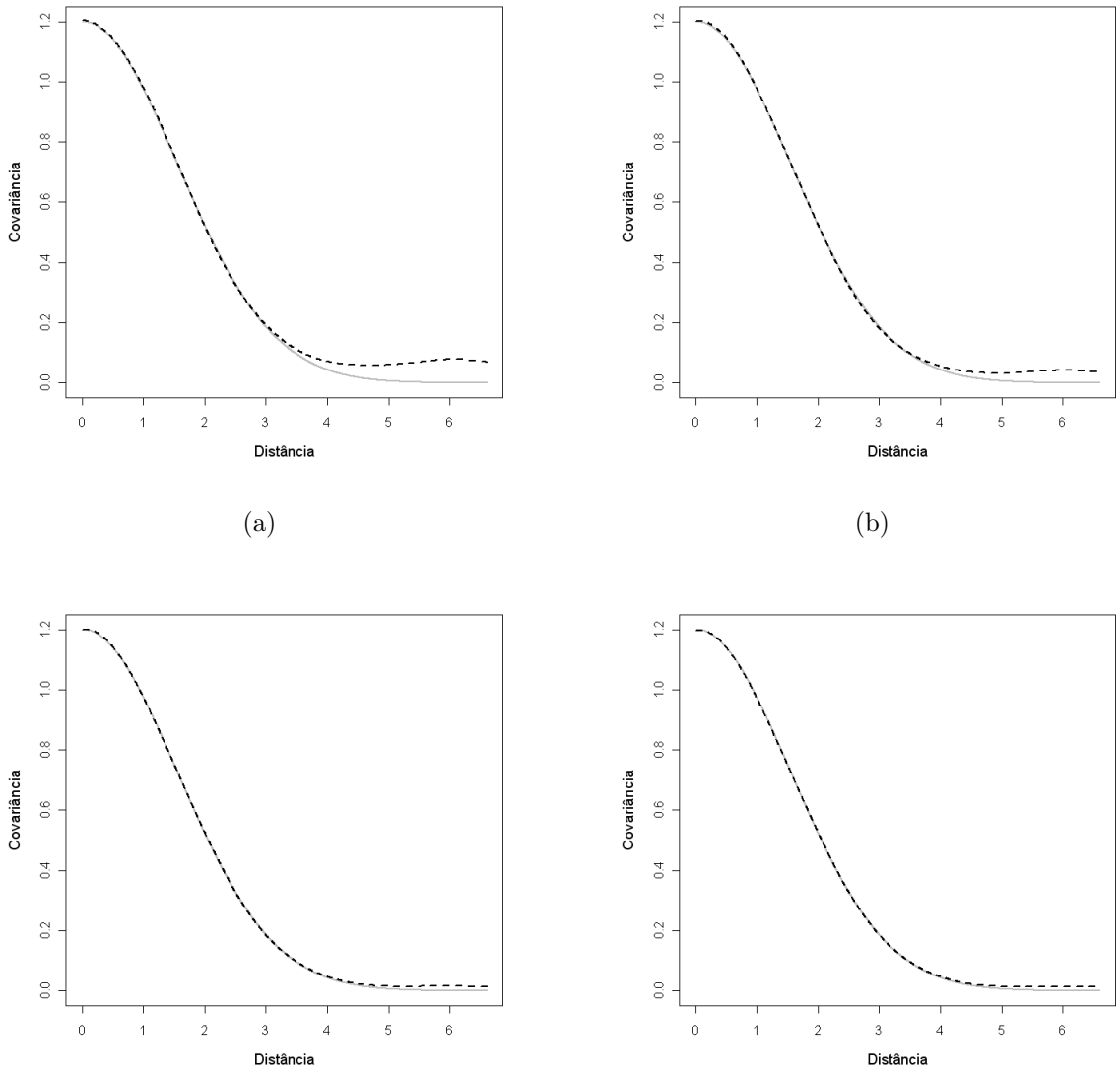
Figura 5.12: Função de covariância Matérn com $\phi = 1,3$ e $\kappa = 1,2$: a linha cheia em cinza representa a função verdadeira e a linha tracejada representa a função estimada para: (a) $n = 25$; 3 bases (b) $n = 100$; 3 bases (c) $n = 500$; 3 bases (d) $n = 1000$; 3 bases.



(a) (b)

(c) (d)

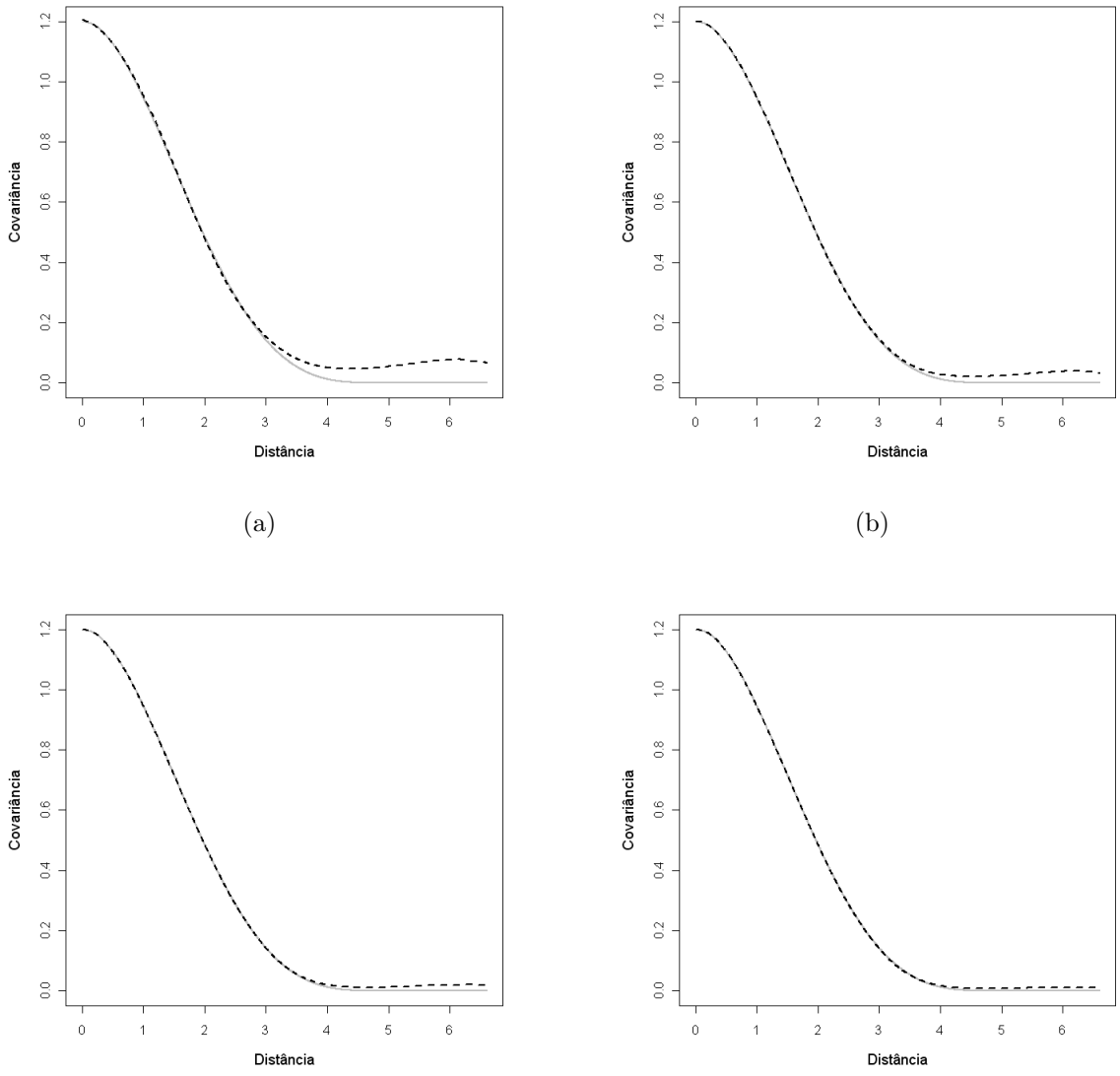
Figura 5.13: Função de covariância esférica com $\phi = 3$: a linha cheia em cinza representa a função verdadeira e a linha tracejada representa a função estimada para: (a) $n = 25$; 6 bases (b) $n = 100$; 6 bases (c) $n = 500$; 6 bases (d) $n = 1000$; 6 bases.



(a) (b)

(c) (d)

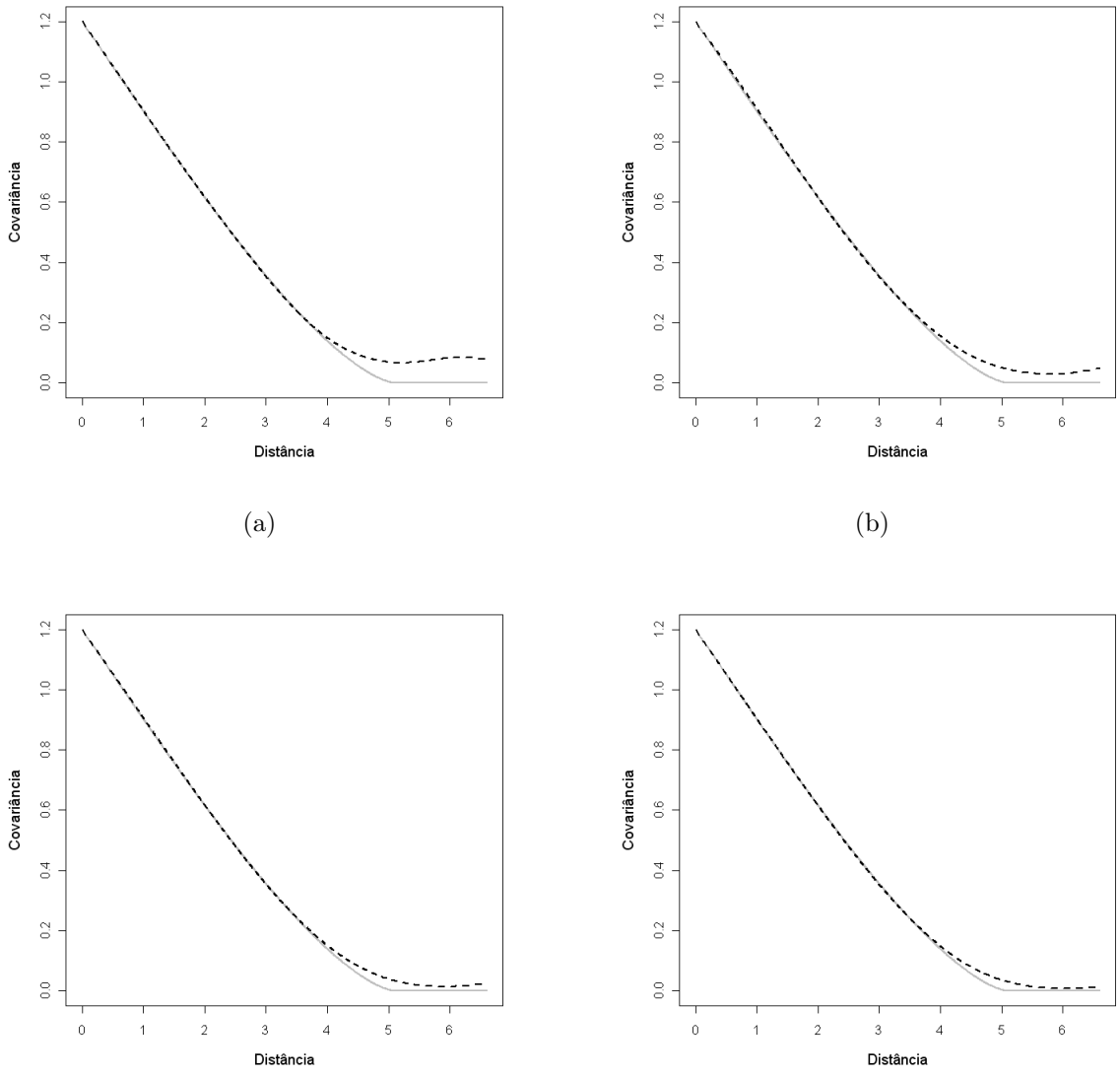
Figura 5.14: Função de covariância Gaussiana com $\phi = 2,2$: a linha cheia em cinza representa a função verdadeira e a linha tracejada representa a função estimada para: (a) $n = 25$; 5 bases (b) $n = 100$; 4 bases (c) $n = 500$; 4 bases (d) $n = 1000$; 4 bases.



(a) (b)

(c) (d)

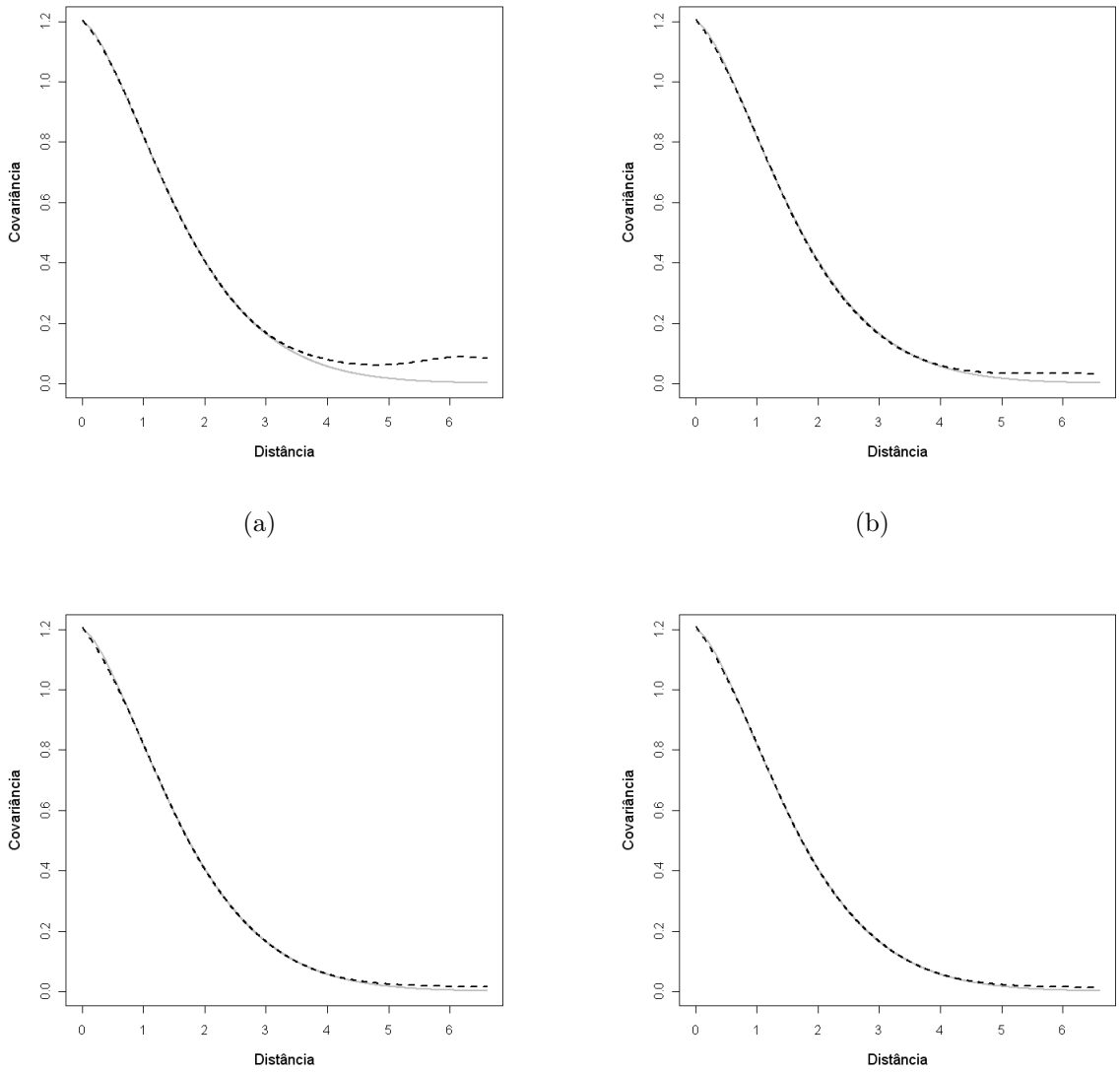
Figura 5.15: Função de covariância cúbica com $\phi = 5$: a linha cheia em cinza representa a função verdadeira e a linha tracejada representa a função estimada para: (a) $n = 25$; 5 bases (b) $n = 100$; 5 bases (c) $n = 500$; 5 bases (d) $n = 1000$; 5 bases.



(a) (b)

(c) (d)

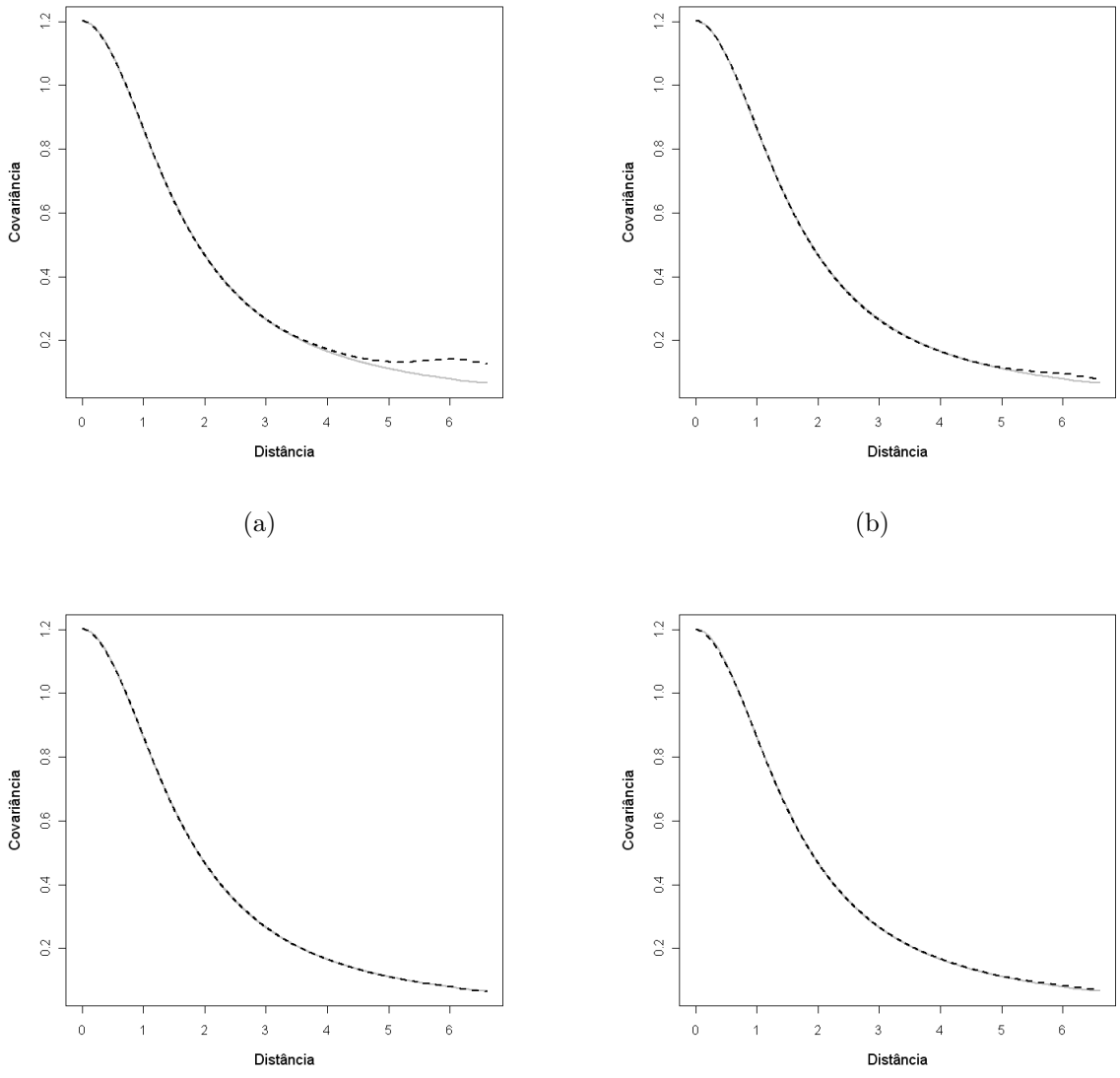
Figura 5.16: Função de covariância circular com $\phi = 5,1$: a linha cheia em cinza representa a função verdadeira e a linha tracejada representa a função estimada para: (a) $n = 25$; 4 bases (b) $n = 100$; 2 bases (c) $n = 500$; 2 bases (d) $n = 1000$; 2 bases.



(a) (b)

(c) (d)

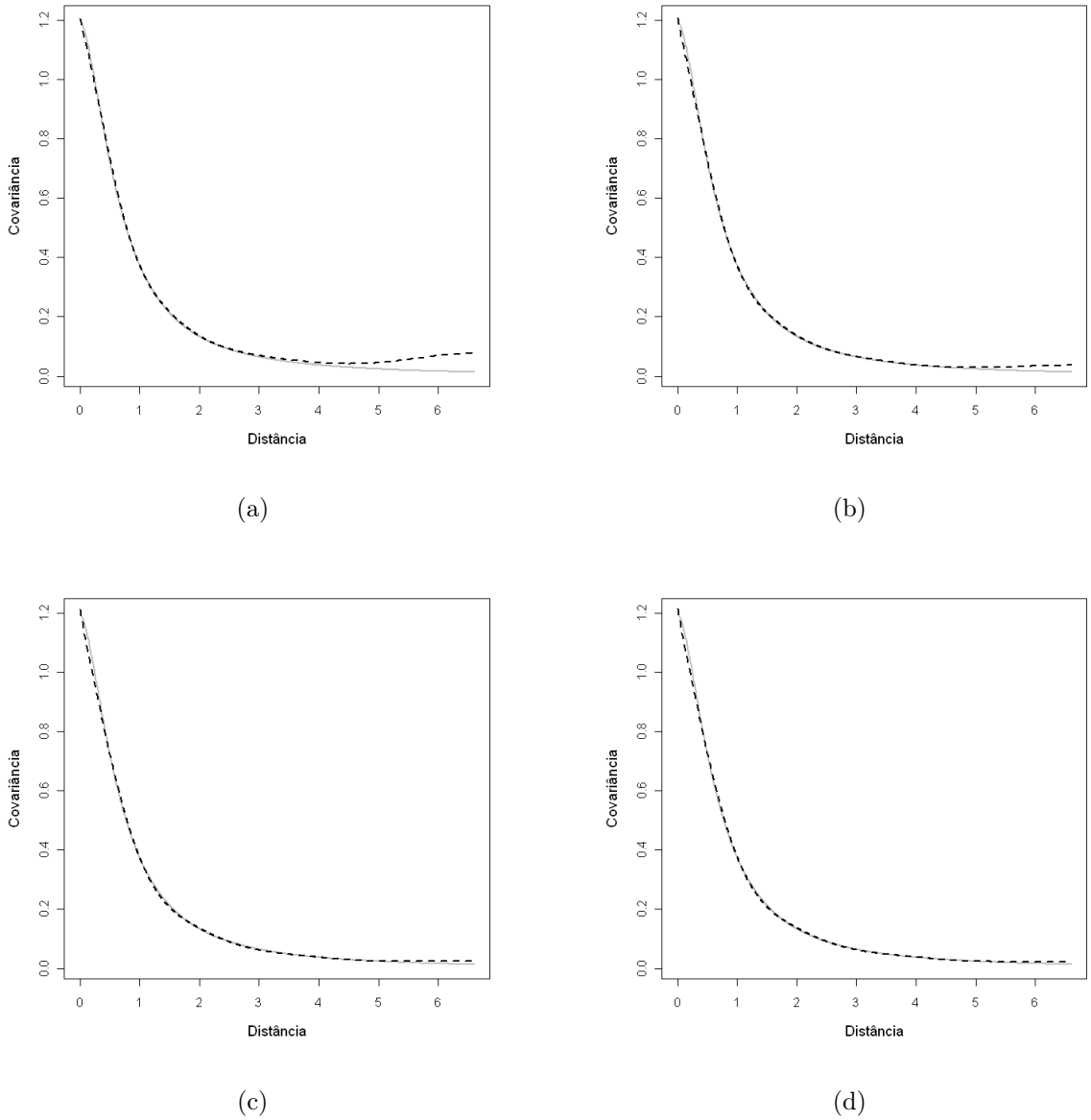
Figura 5.17: Função de covariância exponencial potência com $\phi = 1,9$ e $\kappa = 1,5$: a linha cheia em cinza representa a função verdadeira e a linha tracejada representa a função estimada para: (a) $n = 25$; 3 bases (b) $n = 100$; 2 bases (c) $n = 500$; 2 bases (d) $n = 1000$; 2 bases.



(a) (b)

(c) (d)

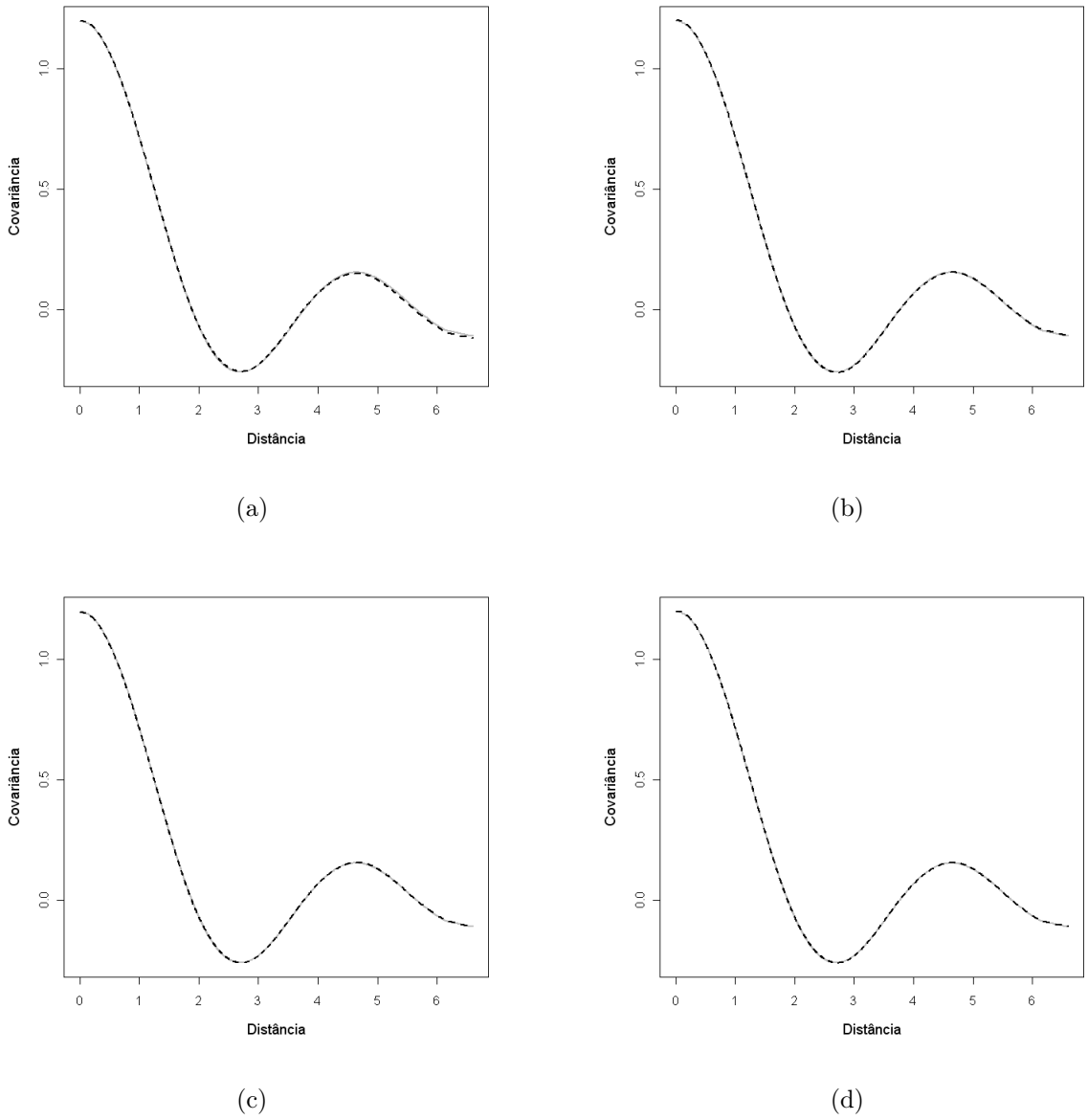
Figura 5.18: Função de covariância quadrática racional com $\phi = 1,6$ e $\kappa = 1$: a linha cheia em cinza representa a função verdadeira e a linha tracejada representa a função estimada para: (a) $n = 25$; 4 bases (b) $n = 100$; 4 bases (c) $n = 500$; 4 bases (d) $n = 1000$; 4 bases.



(a) (b)

(c) (d)

Figura 5.19: Função de covariância quadrática racional generalizada com $\phi = 0.8$ e $\kappa = (2, 1; 1, 6)$: a linha cheia em cinza representa a função verdadeira e a linha tracejada representa a função estimada para: (a) $n = 25$; 8 bases (b) $n = 100$; 7 bases (c) $n = 500$; 6 bases (d) $n = 1000$; 6 bases.



(a) (b)

(c) (d)

Figura 5.20: Função de covariância ondular $\phi = 0.6$: a linha cheia em cinza representa a função verdadeira e a linha tracejada representa a função estimada para: (a) $n = 25$; 6 bases (b) $n = 100$; 6 bases (c) $n = 500$; 6 bases (d) $n = 1000$; 6 bases.

5.11 Combinações de Funções de Covariância

Na seção anterior apresentamos algumas funções, na qual iremos trabalhar. Agora vamos apresentar como podemos construir funções de covariâncias válidas a partir de funções de covariância já obtidas. Se $\mathcal{C}_1(\tau)$ e $\mathcal{C}_2(\tau)$ são funções de covariâncias válidas, então a soma, $\mathcal{C}_1(\tau) + \mathcal{C}_2(\tau)$ e o produto, $\mathcal{C}_1(\tau) \times \mathcal{C}_2(\tau)$, também são funções de covariâncias válidas. Tomando a soma de funções de covariâncias de um modelo linear e uma função de covariância exponencial com variância pequena resulta em um modelo quase linear. Somas de funções de covariâncias com escalas de tamanhos diferentes resulta em funções com ambas de escala grande e escala pequena.

Produtos de funções de covariâncias univariadas para diferentes entradas resulta em funções de covariâncias multivariadas que permitem interações. A seguir apresentamos duas propriedades de funções definidas positivas, no qual podemos construir uma vasta família de covariâncias teóricas que podem ser definidas em termos de esquemas da positividade definida básica:

- Toda combinação linear de covariâncias com coeficientes positivos é uma covariância,

$$\mathcal{C}(\tau) = \sum_{i=1}^{n} \lambda_i^2 \mathcal{C}_i(\tau). \tag{5.13}$$

  Basta considerar o processo aleatório com a soma ponderada de n processos aleatórios independentes $Y_i(s)$ com covariâncias $\mathcal{C}_i(\tau)$. A covariância de

$$Z(s) = \sum_{i=1}^{n} \lambda_i Y_i(s) \tag{5.14}$$

  é dada por (5.13).

- Qualquer produto de covariância é uma covariância,

$$\mathcal{C}(\tau) = \prod_{i=1}^{n} \mathcal{C}_i(\tau) = \mathcal{C}_1(\tau) \times \mathcal{C}_2(\tau) \times \ldots \times \mathcal{C}_n(\tau). \tag{5.15}$$

Aqui consideramos o processo aleatório $Z(s)$ como produto dos processos aleatórios independentes $Y_i(s)$ com covariâncias $\mathcal{C}_i(\tau)$. A covariância de

$$Z(s) = \prod_{i=1}^{n} Y_i(s) \tag{5.16}$$

é dada por (5.15), veja Journel e Huijbregts, (1978).

5.12 Modelos Encaixados

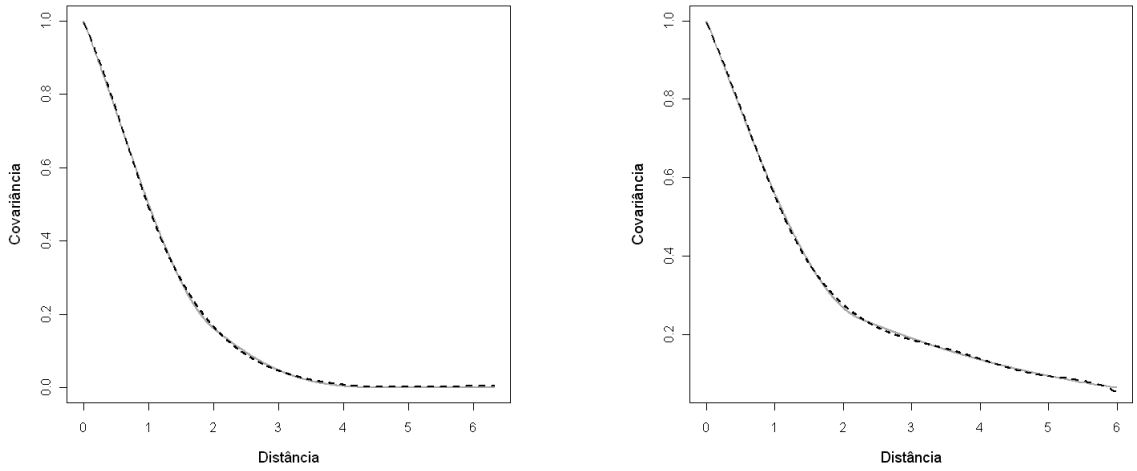
Na prática existem determinados fenômenos em que são necessários modelos de covariâncias mais complexos para explicar suas variações espaciais. McBratney e Webster (1986) observaram que modelos encaixados são necessários para explicar a variação do solo decorrente de fatores independentes de formação. Estes modelos são combinações de modelos simples, denominados encaixados. Por exemplo, um modelo encaixado útil em estudos de mineração e pesquisa de solo é o duplo esférico. McBratney *et al.* (1982) o utilizaram para descrever a variação do cobre e do cobalto no solo. Podemos ter processos com dependência espacial diferente, como duas escalas diferentes que podem ser modelados por um modelo encaixado tal como,

$$\mathcal{C}(\tau) = \begin{cases} \mathcal{C}_{\phi_1}(\tau), & 0 < \tau \leq \phi_1 \\ \mathcal{C}_{\phi_2}(\tau), & \phi_1 < \tau \leq \phi_2 \\ \text{constante}, & \tau > \phi_2, \end{cases} \tag{5.17}$$

onde $\mathcal{C}_{\phi_1}(\tau)$ e $\mathcal{C}_{\phi_2}(\tau)$ são modelos espaciais nas escalas ϕ_1 e ϕ_2, respectivamente. As escalas poderiam ser pequena e ter um alcance longo da dependência espacial. Todas as estruturas encaixadas atuam sobre todas as escalas de distância. Dependendo do fenômeno

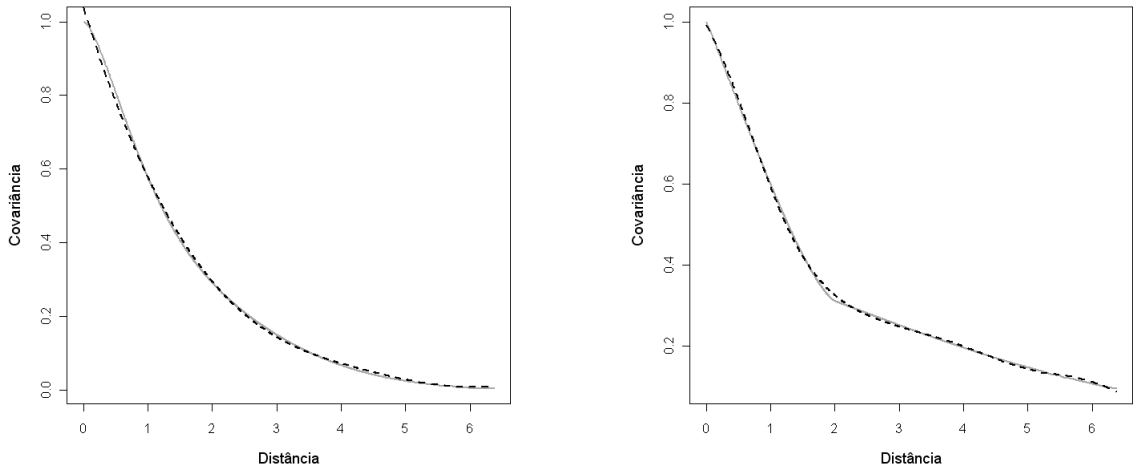
em estudo, outros modelos encaixados são necessários para caracterizar a variabilidade espacial. Por exemplo: duplo exponencial, exponencial com duplo esférico, linear com duplo esférico, entre outros.

As figuras, a seguir, são gráficos das funções de covariância encaixadas estimadas com sua função de covariância encaixada verdadeira. Em todos os casos foram usadas $n = 5000$, localizações $= 64$, $\sigma_1^2 = 0,6$, $\sigma_2^2 = 0,4$ e $\delta = 0,0001$.



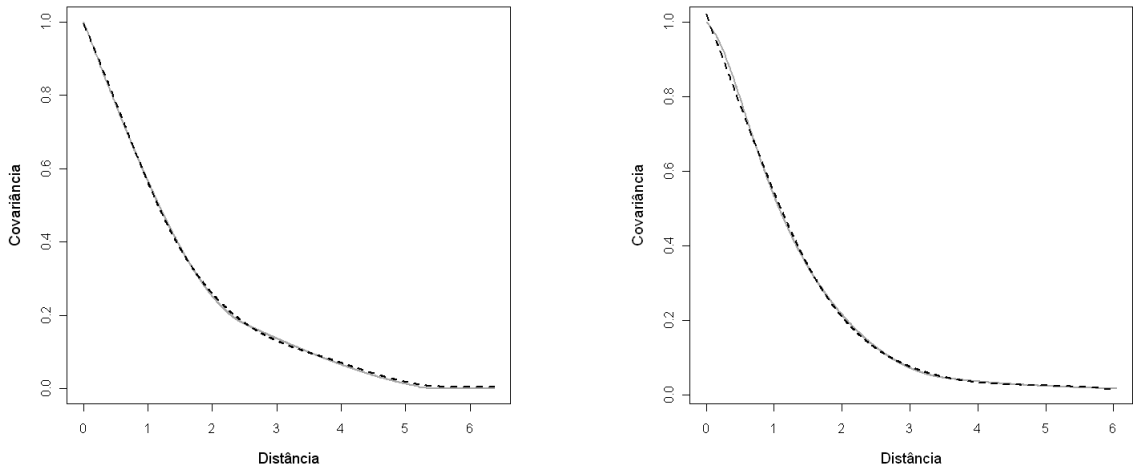
(a) (b)

Figura 5.21: Função de covariância encaixada: (a) esférica ($\phi_1 = 2$) com cúbica ($\phi_2 = 5$) e (b) esférica ($\phi_1 = 2,25$) com exponencial potência ($\phi_2 = 3,75; \kappa = 1,3$).



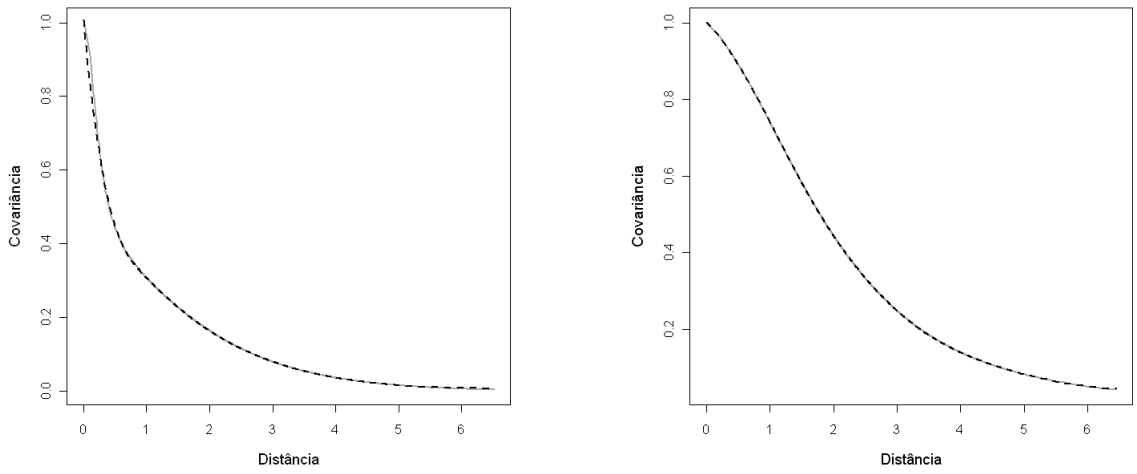
(a) (b)

Figura 5.22: Função de covariância encaixada: (a) exponencial potência ($\phi_1 = 1$) com gaussiana ($\phi_2 = 3$) e (b) circular ($\phi_1 = 2$) com exponencial potência ($\phi_2 = 5$).



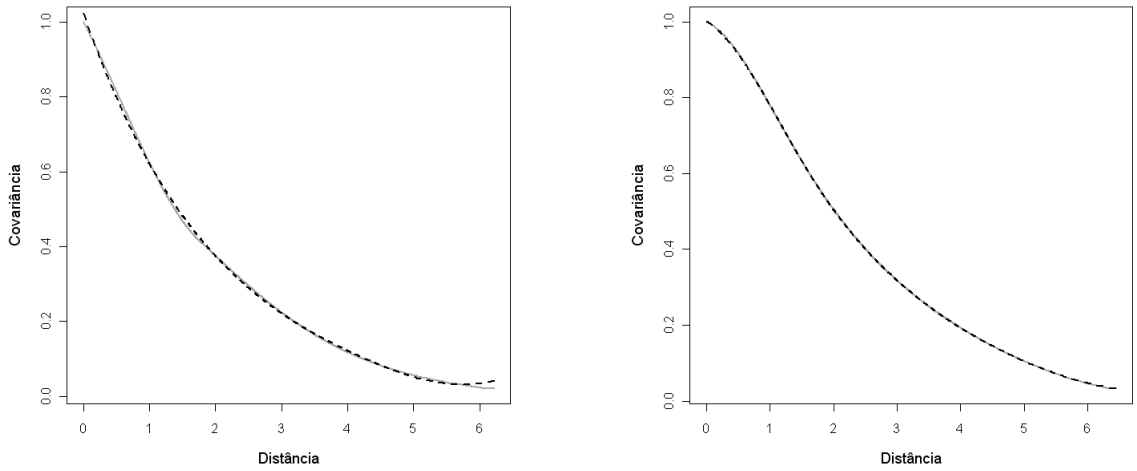
(a) (b)

Figura 5.23: Função de covariância encaixada: a) esférica ($\phi_1 = 2,5$) com circular ($\phi_2 = 5,5$) e b) quadrática racional ($\phi = 1; \kappa = 1$) com esférica ($\phi = 3,5$).



(a) (b)

Figura 5.24: Função de covariância encaixada: a) quadrática racional ($\phi_1 = 0,3; \kappa_1 = 1,5$) com Matérn ($\phi_2 = 1; \kappa_2 = 1,5$) e b) Matérn ($\phi = 0,75; \kappa_1 = 2,5$) com exponencial potência ($\phi = 2,75, \kappa_2 = 1$).



(a) (b)

Figura 5.25: Função de covariância encaixada: a) Matérn ($\phi_1 = 0,75; \kappa_1 = 4,5$) com esférica ($\phi_2 = 1,75$) e b) quadrática racional ($\phi_1 = 2; \kappa = 1,2$) com circular ($\phi_2 = 6,45$).

Capítulo 6

Considerações Finais

Dentro do contexto da geoestatística é comum querermos fazer inferência em um processo Gaussiano. Mas para isso é preciso caracterizar o processo. Surge então o maior desafio, que é obter a função de covariância do processo. Não se sabe qual a forma que a função de covariância assume. Há diversos estimadores propostos, aqueles obtidos por método dos momentos e outros obtidos pela construção de uma função *kernel*. O problema desses estimadores é que eles não são válidos, uma vez que a condição para ser definida positiva da função de covariância falha.

Neste trabalho foi proposto um estimador não-paramétrico para a função de covariância, sendo aproximado por combinações lineares de B-*splines* cúbicas. Ao contrário do que muitos autores fazem, que fixam o número de bases, aqui foi obtido o número de bases das funções B-*splines* por meio de um algoritmo. Os números de bases foram incrementados por um até alcançar um número ótimo, satisfazendo uma pseudo-distância de Hellinger chamada de afinidade. A afinidade mensura o quão próxima está a função de covariância

verdadeira da função de covariância estimada. Quanto mais próxima de um, melhor será o número de bases obtidas, pois o algoritmo vai parar quando as funções verdadeira e estimada forem semelhantes.

Com a finalidade de usar o teorema de Bochner foi calculado a transformada de Fourier do estimador. A partir dos cálculos observou-se que para a função de covariância estimada ser definida positiva bastaria resolver o sistema de quadrados mínimos com restrições nos coeficientes, não negativos. Entretanto, isto somente tem sentido quando a função de covariância fosse positiva. Quando a função de covariância possui valores negativos não se aplica o estimador, pois não garantimos que seja definida positiva, a menos que tenha tamanho amostral grande. Por exemplo, a estimativa da função de covariância ondular, que possui valores negativos, ficou boa para tamanho amostral pequeno, mas não podemos garantir que seja definida positiva. Nos estudos realizados, verificou-se que as estimativas das funções de covariâncias, assintoticamente, ficaram muito boas. Estimamos ainda funções de covariâncias no caso de processos com duas escalas diferentes a partir de um modelo encaixado. As estimativas obtidas foram muito boas.

O caso unidimensional pode ser estendido para múltiplas dimensões por meio da construção de produto tensorial das B-*splines*. Foi dado uma ideia de como seria essa extensão no Capítulo 2.

Referências Bibliográficas

[1] Banerjee, S. (2005). On geodetic distance computations in spatial modeling. *Biometrics*. Vol. 61, 617-625.

[2] Bjϕrnstad, O. N. e Falck, W. (2001). Nonparametric spatial covariance functions: Estimation and testing. *Environmental and Ecological Statistics*. Vol. 8, 53-70.

[3] Chilès, J. P. e Delfiner, P. (1999). *Geostatistics: Modeling Spatial Uncertainty*. Wiley, New York.

[4] Conover, W. J. (1999). *Practical Nonparametric Statistics*. 3.ed. John Wiley & Sons, New York.

[5] Cressie, N. (1993). *Statistics for Spatial Data*. John Wiley & Sons Inc, New York.

[6] De Boor, C. (2001). *A Practical Guide to Splines*. Revised Edition. Springer-Verlag, New York.

[7] Dias, R. (1999). Sequential adaptative nonparametric regression via h-spline. *Communications in Statistics. Simulation and Computation*. Vol. 28, Nº. 2, 501-515.

[8] Efron, B. e Tibshirani R. J. (1993). *An Introduction to the Bootstrap*. Chapman & Hall, New York.

[9] Eilers, P. H. C e Marx, B. D. (1996). Flexibible smoothing with b-splines and penalties. *Statistical Science*, Vol. 11, N°. 2, 89-121.

[10] Hall, P., Fisher, N. I. e Hoffmann, B. (1994). On the nonparametric estimation of covariance functions. *Annals of Statistics*. Vol. 22, 2115-2134.

[11] Hall, P. e Patil, P. (1994). Properties of nonparametric estimators of autocovariance for stationary random fields. *Probability Theory and Related Fields*. Vol. 99, N°. 3, 399-424.

[12] Hastie, T. e Tibshirani, R. (2000). Bayesian backfitting. *Statistical Science*. Vol. 15, N°. 3, 196-223.

[13] Hoeting, J. A., Davis, R. A., Merton, A. A. e Thompson, S. E. (2006). Model selection for geostatistical models. *Ecological Applications*. Vol. 6, N°. 1, 87-98.

[14] Hämmerlin, G. e Hoffmann, K. (1991). *Numerical Mathematics*. Springer-Verlag, New York.

[15] Isaaks, E. H. e Srivastava, R. M. (1989). *An Introduction to Applied Geostatistics*. Oxford University Press, New York.

[16] James, B. J. (2004). *Probabilidade: um curso em nível intermediário*. 3.ed. IMPA. Rio de Janeiro.

[17] Journel, A. G. e Huijbregts, Ch. J. (1978). *Mining Geostatistics*. Academy Press, New York.

[18] McBratney, A. G., Webster, R., McLaren, R. G. e Spiers, R. B. (1982). Regional variation of extractable copper and cobalt in topsoil and south-east Scotland. *Agronomie*. Vol. 2, Nº. 4, 969-982.

[19] McBratney, A. G. e Webster, R. (1986). Choosing functions for semi-variograms and fitting them to sampling estimates. *Journal of Soil Science*. Vol. 37, Nº. 3, 617-639.

[20] Mullen, K. M. e Stokkum, I. H. M. (2007). nnls: The Lawson-Hanson algorithm for non-negative least squares (NNLS). R package version 1.1.

[21] R Development Core Team. (2008). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. http://www.R-project.org, Vienna, Austria

[22] Ramsay, J. O. (1988). Monotone regression splines in action. *Statistical Science*. Vol. 3, Nº. 4, 425-461.

[23] Ribeiro Jr., P. J. e Diggle, P. J. (2000). *Model Based Geostatistics*. 14o SINAPE. ABE, Caxambu-MG.

[24] Ribeiro Jr., P. J. e Diggle, P. J. (2001). geoR: A package for geostatistical analysis. R-NEWS, Vol 1, Nº. 4, 15-18.

[25] Schumaker, L. L. (1972). *Spline Functions and Aproximation Theory*. Birkhauser.

[26] Schumaker, L. L. (1981). *Spline Functions: Basic Theory*. Wiley, New York.

[27] Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman & Hall, London.

[28] Wackernagel, H. (1995). *Multivariate Geostatistics*. Springer-Verlag Inc, Berlin Heidelberg.

[29] Yaglom, A. M. (1987). *Correlation Theory of Stacionary and Related Random Funtions I*. Springer-Verlag, New York.

[30] Yaglom, A. M. (1987). *Correlation Theory of Stacionary and Related Random Funtions II*. Supplementary Notes and References. Springer-Verlag, New York Inc.

Apêndice

Apêndice A

Programa utilizado nesta dissertação: *R-Gui versão 2.7.1*: algoritmo implementado para o estudo do comportamento da distribuição da afinidade e obtenção do ponto de corte.

```
library(splines)
library(MASS)
library(geoR)
#set.seed(8911)
## Funcao de Calculo dos B-splines
bsnormB<-function(x,xl,xr,ndx,bdeg=3){
        dx<-abs(xr-xl)/(ndx+1)
        knots<-sort(c(rep(xl-0.0*dx,3),seq(xl,xr,by=dx),rep(xr+0.0*dx,3)))
        x<-as.single(x)
        B<-spline.des(knots,x,4,0*x)$design
```

```
        list(B=B,knots=knots)
}
## Define Função Afinidade e Afinidade parcial
affparcial<-function(f,g){sum(abs(f*g)/sqrt(sum(f^2)*sum(g^2)))}
## Geracao dos Dados
p<-1000
q<-64
n<-1000
D<-matrix(0,q,q)
predito<-matrix(0,(q*q),p)
meanpred<-rep(0,length=(q*q))
x<-runif(q,min=0,max=5)
y<-runif(q,min=0,max=5)
     for(i in 1:length(x)-1){
         for(j in (i+1):length(y)){
            D[i,j]<-sqrt((x[i]-x[j])^2+(y[i]-y[j])^2)
         }
     }
## matriz simetrica, diag. princ. de zero
D<-D+t(D)
ord<-order(D)
## Construcao da Matriz de Covariancia
phi<-2.0
s2<-1
base<-0.001465
Dgrid<-seq(from=0,to=6,by=base)
```

```
tam<-length(Dgrid)
sqrt(tam)

SigmaGrid<-cov.spatial(matrix(Dgrid,ncol=sqrt(tam)),cov.model =
          "powered.exponential",cov.pars = c(s2, phi),kappa=1.5)
Sigma1<-cov.spatial(D,cov.model="powered.exponential",cov.pars=c(s2,phi),
          kappa=1.5)
TSig<-(array(Sigma1)*array(Sigma1))/(sum((array(Sigma1)*array(Sigma1))*base))

repet<-1000
mydir="hellinger"
for(j in 1:repet) {
 loop<-function(kzero){
  for(k in 1:p){
  ## normal multivariada, vetor de media 1 e matriz Cov. Sigma1(64x64)##
     z<-mvrnorm(n,rep(1,q),Sigma1)
     scorr<-cov(z)    ## matriz de covariancia
     scorr<-array(scorr)
     B<-bsnormB(D,min(D),max(D),ndx=kzero,bdeg=3)
     X<-B$B
     myfit<-lm(scorr~X)
     list(myfit=myfit)
     predito[,k]<-myfit$fitted.values
     list(predito=predito)
 }
 for(i in 1:(q*q)){
```

```
 meanpred[i]<-mean(predito[i,])
 }
 list(meanpred=meanpred)
}
  contador<-0
  d1<-0
  d2<-0
while(contador<=56){
   a<-loop(contador)
   estimativa1<-a$meanpred
 if(contador==0)
   d1<-0.1
   d2<-0.1
   contador<-contador+1
   a<-loop(contador)
   estimativa2<-a$meanpred
   d2<-affparcial(estimativa1,estimativa2)
   meanpred<-a$meanpred
   TSighat<-((a$meanpred)*(a$meanpred))/sum(((a$meanpred)*(a$meanpred))*base)
   d1<-sum(sqrt(TSighat*TSig))*base

   Y1<-d1
   Y2<-d2
write.table(Y1,paste(mydir,"affinity.txt",sep=""),append=TRUE,row.names=FALSE,
           col.names=FALSE)
write.table(Y2,paste(mydir,"parcial.txt",sep=""),append=TRUE,row.names=FALSE,
```

```
            col.names=FALSE)
  }
}
plot(D[ord],array(Sigma1)[ord],xlab="Distância",ylab="Covariância",type="l",
      col="dark grey",lwd=2,cex.lab=1.2)
lines(D[ord],meanpred[ord],col=1,lty=2,lwd=2)

Y1<-as.matrix(read.table("affinity.txt"))
Y11<-matrix(Y1,56,repet)
matplot(Y11,pch=20,xlab="Nós",ylab="Afinidade",col="dark grey",cex.lab=1.2)
Y2<-as.matrix(read.table("parcial.txt"))
Y22<-matrix(Y2,56,repet)
matplot(Y22,pch=20,xlab="Nós",ylab="Afinidade parcial",col="dark grey",
        cex.lab=1.2)

## afinidade
mafin<-mean(Y1)
afin<-c(colMeans(Y11))
alpha<-mafin*(((mafin*(1-mafin))/var(afin))-1)
betha<-(1-mafin)*(((mafin*(1-mafin))/var(afin))-1)
betaparam<-fitdistr(afin,'beta',list(shape1=alpha,shape2=betha))$estimate
beta<-rbeta(1000,shape1=betaparam[1],shape2=betaparam[2])
normparam<-fitdistr(afin,"normal")$estimate
norm<-rnorm(1000,mean=normparam[1],sd=normparam[2])
hist(afin,prob=T,main=paste(''),xlab='Afinidade',ylab='Densidade',
       cex.lab=1.2)
```

```
lines(sort(norm),dnorm(sort(norm),mean=normparam[1],sd=normparam[2]),
        col="8",lw=2)
lines(sort(beta),dbeta(sort(beta),betaparam[1],betaparam[2]),lty=2,lw=2)
lines(density(afin,bw='bcv'),col=1,lty=3,lw=2) #est. de dens. por kernel
```

Programa utilizado nesta dissertação: *R-Gui versão 2.7.1*: algoritmo implementado para o estudo do ponto de corte.

```
{
repet<-500
Y1<-as.matrix(read.table("pwexphellingerafinidade-p100q32n500ndx32r500
      phi2k15.txt"))
Y11<-matrix(Y1, 33, repet)

## afinidade
afin<-c(colMeans(Y11))
mafin<-mean(afin)
alpha<-mafin*(((mafin*(1-mafin))/var(afin))-1)
betha<-(1-mafin)*(((mafin*(1-mafin))/var(afin))-1)
q_hat<-qbeta(0.01,alpha,betha,ncp=0,lower.tail=FALSE,log.p=FALSE)

mafin<-c(0,0)
alpha<-c(0,0)
betha<-c(0,0)
q_boot<-c(0,0)
```

```
B<-2000
n<-500
boot <-matrix(0,B,n)
for(i in 1:B){
  for(j in 1:n){
      boot[i,]<-sample(afin,replace=T)
  }
}
for(i in 1:B){
 mafin[i]<-mean(boot[i,])
 alpha[i]<-mafin[i]*(((mafin[i]*(1-mafin[i]))/var(boot[i,]))-1)
 betha[i]<-(1-mafin[i])*(((mafin[i]*(1-mafin[i]))/var(boot[i,]))-1)
 q_boot[i]<-qbeta(0.01,alpha[i],betha[i],ncp=0,lower.tail=FALSE,log.p=FALSE)
}
q_til<-mean(q_boot)
q_var<-var(q_boot)
vies<-q_til-q_hat
boxplot(q_boot)
}
```

Apêndice B

Programa utilizado nesta dissertação: *R-Gui versão 2.7.1*: algoritmo implementado para a estimação da função de covariância.

```
{
library(splines)
```

```
library(MASS)
library(geoR)
library(nnls)
## Funcao de Calculo dos B-splines ##
bsnormB<-function(x,xl,xr,ndx,bdeg=3){
     dx<-abs(xr-xl)/(ndx+1)
  knots<-sort(c(rep(xl-0.0*dx,3),seq(xl,xr,by=dx),rep(xr+0.0*dx,3)))
  x    <-as.single(x)
  B    <-spline.des(knots,x,4,0*x)$design
  list(B=B,knots=knots)
}
## Define Função Afinidade ##
affinity<-function(f,g){sum(abs(f*g)/sqrt(sum(f^2)*sum(g^2)))}
## Geracao dos Dados ##
p<-1000
q<-64
n<-1000
D<-matrix(0, q, q)
predito<-matrix(0,(q*q),p)
meanpred<-rep(0,length=(q*q))
x<-runif(q,min=0,max=5)
y<-runif(q,min=0,max=5)
     for(i in 1:length(x)-1){
         for(j in (i+1):length(y)){
             D[i,j]<-sqrt((x[i]-x[j])^2+(y[i]-y[j])^2)
         }
```

```
     }
## matriz simetrica, diag princ de zero ##
D<-D+t(D)
ord<-order(D)
## Construcao da Matriz de Covariancia ##
## parametros da covariancia ##
phi<-1.6
s2<-1
##1# matern ##
Sigma1<-cov.spatial(D,cov.model="matern",cov.pars=c(s2,phi),kappa=1.5)
##02# powered.exponential (ou estavel) ##
#Sigma1<-cov.spatial(D,cov.model="stable",cov.pars=c(s2,phi),kappa=1)
##03# Cauchy (ou quradratica racional) ##
#Sigma1<-cov.spatial(D,cov.model="cauchy",cov.pars=c(s2,phi),kappa=0.5)
##04# gencauchy (generalised Cauchy)##
#Sigma1<-cov.spatial(D,cov.model="gencauchy",cov.pars=c(s2,phi),kappa=1)
##07# exponential ##
#Sigma1<-cov.spatial(D,cov.model="exponential",cov.pars=c(s2,phi))
##08# gaussian ##
#Sigma1<-cov.spatial(D,cov.model="gaussian",cov.pars=c(s2,phi))
##09# spherical ##
#Sigma1<-cov.spatial(D,cov.model="spherical",cov.pars=c(s2,phi))
##10# circular ##
#Sigma1<-cov.spatial(D,cov.model="circular",cov.pars=c(s2,phi))
##11# cubic ##
#Sigma1<-cov.spatial(D,cov.model="cubic",cov.pars=c(s2,phi))
```

```
##12# wave ##
Sigma1<-cov.spatial(D,cov.model="wave",cov.pars=c(s2,phi))


loop<-function(kzero){
   for(k in 1:p){
   ##normal multivariada, vetor de media 1 e matriz Cov. Sigma1(64x64)##
      z<-mvrnorm(n,rep(1,q),Sigma1)
      scorr<-cov(z)             ## matriz de covariancia ##
      scorr<-c(array(scorr))
      BS<-bsnormB(D,min(D),max(D),ndx=kzero)
      B<-BS$B
      myfit<-nnls(B,scorr)
      predito[,k]<-myfit$fitted
   }
   for(i in 1:(q*q)){
      meanpred[i]<-mean(predito[i,])
   }
   list(meanpred=meanpred)
}
delta<-0.0001
contador<-1
d<-0
while(d < 1-delta & contador <= 15){
      a<-loop(contador)
      estimativa1<-a$meanpred
      if(contador == 1)
```

```
        d<-0.1
        contador<-contador + 1
        a<-loop(contador)
        estimativa2<-a$meanpred
        d<-affinity(estimativa1,estimativa2)
        print(c('A afinidade:',d))
        print(c('no. bases:',contador))
        meanpred<-a$meanpred
}
plot(D[ord],array(Sigma1)[ord],xlab="Distância",ylab="Covariância",
        type="l",col=8,lwd=2)
lines(D[ord],meanpred[ord],col=1,lty=2,lwd=1)
}
```