Estimação de Tipologia para Dados Funcionais Agrupados

Este exemplar corresponde à redação final da dissertação devidamente corrigida e defendida por Angelo Martarelli Filho e aprovada pela comissão julgadora.

Campinas, 07 de Abril de 2006.

Profa. Dra. Nancy Lopes Garcia
Orientadora

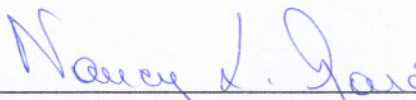
Prof. Dr. Ronaldo Dias
Co-orientador

Banca Examinadora
1. Profa. Dra. Nancy Lopes Garcia (Orientadora) – IMECC/UNICAMP.
2. Prof. Dr. Dani Gamerman – DME/UFRJ.
3. Prof. Dr. Armando Milioni – ITA.

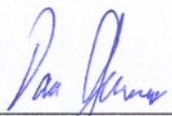
Dissertação apresentada ao Instituto de Matemática, Estatística e Computação Científica, UNICAMP, como requisito parcial para obtenção do Título de Mestre em Estatística.

Dissertação de Mestrado defendida em 07 de abril de 2006 e aprovada

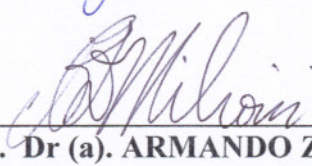
Pela Banca Examinadora composta pelos Profs. Drs.



Prof. (a). Dr (a). NANCY LOPES GARCIA



Prof. (a). Dr (a). DANI GAMERMAN



Prof. (a). Dr (a). ARMANDO ZEFERINO MILIONI

FICHA CATALOGRÁFICA ELABORADA PELA BIBLIOTECA DO IMECC DA UNICAMP
Bibliotecária: Maria Júlia Milani Rodrigues – CRB8a / 2116

Martarelli Filho, Angelo
M36e Estimação de tipologia para dados funcionais agrupados / Angelo Martarelli Filho -- Campinas, [S.P. :s.n.], 2006.

Orientadores : Nancy Lopes Garcia; Ronaldo Dias
Dissertação (mestrado) - Universidade Estadual de Campinas, Instituto de Matemática, Estatística e Computação Científica.

1. Estatística não paramétrica. 2. Modelos lineares (Estatística). 3. Análise multivariada. I. Garcia, Nancy Lopes. II. Dias, Ronaldo. III. Universidade Estadual de Campinas. Instituto de Matemática, Estatística e Computação Científica. VI. Título.

Título em inglês: Typology estimation for grouped functional data

Palavras-chave em inglês (Keywords): 1. Nonparametric statistics. 2. Linear models (Statistics). 3. Multivariate analysis.

Área de concentração: Estatística computacional, Sub-área: Estimação não-paramétrica

Titulação: Mestre em Estatística

Banca examinadora: Prof. Dr. Dani Gamerman (DME-UFRJ)
Prof. Dr. Armando Milioni (ITA)
Profa. Dra. Nancy Lopes Garcia (IMECC-UNICAMP)
Prof. Dr. Filidor Vilca-Labra (IMECC-UNICAMP)

Data da defesa: 07/04/2006

UNIVERSIDADE ESTADUAL DE CAMPINAS

INSTITUTO DE MATEMÁTICA, ESTATÍSTICA E COMPUTAÇÃO CIENTÍFICA

DEPARTAMENTO DE ESTATÍSTICA

Estimação de Tipologia para Dados Funcionais Agrupados

Angelo Martarelli Filho

Orientadora: Profa. Dra. Nancy Lopes Garcia

Co-orientador: Prof. Dr. Ronaldo Dias

Campinas – SP

2006

Aos meus pais, Angelo e Beatriz.

Agradecimentos

Agradeço aos meus pais, Angelo Martarelli e Beatriz Scaramucci Fontes Martarelli, por me apoiarem em todas as minhas decisões e me darem todo o suporte necessário. Um sincero agradecimento também a Priscila, minha querida irmã.

A Daisy, pelo apoio e compreensão por eu estar longe a maior parte do tempo.

Aos meus orientadores, Profa. Dra. Nancy Lopes Garcia e Prof. Dr. Ronaldo Dias, pela paciência e confiança em meu trabalho.

Aos meus amigos, da graduação em Estatística, mestrado em Estatística e da minha terra natal (Gália-SP), que, de uma forma ou de outra, contribuíram para o meu sucesso neste trabalho.

Aos que participaram de forma mais constante no meu dia-a-dia, seja no laboratório, em casa ou num momento de descontração: André, Benilton Carvalho, Camila Boreli, Camila Esteves, Daniel Takata, Fernanda Gadine, Fernando Ferraz, Gabriel Coelho, Guilherme Pereira, Jaqueline Barbão, Lori Grandin, Luciana Fontes, Melissa Souza, Pleyciene, Rodrigo Tsai, Sergio Pimentel e Thais Castro.

Ao Departamento de Estatística pelo amparo e suporte.

Aos funcionários da Secretaria de Pós-graduação do IMECC representados na pessoa da Sra. Tânia.

À CAPES, pelo apoio financeiro, fundamental para o desenvolvimento desse projeto.

Resumo

Neste trabalho abordamos o problema de estimação de dados funcionais quando as curvas não são observadas individualmente. Temos uma população dividida em subpopulações de tamanho conhecido, e as observações são somas de todas as observações funcionais individuais em todas as subpopulações observadas a intervalos de tempo fixos. Utilizando expansão em bases B-splines, é possível recuperar a curva média de cada subpopulação (tipologia), bem como a estrutura de variância e covariância das curvas. Estudos de simulação sugerem que o método estima bem as curvas mesmo com poucas replicações e é assintoticamente consistente. Aplicações para um problema real de curvas de carga de energia elétrica são apresentadas.

Abstract

In this work we address the problem of estimating functional data when the curves are not individually observed. That is, the observations are the sum of all curves for the individuals in the population. Consider a population divided into subpopulations of known sizes. The objective of this work is to estimate the mean curve for each subpopulation (tipology) as well as the covariance structure. We propose an estimation method based on B-splines expansion. Simulation studies suggest that the method is suitable even with few replications. Moreover, it appears to be consistent. Application to a real data set is presented.

Sumário

Lista de Tabelas

Lista de Figuras

Introdução

A filosofia básica na análise de dados funcionais é considerar os dados observados de funções como entidades únicas, ao invés de meramente uma seqüência de observações individuais. O termo *funcional* refere-se à estrutura intrínseca dos dados mas, na prática, eles são usualmente observados e registrados discretamente. O registro de uma observação funcional x consiste em T pares (t_j, z_j), onde z_j equivale a $x(t_j)$ para $j = 1, 2, \ldots, T$.

Existem técnicas muito difundidas na literatura para a análise de dados funcionais quando as curvas são observadas individualmente, como a análise de variância funcional e a representação por funções alisadas, que podem ser vistas em Ramsay e Silverman (1997). Em alguns casos, não conhecemos ou não é interessante medir as observações funcionais para cada indivíduo da amostra, mas podemos obter facilmente observações funcionais referentes a soma de observações individuais. Ou seja, conseguimos observar funções referentes a grupos de indivíduos não necessariamente do mesmo tipo. Por exemplo, na distribuição de energia elétrica em cidades, pode ser conveniente medir a demanda de energia em transformadores de energia (responsáveis pelo fornecimento de energia a um grupo de consumidores), e não diretamente nas residências, comércios entre outros.

Um problema real apresentado por uma companhia de distribuição de energia elétrica da região de Campinas e a restrita quantidade de teoria para analisar dados funcionais agrupados motivaram o desenvolvimento deste trabalho. O interesse da companhia é estimar a curva típica (tipologia) e a estrutura de variância e covariância para os diferentes tipos de consumidores, conhecendo apenas as *curvas de consumo* de energia observadas em transformadores de energia (variável resposta funcional agrupada) e seus *mercados* (conjunto de freqüências que os diferentes tipos de consumidores apresentam em um determinado transformador). Um método capaz de estimar as quantidades de interesse

usando apenas estas duas informações é de grande interesse para a companhia, pois elas são facilmente observadas e com custo baixo. Neste caso, a companhia só precisaria observar as curvas nos transformadores (variável resposta), pois seus mercados são conhecidos.

Neste trabalho, optamos por uma abordagem não paramétrica através da representação em funções de base, pois a quantidade de parâmetros diminui consideravelmente e obtemos estimativas de funções diretamente. Para a estimação das tipologias, colocamos o problema como um sistema linear utilizando bases B-splines. Na estimação das superfícies de variância e covariância para as tipologias, encontramos um sistema linear em função do produto tensorial de B-splines. Com isso, conseguimos propor um modelo linear com duas equações.

No Capítulo 1, fizemos uma breve descrição do problema apresentado pela companhia de distribuição de energia elétrica. Em seguida, introduzimos a teoria de splines e descrevemos o método de estimação para as tipologias e superfície de variância e covariância. O Capítulo 2 foi dedicado ao estudo de simulações, onde criamos vários cenários para observar o comportamento do modelo quanto ao número de replicações (dias), consumidores e transformadores. No Capítulo 3, aplicamos a metodologia em dados fornecidos pela companhia. Colocamos em apêndice o programa em R que foi utilizado neste trabalho.

1 *Dados Funcionais Agrupados por Sub-Populações*

1.1 Descrição do problema

Os consumidores de energia elétrica da região de Campinas se dividem em diversos subgrupos os quais são classificados de acordo com seu ramo de atividade, destacando-se os consumidores *residenciais*, *comerciais* e *industriais*.

Considere que a nossa população de interesse é dividida em micro-regiões com aproximadamente 50 consumidores. Cada micro-região é abastecida por um transformador responsável pela distribuição de energia elétrica para os seus respectivos consumidores. Toda a carga distribuída por um transformador é medida continuamente e registrada a cada 15 minutos. Por exemplo, a i-ésima micro-região poderia ser composta por 50 consumidores, onde 10 são residenciais, 15 comerciais e 25 industriais. Neste exemplo, o consumo total $Y_i(t)$ no tempo t para o i-ésimo transformador seria descrito pela equação,

$$Y_i(t) = \sum_{n_1=1}^{10} R_{i,n_1}(t) + \sum_{n_2=1}^{15} C_{i,n_2}(t) + \sum_{n_3=1}^{25} I_{i,n_3}(t),$$

onde $R_{i,n_1}(t)$, $C_{i,n_2}(t)$ e $I_{i,n_3}(t)$ representam o consumo individual do n_1-ésimo, n_2-ésimo, n_3-ésimo consumidor residencial, comercial e industrial respectivamente.

Cada micro-região (transformador) possui um mercado diferente, ou seja, tem diferentes freqüências de consumidores. No exemplo acima, o mercado seria 10 consumidores residenciais, 15 comerciais e 25 industriais. O domínio das funções estudadas no pro-

blema é o intervalo $[0, 24]$, que está expresso em horas e corresponde a um dia. Devido a dificuldade de observar funções contínuas na prática, os dados foram registrados a cada 15 minutos perfazendo um total de 96 observações diárias. A Figura 1.1.1 ilustra uma possível observação funcional de demanda de energia num transformador.

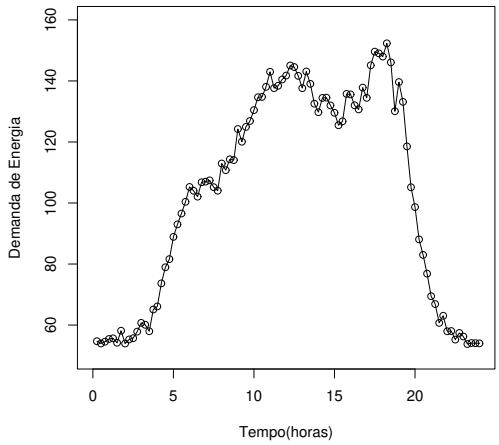


Figura 1.1.1: Exemplo fictício de demanda de energia num transformador.

A companhia coleta as observações funcionais agrupadas em dias consecutivos (dados cíclicos), ou seja, para cada transformador são observadas seqüências de $J \times 96$ pontos, referentes a J dias (observações funcionais). Portanto, os pontos observados no final de um dia são correlacionados com os primeiros pontos do dia seguinte. Este fato faz com que exista correlação entre observações funcionais subsequentes, entretanto não vamos levar este fato em consideração na modelagem, pois a companhia de energia não tem interesse nas extremidades das observações funcionais pois a demanda de energia nestes dois períodos é pequena e praticamente constante, não justificando a complicação do modelo.

A companhia gostaria de obter uma metodologia capaz de estimar a tipologia e uma medida de dispersão para cada tipo de consumidor, utilizando apenas os dados funcionais agrupados observados nos transformadores e seus mercados. Neste trabalho, descreveremos um método não paramétrico para o problema que utiliza a representação em bases splines.

1.2 O Espaço de splines

Devido à sua estrutura simples e boas propriedade de aproximação, polinômios algébricos são amplamente usados, na prática, na aproximação de funções. A precisão dessa aproximação depende essencialmente do grau do polinômio e do tamanho do intervalo considerado $[a, b]$. Visto que o custo com operações computacionais aumenta conforme o grau do polinômio, é aconselhável usar polinômios de grau reduzido. Essencialmente, na busca da precisão desejada, nós temos que nos restringir a intervalos pequenos. Com este propósito, usualmente dividimos o intervalo original $[a, b]$ em subintervalos suficientemente pequenos $\{[x_k, x_{k+1}]\}_{k=0}^{K}$, e então usamos polinômios com grau reduzido p_k para aproximar a função desejada sobre $[x_k, x_{k+1}]$, $k = 0, \ldots, K$. Esse procedimento produz uma função de aproximação polinomial por partes $s(x)$,

$$s(x) \equiv p_k(x), \quad \text{para} \quad x \in [x_k, x_{k+1}], \qquad k = 0, \ldots, K.$$

Em geral, os polinômios $\{p_k(x)\}$ não são construídos independentemente uns dos outros. Uma propriedade desejável é que eles se unam de forma suave em $x_1, \ldots, x_n$, isto é, todas as derivadas de p_{k-1} e p_k, acima de uma certa ordem, coincidam em x_k. Como resultado nós introduziremos um alisamento, função polinomial por partes, chamado função spline. Mais detalhes sobre splines podem ser encontrados em de Boor (1978), Eubank (1988) e Green e Silverman (1994).

Definição 1. *A função $s(x)$ é chamada de função spline no intervalo $[a, b]$ (ou simplesmente "spline") de grau r com nós $\{x_k\}_1^K$ se $a =: x_0 \leq x_1 \leq \cdots \leq x_K \leq x_{K+1} := b$ e*

1. *para cada $k = 0, \ldots, K$, $s(x)$ coincide em (x_k, x_{k+1}) com um polinômio de grau menor ou igual a r;*

2. *$s(x), s'(x), \ldots, s^{r-1}(x)$ são funções contínuas em (a, b).*

Todo polinômio algébrico é uma função spline com nós $\{a, b\}$. Nota-se da definição

que a r-ésima derivada de um spline de grau r com nós $\{x_k\}_1^K$ é uma função constante por partes com quebras, eventualmente, em $x_1, \ldots, x_K$. Reciprocamente, a r-ésima função primitiva de uma função constante por partes é um spline de grau r.

Nós denotamos por $S_r(x_1, \ldots, x_K)$ a classe de todas as funções spline de grau r com nós em $x_1, \ldots, x_K$. Claramente, fixados $\{x_k\}_1^K$, $S_r(x_1, \ldots, x_K)$ é um espaço linear.

Um exemplo simples de função spline é a chamada função poder truncada, dada por

$$(x-t)_+^r := \begin{cases} (x-t)^r & para \;\; x \geq t \\ 0 & para \;\; x < t. \end{cases}$$

1.3 Modelo Funcional

Utilizaremos a terminologia introduzida na Seção 1.1 para facilitar a descrição da metodologia. Nesta seção, vamos propor um modelo linear para estimar a curva típica de consumo (tipologia) e a superfície de variância e covariância para o c-ésimo tipo de consumidor, onde $c = 1, 2, \ldots, C$. Com esta abordagem, conseguimos estimar as quantidades de interesse conhecendo apenas as demandas de energia observadas em transformadores (dados funcionais agrupados) e seus mercados (freqüência com que cada tipo de consumidor aparece nos transformadores). Como podemos perceber, não observamos as curvas de consumo de energia individuais em momento algum, portanto com este procedimento conseguimos estimar as tipologias e suas medidas de dispersão para os diferentes tipos de consumidores sem conhecimento prévio do comportamento individual das curvas e superfícies. Para que o modelo proposto apresente solução, o número de transformadores observados tem que ser maior ou igual a C (número de tipos de consumidores). Outra exigência é que os transformadores possuam mercados com proporções diferentes.

A Equação (1.3.1) mostra uma expressão para o problema geral. Em nossa abordagem, as variáveis $y_{c,j,n_c}(t)$ representam o consumo de energia do n_c-ésimo consumidor, no tempo t, do tipo c, no dia j e serão consideradas variáveis aleatórias funcionais independentes para $j = 1, 2, \ldots, J$ e identicamente distribuídas para c fixo. As constantes $N_{1,i}, \ldots, N_{C,i}$

caracterizam o mercado do transformador i. Portanto, a demanda do transformador i, no dia j e no tempo t pode ser representada por,

$$Y_{i,j}(t) = \sum_{c=1}^{C} \sum_{n_c=1}^{N_{c,i}} y_{c,j,n_c}(t). \tag{1.3.1}$$

Neste trabalho, estamos denotando como *tipologia* do consumidor tipo c, $c = 1, 2, \ldots, C$, sua curva média ou típica, onde todas as observações individuais dos consumidores tipo c podem ser consideradas como uma soma de sua tipologia mais uma perturbação aleatória. Com isso, podemos escrever o consumo individual de energia elétrica conforme a expressão,

$$y_{c,j,n_c}(t) = \alpha_c(t) + \varepsilon_{c,j,n_c}(t),$$

onde $\alpha_c(t)$ é a tipologia do consumidor tipo c e $\varepsilon_{c,j,n_c}(t)$ uma perturbação aleatória. Vamos considerar neste trabalho que $\varepsilon_{c,j,n_c}(t)$ são processos gaussianos de média zero independentes para $j = 1, 2, \ldots, J$ e identicamente distribuídos para c fixo. Com isso, conseguimos postular um modelo linear para as tipologias $\alpha_c(t)$, $c = 1, \ldots, C$, cuja equação envolva apenas as observações funcionais agrupadas coletadas em transformadores de energia e seus mercados. Assim,

$$Y_{i,j}(t) = \sum_{c=1}^{C} N_{c,i}\alpha_c(t) + \boldsymbol{\varepsilon}_{i,j}(t), \tag{1.3.2}$$

onde

$$\boldsymbol{\varepsilon}_{i,j}(t) = \sum_{c=1}^{C} \sum_{n_c=1}^{N_{c,i}} \varepsilon_{c,j,n_c}(t). \tag{1.3.3}$$

Como exemplo, a Expressão (1.3.2) para $c = 2$ é,

$$Y_{i,j}(t) = N_{1,i}\alpha_1(t) + N_{2,i}\alpha_2(t) + \boldsymbol{\varepsilon}_{i,j}(t),$$

ou na forma matricial,

$$Y_{i,j}(t) = (N_{1,i} \quad N_{2,i}) \begin{pmatrix} \alpha_1(t) \\ \alpha_2(t) \end{pmatrix} + \boldsymbol{\varepsilon}_{i,j}(t).$$

Com a ajuda da Expressão (1.3.3), percebemos que os erros $\boldsymbol{\varepsilon}_{i,j}(t)$ resultam da somatória de erros provenientes dos C tipos de consumidores. O número de integrantes de cada tipo de consumidor na somatória é definido pelos mercados dos transformadores, ou seja, pelos valores de $N_{c,i}$, $c = 1, 2, \ldots, C$. Como os erros $\varepsilon_{c,j,n_c}(t)$ são considerados processos gaussianos de média zero independentes para $j = 1, 2, \ldots, J$ e identicamente distribuídos para c fixo, as hipóteses de independência e homoscedasticidade continuam sendo satisfeitas para i fixo, mas as curvas de diferentes transformadores não possuem mais a mesma estrutura de variância e covariância devido a diferença entre os mercados $N_{c,i}$, $i = 1, 2, \ldots, I$.

A Expressão (1.3.2) é bastante conhecida na área de análise de regressão. Nestes casos, geralmente usa-se a técnica de mínimos quadrados ordinários para estimar os parâmetros, mas a eficiência deste método está condicionada à adequação de algumas hipóteses, onde as mais importantes são a independência e homoscedasticidade dos erros. Como foi visto, a hipótese de homoscedasticedade não é satisfeita, portanto não podemos usar mínimos quadrados ordinários neste caso. Uma alternativa é utilizar mínimos quadrados generalizados, com essa técnica conseguimos incorporar a estrutura de variância e covariância dos erros à análise.

Até este momento, definimos a primeira equação do modelo que estamos propondo, que é responsável pela estimação das tipologias. Queremos encontrar agora uma segunda equação para estimar uma medida de dispersão para as tipologias obtidas através da Equação (1.4.5). Na estimação de um parâmetro pontual, geralmente usamos a variância como medida de precisão. Quando trabalhamos com curvas, uma maneira análoga de medir a precisão é estimar uma superfície de variância e covariância. Para facilitar a notação, usaremos expressão vetorial,

$$\vec{Y}_i(t) = \sum_{c=1}^{C} \sum_{n_c=1}^{N_{c,i}} \vec{y}_{c,n_c}(t), \tag{1.3.4}$$

onde,

$$\vec{Y}_i(t) = \begin{pmatrix} Y_{i,1}(t) \\ Y_{i,2}(t) \\ \vdots \\ Y_{i,J}(t) \end{pmatrix}_J \quad \text{e} \quad \vec{y}_{c,n_c}(t) = \begin{pmatrix} y_{c,1,n_c}(t) \\ y_{c,2,n_c}(t) \\ \vdots \\ y_{c,J,n_c}(t) \end{pmatrix}_J ,$$

para $c = 1, 2, \ldots, C$ e $t = t_1, t_2, \ldots, t_T$.

Com isso, a matriz de variância e covariância para o i-ésimo transformador pode ser escrita como,

$$Z_i = \begin{pmatrix} Cov[\vec{Y}_i(t_1), \vec{Y}_i(t_1)] & Cov[\vec{Y}_i(t_1), \vec{Y}_i(t_2)] & \ldots & Cov[\vec{Y}_i(t_1), \vec{Y}_i(t_T)] \\ Cov[\vec{Y}_i(t_2), \vec{Y}_i(t_1)] & Cov[\vec{Y}_i(t_2), \vec{Y}_i(t_2)] & \ldots & Cov[\vec{Y}_i(t_2), \vec{Y}_i(t_T)] \\ \vdots & \vdots & \ddots & \vdots \\ Cov[\vec{Y}_i(t_T), \vec{Y}_i(t_1)] & Cov[\vec{Y}_i(t_T), \vec{Y}_i(t_2)] & \ldots & Cov[\vec{Y}_i(t_T), \vec{Y}_i(t_T)] \end{pmatrix}_{T \times T} ,$$

onde

$$Z_i(t, s) = Cov[\vec{Y}_i(t), \vec{Y}_i(s)],$$

para t e $s = t_1, t_2, \ldots, t_T$.

Quando estimamos Z_i, na verdade estamos estimando pontos da superfície de variância e covariância para as observações funcionais agrupadas do i-ésimo transformador, $i = 1, 2, \ldots, I$. Análogo ao problema encontrado na estimação das tipologias, podemos estimar a matriz de variância e covariância $\hat{Z}_i$, mas o nosso objetivo é estimar uma superfície de variância e covariância para cada tipo de consumidor conhecendo apenas $\hat{Z}_i$ e os mercados. Portanto, o primeiro passo em busca de uma solução é estabelecer uma relação entre a matriz $\hat{Z}_i$ e as matrizes de variância e covariância para cada tipo de consumidor, $c = 1, 2, \ldots, C$. Para isso, podemos escrever a $Cov[\vec{Y}_i(t), \vec{Y}_i(s)]$ como função dos consumos individuais com a ajuda da Expressão (1.3.4),

$$Cov[\vec{Y}_i(t), \vec{Y}_i(s)] = Cov\left[\sum_{c=1}^{C}\sum_{n_c=1}^{N_{c,i}} \vec{y}_{c,n_c}(t), \sum_{c=1}^{C}\sum_{n_c=1}^{N_{c,i}} \vec{y}_{c,n_c}(s)\right].$$

onde t e $s = t_1, t_2, \ldots, t_T$. Após algumas manipulações algébricas, chegamos à relação,

$$Cov[\vec{Y}_i(t), \vec{Y}_i(s)] = \sum_{c=1}^{C} N_{c,i} Cov[\vec{y}_c(t), \vec{y}_c(s)]. \tag{1.3.5}$$

Como podemos ver na expressão acima, encontramos uma relação linear entre as covariâncias. Uma maneira de postular uma equação para estimar a superfície de variância e covariância para cada tipo de consumidor, é usando a seguinte relação,

$$\hat{Z}_i(t, s) = Z_i(t, s) + \xi_i(t, s),$$

onde supomos que os erros $\xi_i(t_1, t_2)$ têm distribuição normal e são independêntes e identicamente distribuídos.

Para facilitar a notação, definiremos $\delta_c(t, s) = Cov[\vec{y}_c(t), \vec{y}_c(s)]$ para $c = 1, 2, \ldots, C$. Com isso, podemos postular a equação linear,

$$\hat{Z}_i(t, s) = \sum_{c=1}^{C} N_{c,i} \delta_c(t, s) + \xi_i(t, s). \tag{1.3.6}$$

Com a abordagem proposta pela Expressão (1.3.6), cada transformador produz apenas um conjunto de estimativas ($\hat{Z}_i(t, s)$, para t e $s = t_1, t_2, \ldots, t_T$) que serão usadas como resposta no modelo. Entretanto, cada transformador fornece uma relação diferente (caracterizada pelos mercados $N_{c,i}$, $i = 1, 2, \ldots, I$) entre as respostas $\hat{Z}_i(t, s)$ e os parâmetros de interesse $\delta_c(t, s)$, $c = 1, 2, \ldots, C$. Com isso, o aumento do número de transformadores melhora a qualidade do ajuste das superfícies de variância e covariância. Análogo ao que acontece em álgebra linear, o modelo só apresenta solução quando o número de transformadores for maior ou igual ao de tipos de consumidores ($I \geq C$) e os mercados dos transformadores apresentam freqüências relativas diferentes. No Capítulo 2, estudaremos o comportamento do modelo com relação ao número de transformadores através de simulações.

1.4 Estimação por Expansão em Bases

Vamos supor que são observados T pontos por dia para representar as observações funcionais agrupadas, neste caso a estimação das tipologias através da Expressão (1.3.2) implicaria na estimação de $C \times T$ parâmetros, onde a c-ésima tipologia seria representada pelos pontos estimados $\hat{\alpha}_c(t)$, para $t = t_1, t_2, \ldots, t_T$. Com esta abordagem, o número de parâmetros necessários para representar as tipologias é muito grande e cresce proporcionalmente com T, prejudicando assim a qualidade das estimativas. Neste seção, vamos propor uma abordagem alternativa através da representação das tipologias em um espaço de funções, como o espaço de splines cúbicos $S_3(x_1, \ldots, x_K)$ introduzido da Seção 1.2, que além de boas propriedades de aproximação, diminui consideravelmente o número de parâmetros, comparado com a estimação dos $\alpha_c(t)$ através da Expressão (1.3.2). Além disso, fornece estimativas funcionais, ao invés de apenas T pontos isolados para representar cada tipologia. Nós iremos utilizar as bases B-splines, mais detalhes sobre B-splines podem ser encontrados em de Boor (1978), Eubank (1988) e Green e Silverman (1994). A Expressão (1.4.1) mostra a representação dos $\alpha_c(t)$, $t \in [0, 24]$, como uma combinação linear de K bases B-splines,

$$\alpha_c(t) = \sum_{k=1}^{K} b_{c,k} B_k(t), \quad \text{para} \quad c = 1, 2, \ldots, C. \tag{1.4.1}$$

onde K é o número de bases, $B_k(t)$ representa a k-ésima base B-Splines avaliada no ponto t e $b_{c,k}$ são os coeficientes a serem determinados. A escolha do número de nós e suas posições são de grande importância, pois influenciam diretamente na forma e posição das funções de bases e consequentemente nos resultados obtidos. Nesta abordagem, estamos supondo que as tipologias podem ser escritas como uma combinação linear de funções de base. Note que esta é uma abordagem não paramétrica. Os coeficientes $b_{c,k}$ não são parâmetros, pois dependem dos nós e estes são arbitrários, não determinados pelo modelo probabilístico. A Figura (1.4.1) mostra as cinco funções B-splines obtidas com um nó interno localizado no centro do intervalo $[0, 24]$.

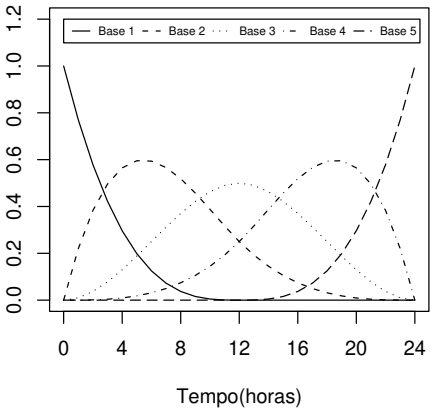


Figura 1.4.1: Gráfico com as cinco funções B-splines obtidas com um nó interno localizado no centro do intervalo [0,24].

Colocando a Expressão (1.4.1) em forma matricial, conseguimos ver melhor a relação existente entre os $\alpha_c(t)$, $t = t_1, t_2, \ldots, t_T$, e os $b_{c,k}$, $k =, 1, 2, \ldots, K$.

$$\begin{pmatrix} \alpha_c(t_1) \\ \vdots \\ \alpha_c(t_T) \end{pmatrix}_T = \begin{pmatrix} B_1(t_1) & \cdots & B_K(t_1) \\ \vdots & & \vdots \\ B_1(t_T) & \cdots & B_K(t_T) \end{pmatrix}_{T\times K} \begin{pmatrix} b_{c,1} \\ \vdots \\ b_{c,K} \end{pmatrix}_K . \tag{1.4.2}$$

para $c = 1, 2, \ldots, C$, onde,

$$B = \begin{pmatrix} B_1(t_1) & \cdots & B_K(t_1) \\ \vdots & & \vdots \\ B_1(t_T) & \cdots & B_K(t_T) \end{pmatrix}_{T\times K} .$$

A matriz B não depende do tipo de consumidor, pois as bases utilizadas para estimar todos os C tipos de consumidores são as mesmas. Analisando a Expressão (1.4.2), podemos notar que os T pontos observados no intervalo $[0, 24]$ (um dia) são utilizados apenas para estimar os parâmetros $b_{c,k}$ (que não dependem mais de t), pois a partir das estimativas $\hat{b}_{c,k}$ conseguimos estimar as tipologias $\hat{\alpha}_c(t)$ para qualquer $t \in [0, 24]$ através

da relação,

$$\hat{\alpha}_c(t) = \left(\begin{array}{ccc} B_1(t) & \cdots & B_K(t) \end{array} \right)_{1\times K} \left(\begin{array}{c} \hat{b}_{c,1} \\ \vdots \\ \hat{b}_{c,K} \end{array} \right)_K ,$$

e não somente os pontos $\hat{\alpha}_c(t)$ para $t = t_1, t_2, \ldots, t_T$.

Como podemos ver na Expressão (1.4.1), as tipologias $\alpha_c(t)$ nos pontos $t = t_1, t_2, \ldots, t_T$ foram escritas como uma combinação linerar de bases B-splines. Portanto, com a substituição de (1.4.1) em (1.3.2) chegamos a expressão,

$$Y_{i,j}(t) = \sum_{c=1}^{C} \sum_{k=1}^{K} N_{c,i} b_{c,k} B_k(t) + \boldsymbol{\varepsilon}_{i,j}(t), \tag{1.4.3}$$

que será utilizada na estimação dos parâmetros $b_{c,k}$.

A Expressão (1.4.3) pode ser colocada em forma de um sistema linear usual e, com isso, conseguimos definir a matriz de desenho e os vetores resposta, de coeficientes e de erros. Para cada observação funcional agrupada $Y_{i,j}(t)$, temos a relação,

$$Y_{i,j}(t) = (N_{1,i}B_1(t), \cdots, N_{1,i}B_K(t), \cdots, N_{C,i}B_1(t), \cdots, N_{C,i}B_K(t)) \left(\begin{array}{c} b_{1,1} \\ \vdots \\ b_{1,K} \\ \vdots \\ b_{C,1} \\ \vdots \\ b_{C,K} \end{array} \right) + \boldsymbol{\varepsilon}_{i,j}(t).$$

Expandindo a expressão acima para $t = t_1, t_2, \ldots, t_T$, $j = 1, 2, \ldots, J$ e $i = 1, 2, \ldots, I$, chegamos ao sistema linear contendo todos os dados,

$$\begin{pmatrix} Y_{1,1}(t_1) \\ \vdots \\ Y_{1,1}(t_T) \\ \vdots \\ Y_{1,J}(t_1) \\ \vdots \\ Y_{1,J}(t_T) \\ \vdots \\ Y_{I,1}(t_1) \\ \vdots \\ Y_{I,1}(t_T) \\ \vdots \\ Y_{I,J}(t_1) \\ \vdots \\ Y_{I,J}(t_T) \end{pmatrix} = \begin{pmatrix} N_{1,1}B_1(t_1)\cdots N_{1,1}B_K(t_1)\cdots N_{C,1}B_1(t_1)\cdots N_{C,1}B_K(t_1) \\ \vdots \\ N_{1,1}B_1(t_T)\cdots N_{1,1}B_K(t_T)\cdots N_{C,1}B_1(t_T)\cdots N_{C,1}B_K(t_T) \\ \vdots \\ N_{1,1}B_1(t_1)\cdots N_{1,1}B_K(t_1)\cdots N_{C,1}B_1(t_1)\cdots N_{C,1}B_K(t_1) \\ \vdots \\ N_{1,1}B_1(t_T)\cdots N_{1,1}B_K(t_T)\cdots N_{C,1}B_1(t_T)\cdots N_{C,1}B_K(t_T) \\ \vdots \\ N_{1,I}B_1(t_1)\cdots N_{1,I}B_K(t_1)\cdots N_{C,I}B_1(t_1)\cdots N_{C,I}B_K(t_1) \\ \vdots \\ N_{1,I}B_1(t_T)\cdots N_{1,I}B_K(t_T)\cdots N_{C,I}B_1(t_T)\cdots N_{C,I}B_K(t_T) \\ \vdots \\ N_{1,I}B_1(t_1)\cdots N_{1,I}B_K(t_1)\cdots N_{C,I}B_1(t_1)\cdots N_{C,I}B_K(t_1) \\ \vdots \\ N_{1,I}B_1(t_T)\cdots N_{1,I}B_K(t_T)\cdots N_{C,I}B_1(t_T)\cdots N_{C,I}B_K(t_T) \end{pmatrix} \begin{pmatrix} b_{1,1} \\ \vdots \\ b_{1,K} \\ \vdots \\ b_{C,1} \\ \vdots \\ b_{C,K} \end{pmatrix} + \begin{pmatrix} \boldsymbol{\varepsilon}_{1,1}(1) \\ \vdots \\ \boldsymbol{\varepsilon}_{1,1}(t_T) \\ \vdots \\ \boldsymbol{\varepsilon}_{1,J}(t_1) \\ \vdots \\ \boldsymbol{\varepsilon}_{1,J}(t_T) \\ \vdots \\ \boldsymbol{\varepsilon}_{I,1}(t_1) \\ \vdots \\ \boldsymbol{\varepsilon}_{I,1}(t_T) \\ \vdots \\ \boldsymbol{\varepsilon}_{I,J}(t_1) \\ \vdots \\ \boldsymbol{\varepsilon}_{I,J}(t_T) \end{pmatrix}. \tag{1.4.4}$$

Portanto, a equação usada para estimar as tipologias é dada por,

$$Y_{IJT} = X_{IJT\times CK}\Theta_{CK} + \varepsilon_{IJT},$$

ou simplesmente,

$$Y = X\Theta + \varepsilon. \tag{1.4.5}$$

onde I, J, T, K indicam o número de transformadores, dias (replicações), pontos observados durante o dia e bases B-splines respectivamente.

Conforme observamos, os erros $\boldsymbol{\varepsilon}_{i,j}(t)$ não satisfazem a hipótese de homoscedasticidade, portanto vamos estimar os coeficientes Θ da Expressão (1.4.5) usando mínimos quadrados generalizados.

Mínimos quadrados generalizados. Suponha que conhecemos uma matriz simétrica definida positiva $\boldsymbol{\Sigma}$, tal que a matriz de variância e covariância para o vetor de erros $\boldsymbol{\varepsilon}$ seja dada por $Var(\boldsymbol{\varepsilon}) = \sigma^2\boldsymbol{\Sigma}$, com $\sigma^2 > 0$, mas não necessariamente conhecido. Podemos esperar que nestas circunstâncias o estimador de mínimos quadrados ordinários de Θ,

embora não viciado, não será mais estimador de mínima variância, pois ignora informação.

Usaremos o símbolo $\hat{\Theta}$ para denotar o estimador de mínimos quadrados generalizados de Θ. O estimador $\hat{\Theta}$ é encontrado minimizando a função soma de quadrados ponderada (pela matriz $\mathbf{\Sigma}^{-1}$) dos resíduos,

$$SQR(\Theta) = (Y - \mathbf{X}\Theta)^T \mathbf{\Sigma}^{-1} (Y - \mathbf{X}\Theta).$$

Com esta abordagem, algumas observações tornam-se mais importantes que outras. Em particular, resíduos correspondentes a erros com uma variância maior são menos importantes no cálculo da soma de quadrados de resíduos generalizados. O estimador de mínimos quadrados generalizados é dado por,

$$\hat{\Theta} = (\mathbf{X}^T \hat{\mathbf{\Sigma}}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \hat{\mathbf{\Sigma}}^{-1} Y. \tag{1.4.6}$$

Para evitar a inversão de matrizes, estimaremos os coeficientes através da resolução do sistema linear,

$$\mathbf{X}^T \hat{\mathbf{\Sigma}}^{-1} Y = (\mathbf{X}^T \hat{\mathbf{\Sigma}}^{-1} \mathbf{X}) \hat{\Theta}. \tag{1.4.7}$$

O ponto mais delicado na aplicação de mínimos quadrados generalizados é o cálculo da inversa da matriz estimada de variância e covariância dos erros $\hat{\mathbf{\Sigma}}$, pois pode causar problemas numéricos devido a sua alta dimensão. Para a Equação (1.4.5), a matriz de variância e covariância $\hat{\mathbf{\Sigma}}$ é bloco diagonal devido as hipóteses de independência de $Y_{i,j}(t)$, para $i = 1, 2, \ldots, I$ e $j = 1, 2, \ldots, J$, facilitando bastante os cálculos da matriz inversa. Devido ao fato de considerarmos as observações funcionais agrupadas provenientes do mesmo transformador independentes e identicamente distribuídas, estimaremos a matriz de variância e covariância para cada transformador usando os dias observados (replicações). Com isso, cada observação funcional $Y_{i,j}$ está relacionada a sua respectiva matriz

de variância e covariância Z_i, $i = 1, 2, \ldots, I$,

$$\vec{Y}_{i,j} = \begin{pmatrix} Y_{i,j}(t_1) \\ Y_{i,j}(t_2) \\ \vdots \\ Y_{i,j}(t_T) \end{pmatrix}_T \quad \text{e} \quad Z_i = \begin{pmatrix} Z_i(t_1,t_1) & Z_i(t_1,t_2) & \ldots & Z_1(t_1,t_T) \\ Z_i(t_2,t_1) & Z_i(t_2,t_2) & \ldots & Z_i(t_2,t_T) \\ \vdots & \vdots & \ddots & \vdots \\ Z_i(t_T,t_1) & Z_i(t_T,t_2) & \ldots & Z_i(t_T,t_T) \end{pmatrix}_{T\times T}, \tag{1.4.8}$$

cujos elementos da matriz Z_i são estimados através da expressão,

$$\hat{Z}_i(t,s) = \sum_{j=1}^{J} \frac{(Y_{i,j}(t) - \hat{Y}_i(t))(Y_{i,j}(s) - \hat{Y}_i(s))}{J-1}, \tag{1.4.9}$$

onde t e $s = t_1, t_2, \ldots, t_T$.

Observando a Expressão (1.4.9), podemos perceber que a estimação de matriz Z_i envolve os valores estimados $\hat{Y}_i(t)$, $t = t_1, t_2, \ldots, t_T$, cuja estimação depende da matriz $\hat{Z}_i$. Portanto, utilizaremos um método iterativo de estimação. Para isso, vamos definir o valor inicial $\hat{Z}_i^{(0)}$, $i = 1, 2, \ldots, I$, cujos valores são calculados pela expressão,

$$\hat{Z}_i^{(0)}(t,s) = \sum_{j=1}^{J} \frac{(Y_{i,j}(t) - \bar{Y}_i(t))(Y_{i,j}(s) - \bar{Y}_i(s))}{J-1}.$$

Com isso, estimamos $\hat{Z}_i$ através do algoritmo abaixo:

1. $m \leftarrow 0$
2. Estimar $\hat{Y}_i^{(m+1)}(t)$, $t = t_1, t_2, \ldots, t_T$, usando $\hat{Z}_i^{(m)}$, $i = 1, 2, \ldots, I$;
3. Estimar $\hat{Z}_i^{(m+1)}$ usando $\hat{Y}_i^{(m+1)}(t)$ através da expressão (1.4.9), para $i = 1, 2, \ldots, I$;
4. Se $\hat{Z}_i^{(m)}(t,s) - \hat{Z}_i^{(m+1)}(t,s) < Erro$, para $i = 1, 2, \ldots, I$, t e $s = t_1, t_2, \ldots, t_T$;
   Então $\hat{Z}_i \leftarrow \hat{Z}_i^{(m+1)}$, fim.
   Senão $m \leftarrow m + 1$, volta ao passo 2;

Agora, com a ajuda da equação linear (1.4.4) e da Expressão (1.4.8), conseguimos

definir a matriz de variância e covariância dos erros $\mathbf{\Sigma}$. Para o i-ésimo transformador,

$$\vec{Y}_i = \begin{pmatrix} \vec{Y}_{i,1} \\ \vec{Y}_{i,2} \\ \vdots \\ \vec{Y}_{i,J} \end{pmatrix}_{JT} \qquad \mathbf{\Sigma}_i = \begin{pmatrix} Z_i & 0 & \dots & 0 \\ 0 & Z_i & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & Z_i \end{pmatrix}_{JT \times JT}, \tag{1.4.10}$$

onde os elementos do vetor $\vec{Y}_i$ e da matriz $\mathbf{\Sigma}_i$ são dados pela Expressão (1.4.8). Como podemos ver na Expressão (1.4.10), a suposição de independência entre as observações funcionais agrupadas do mesmo transformador faz com que a matriz $\mathbf{\Sigma}_i$ seja bloco diagonal.

Com isso, chegamos a matriz de variância e covariância dos erros $\mathbf{\Sigma}$,

$$Y = \begin{pmatrix} \vec{Y}_1 \\ \vec{Y}_2 \\ \vdots \\ \vec{Y}_I \end{pmatrix}_{IJT} \qquad \mathbf{\Sigma} = \begin{pmatrix} \mathbf{\Sigma}_1 & 0 & \dots & 0 \\ 0 & \mathbf{\Sigma}_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \mathbf{\Sigma}_I \end{pmatrix}_{IJT \times IJT}. \tag{1.4.11}$$

onde os elementos do vetor Y e da matriz $\mathbf{\Sigma}$ são dados pela Expressão (1.4.10).

O fato da matriz $\mathbf{\Sigma}$ ser bloco diagonal ajuda bastante no cálculo de sua inversa, pois implica apenas na inversão das matrizes Z_i, $i = 1, 2, \dots, I$, que compõem a diagonal da matriz de variância e covariância dos erros $\mathbf{\Sigma}$. As matrizes foram invertidas usando o método de QR, através do comando *solve()* do pacote estatístico *R*.

Portanto, temos agora todos os elementos necessários para a estimação das tipologias no espaço de funções. O vetor de parâmtros Θ será estimado através da resolução do sistema linear dado pela Expressão (1.4.7),

$$\mathbf{X}^T \hat{\mathbf{\Sigma}}^{-1} Y = (\mathbf{X}^T \hat{\mathbf{\Sigma}}^{-1} \mathbf{X}) \hat{\Theta},$$

pelo método de QR, onde $\hat{\mathbf{\Sigma}}$ é dada por expressão similar à (1.4.11) utilizando $\hat{Z}_i$ dado por (1.4.8).

A estimação da superfície de variância e covariância atráves do sistema linear (1.3.6), implicaria na estimação de T^2 parâmetros para representar a superfície para cada tipo de consumidor e, apesar disso, só teríamos estimativas pontuais. Como alternativa a esses inconvenientes, utilizaremos a representação em um espaço de funções, pois com isso diminuimos consideravelmente no número de parâmetros, aumentamos os graus de liberdade do modelo e estimamos superfícies $\hat{\delta}_c(t,s)$ para t e $s \in [0,24]$.

Utilizaremos o espaço de splines bidimensional construído através do produto tensorial de bases splines unidimensionais. Mais detalhes sobre o produto tensorial de splines podem ser encontrados em Schumaker (1981). Trabalharemos com o espaço de splines cúbicos $S_3(x_1,\ldots,x_K)$ conforme definimos na seção anterios e com as bases B-splines, como na estimação das tipologias. Devido ao formato particular da superfície de interesse (simétrica), usaremos a mesma quantidade de bases (K) e o mesmo posicionamento dos nós nas duas dimensões (linhas e colunas). Com isso, podemos usar apenas a $Cov[Y_i(t), Y_i(s)]$ para $(t-s) \geq 0$. Usando a matriz inteira chegaríamos ao mesmo resultado, mas com um custo computacional maior.

A expressão abaixo mostra a representação dos $\delta_c(t,s)$, t e $s = t_1, t_2, \ldots, t_T$, como uma combinação linear do produto tensorial de bases B-splines,

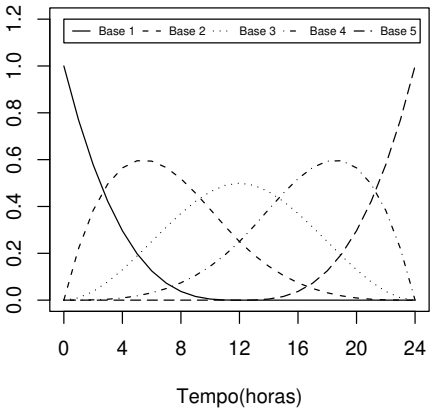
$$\delta_c(t,s) = \sum_{k_1=1}^{K}\sum_{k_2=1}^{K} b_{c,k_1,k_2} B_{k_1}(t) B_{k_2}(s) = \sum_{k_1=1}^{K}\sum_{k_2=1}^{K} b_{c,k_1,k_2} B_{k_1,k_2}(t,s), \tag{1.4.12}$$

para $c = 1, 2, \ldots, C$.

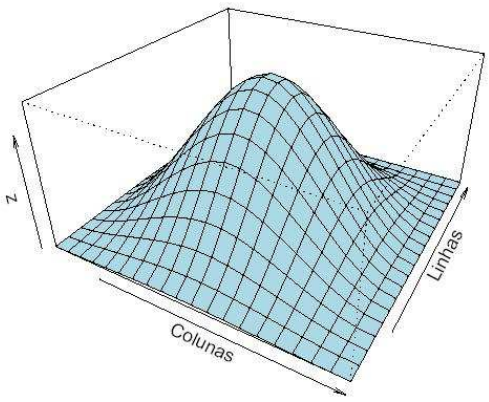
Os dois gráficos mostrados na Figura (1.4.2) foram construídos para facilitar a visualização do produto tensorial das bases B-splines. A Figura (1.4.2-(a)) mostra as cinco bases B-splines obtidadas com um nó interno colocado no centro do intervalo $[0,24]$.

A Expressão (1.4.12) envolve o produto tensorial $B_{k_1}(t)B_{k_2}(s)$, para k_1 e $k_2 = 1, 2, \ldots, K$, onde cada produto produz uma superfície resultante. Para ilustrar, construímos a Figura (1.4.2-(b)), que é a superfície obtida com o produto entre a Base 3 com ela mesma, ou seja, $B_3(t)B_3(s)$.

Colocando a Expressão (1.4.12) em forma matricial, conseguimos ver melhor a relação



(a)



(b)

Figura 1.4.2: Exemplo para ilustrar o produto tensorial entre bases B-splines.

existente entre os $\delta_c(t,s)$, t e $s = t_1, t_2, \ldots, t_T$, e os b_{c,k_1,k_2}, k_1 e $k_2 =, 1, 2, \ldots, K$.

$$\begin{pmatrix} \delta_c(t_1,t_1) \\ \delta_c(t_2,t_1) \\ \delta_c(t_2,t_2) \\ \delta_c(t_3,t_1) \\ \vdots \\ \delta_c(t_T,t_T) \end{pmatrix} = \begin{pmatrix} B_{1,1}(t_1,t_1) & \cdots & B_{1,K}(t_1,t_1) & \cdots & B_{K,1}(t_1,t_1) & \cdots & B_{K,K}(t_1,t_1) \\ B_{1,1}(t_2,t_1) & \cdots & B_{1,K}(t_2,t_1) & \cdots & B_{K,1}(t_2,t_1) & \cdots & B_{K,K}(t_2,t_1) \\ B_{1,1}(t_2,t_2) & \cdots & B_{1,K}(t_2,t_2) & \cdots & B_{K,1}(t_2,t_2) & \cdots & B_{K,K}(t_2,t_2) \\ B_{1,1}(t_3,t_1) & \cdots & B_{1,K}(t_3,t_1) & \cdots & B_{K,1}(t_3,t_1) & \cdots & B_{K,K}(t_3,t_1) \\ \vdots & & \vdots & & \vdots & & \\ B_{1,1}(t_T,t_T) & \cdots & B_{1,K}(t_T,t_T) & \cdots & B_{K,1}(t_T,t_T) & \cdots & B_{K,K}(t_T,t_T) \end{pmatrix} \begin{pmatrix} b_{c,1,1} \\ \vdots \\ b_{c,1,K} \\ \vdots \\ b_{c,K,1} \\ \vdots \\ b_{c,K,K} \end{pmatrix} \quad (1.4.13)$$

para $c = 1, 2, \ldots, C$, onde,

$$\mathbf{B} = \begin{pmatrix} B_{1,1}(t_1,t_1) & \cdots & B_{1,K}(t_1,t_1) & \cdots & B_{K,1}(t_1,t_1) & \cdots & B_{K,K}(t_1,t_1) \\ B_{1,1}(t_2,t_1) & \cdots & B_{1,K}(t_2,t_1) & \cdots & B_{K,1}(t_2,t_1) & \cdots & B_{K,K}(t_2,t_1) \\ B_{1,1}(t_2,t_2) & \cdots & B_{1,K}(t_2,t_2) & \cdots & B_{K,1}(t_2,t_2) & \cdots & B_{K,K}(t_2,t_2) \\ B_{1,1}(t_3,t_1) & \cdots & B_{1,K}(t_3,t_1) & \cdots & B_{K,1}(t_3,t_1) & \cdots & B_{K,K}(t_3,t_1) \\ \vdots & & \vdots & & \vdots & & \\ B_{1,1}(t_T,t_T) & \cdots & B_{1,K}(t_T,t_T) & \cdots & B_{K,1}(t_T,t_T) & \cdots & B_{K,K}(t_T,t_T) \end{pmatrix} \quad (1.4.14)$$

Estamos trabalhando com B-splines cúbicos ($r = 3$). O valor de r está diretamente relacionado ao número de colunas da matriz $\mathbf{B}$, e conseqüentemente com o número de

parâmetros no espaço de funções. Quando $K \leq r+1$, temos a matriz mostrada na expressão (1.4.14) com K^2 colunas. Para $K > r+1$, as colunas B_{k_1,k_2} para $|k_1 - k_2| > r+1$ se anulam, fazendo com que o número de parâmetros diminua.

A matriz $\mathbf{B}$ não depende do tipo de consumidor, pois as bases (produto tensorial) utilizadas para estimar todos os C tipos de consumidores são as mesmas. Analisando a Expressão (1.4.13), podemos notar que os T^2 pontos estimados ($\hat{Z}_i(t,s)$) são utilizados apenas para estimar os parâmetros b_{c,k_1,k_2} (que não dependem mais de t), pois a partir das estimativas $\hat{b}_{c,k_1,k_2}$ conseguimos estimar a superficie de variância e covariância $\hat{\delta}_c(t,s)$, t e $s \in [0,24]$, através da relação,

$$\hat{\delta}_c(t,s) = \left(\begin{array}{ccc} B_{1,1}(t,s) & \cdots & B_{K,K}(t,s) \end{array} \right)_{1\times K^2} \left(\begin{array}{c} \hat{b}_{c,1,1} \\ \vdots \\ \hat{b}_{c,K,K} \end{array} \right)_{K^2\times 1},$$

e não somente alguns pontos isolados (matriz com T^2 pontos).

Como podemos ver na Expressão (1.4.12), as covariâncias $\delta_c(t,s)$, para t e $s = t_1, t_2, \ldots, t_T$, foram escritas como uma combinação linerar do produto tensorial de bases B-splines. Portanto, com a substituição de (1.4.12) em (1.3.6) chegamos a expressão,

$$Z_i(t,s) = \sum_{c=1}^{C} \sum_{k_1=1}^{K} \sum_{k_2=1}^{K} N_{c,i} b_{c,k_1,k_2} B_{k_1,k_2}(t,s) + \xi_i(t,s). \tag{1.4.15}$$

que será utilizada na estimação dos parâmetros b_{c,k_1,k_2}.

A Expressão (1.4.15) pode ser colocada em forma de um sistema linear usual e, com isso, conseguimos definir a matriz de desenho e os vetores resposta, de coeficientes e de erros. Para cada ponto da matriz $\hat{Z}_i(t,s)$, temos a relação,

$$\hat{Z}_i(t,s) = (N_{1,i}B_{1,1}(t,s), \cdots, N_{1,i}B_{K,K}(t,s), \cdots, N_{C,i}B_{1,1}(t,s), \cdots, N_{C,i}B_{K,K}(t,s)) \begin{pmatrix} b_{1,1,1} \\ \vdots \\ b_{1,K,K} \\ \vdots \\ b_{C,1,1} \\ \vdots \\ b_{C,K,K} \end{pmatrix} + \xi_i(t,s).$$

Expandindo a expressão acima para $t = t_1, t_2, \ldots, t_T$, $s \leq t$ e $i = 1, \ldots, I$, chegamos ao sistema linear contendo todos os dados,

$$Z = W\Delta + \xi. \tag{1.4.16}$$

onde,

$$Z = \begin{pmatrix} \hat{Z}_1(t_1,t_1) \\ \hat{Z}_1(t_2,t_1) \\ \hat{Z}_1(t_2,t_2) \\ \hat{Z}_1(t_3,t_1) \\ \vdots \\ \hat{Z}_1(t_T,t_T) \\ \vdots \\ \hat{Z}_I(t_1,t_1) \\ \hat{Z}_I(t_2,t_1) \\ \hat{Z}_I(t_2,t_2) \\ \hat{Z}_I(t_3,t_1) \\ \vdots \\ \hat{Z}_I(t_T,t_T) \end{pmatrix}, \qquad \xi = \begin{pmatrix} \xi_1(t_1,t_1) \\ \xi_1(t_2,t_1) \\ \xi_1(t_2,t_2) \\ \xi_1(t_3,t_1) \\ \vdots \\ \xi_1(t_T,t_T) \\ \vdots \\ \xi_I(t_1,t_1) \\ \xi_I(t_2,t_1) \\ \xi_I(t_2,t_2) \\ \xi_I(t_3,t_1) \\ \vdots \\ \xi_I(t_T,t_T) \end{pmatrix}, \qquad \Delta = \begin{pmatrix} b_{1,1,1} \\ \vdots \\ b_{1,K,K} \\ \vdots \\ b_{C,1,1} \\ \vdots \\ b_{C,K,K} \end{pmatrix},$$

$$W = \begin{pmatrix}
N_{1,1}B_{1,1}(t_1,t_1),\cdots,N_{1,1}B_{K,K}(t_1,t_1),\cdots,N_{C,1}B_{1,1}(t_1,t_1),\cdots,N_{C,1}B_{K,K}(t_1,t_1)) \\
N_{1,1}B_{1,1}(t_2,t_1),\cdots,N_{1,1}B_{K,K}(t_2,t_1),\cdots,N_{C,1}B_{1,1}(t_2,t_1),\cdots,N_{C,1}B_{K,K}(t_2,t_1)) \\
N_{1,1}B_{1,1}(t_2,t_2),\cdots,N_{1,1}B_{K,K}(t_2,t_2),\cdots,N_{C,1}B_{1,1}(t_2,t_2),\cdots,N_{C,1}B_{K,K}(t_2,t_2)) \\
N_{1,1}B_{1,1}(t_3,t_1),\cdots,N_{1,1}B_{K,K}(t_3,t_1),\cdots,N_{C,1}B_{1,1}(t_3,t_1),\cdots,N_{C,1}B_{K,K}(t_3,t_1)) \\
\vdots \\
N_{1,1}B_{1,1}(t_T,t_T),\cdots,N_{1,1}B_{K,K}(t_T,t_T),\cdots,N_{C,1}B_{1,1}(t_T,t_T),\cdots,N_{C,1}B_{K,K}(t_T,t_T)) \\
\vdots \\
N_{1,I}B_{1,1}(t_1,t_1),\cdots,N_{1,I}B_{K,K}(t_1,t_1),\cdots,N_{C,I}B_{1,1}(t_1,t_1),\cdots,N_{C,I}B_{K,K}(t_1,t_1)) \\
N_{1,I}B_{2,1}(t_1,t_1),\cdots,N_{1,I}B_{K,K}(t_2,t_1),\cdots,N_{C,I}B_{1,1}(t_2,t_1),\cdots,N_{C,I}B_{K,K}(t_2,t_1)) \\
N_{1,I}B_{1,1}(t_2,t_2),\cdots,N_{1,I}B_{K,K}(t_2,t_2),\cdots,N_{C,I}B_{1,1}(t_2,t_2),\cdots,N_{C,I}B_{K,K}(t_2,t_2)) \\
N_{1,I}B_{1,1}(t_3,t_1),\cdots,N_{1,I}B_{K,K}(t_3,t_1),\cdots,N_{C,I}B_{1,1}(t_3,t_1),\cdots,N_{C,I}B_{K,K}(t_3,t_1)) \\
\vdots \\
N_{1,I}B_{1,1}(t_T,t_T),\cdots,N_{1,I}B_{K,K}(t_T,t_T),\cdots,N_{C,I}B_{1,1}(t_T,t_T),\cdots,N_{C,I}B_{K,K}(t_T,t_T))
\end{pmatrix}.$$

Devido ao número restrito de observações, usaremos mínimos quadrados ordinários, pois neste caso não é viável aplicarmos o método de mínimos quadrados generalizados como na estimação das tipologias. Apesar disso, conseguimos estimativas não viciadas das superfícies de variância e covariância para cada tipo de consumidor, entretanto não mais de mínima variância, mas aceitáveis dada a complexidade do problema. Estimaremos o vetor de parâmetros através do método de QR a partir do sistema linear (1.4.16).

Portanto, os sistemas lineares usados na estimação do modelo proposto é dado pelas equação,

$$Y = X\Theta + \varepsilon,$$

$$Z = W\Delta + \xi,$$

que estimam as tipologias e as superfícies de variância e covariância respectivamente, para cada tipo de consumidor.

2 *Simulações*

Utilizaremos a terminologia e intuição das curvas de carga de transformadores neste capítulo e aplicaremos a metodologia apresentada no capítulo anterior a conjuntos de dados simulados. Na primeira seção explicaremos em detalhes o processo de geração dos dados. Nas Seções 2.2 e 2.3 mostraremos os resultados para 2 e 3 tipos de consumidores respectivamente e, na Seção 2.4 apresentaremos algumas conclusões.

2.1 Processo de geração dos dados

Nesta seção, descreveremos o procedimento empregado para simular o problema descrito na Seção 1.1. Como foi visto, os consumidores de energia estão divididos em categorias conforme seu ramo de atividade, que por sua vez são caracterizados pela tipologia, representadas por funções no intervalo $[0, 24]$. Os conjuntos de dados serão gerados conforme a Expressão (1.3.1), utilizando as três tipologias hipotéticas apresentadas na Figura 2.1.1. Essas tipologias estão representando três diferentes padrões de consumo, onde as curvas 1 e 2 poderiam representar o consumo residencial para duas classes sociais e a curva 3 o consumo médio de energia de estabelecimentos comerciais.

Vamos considerar o modelo normal p-variado denotado por, $N_p(\alpha, \delta)$, onde α é o vetor de médias com p elementos e δ é a matriz $p \times p$ de variância e covariância de α. Mais detalhes sobre o modelo normal multivariado pode ser encontrado em Johnson e Wichern (1998).

Similar ao caso real, as observações funcionais agrupadas $Y(t)$ simuladas neste capítulo

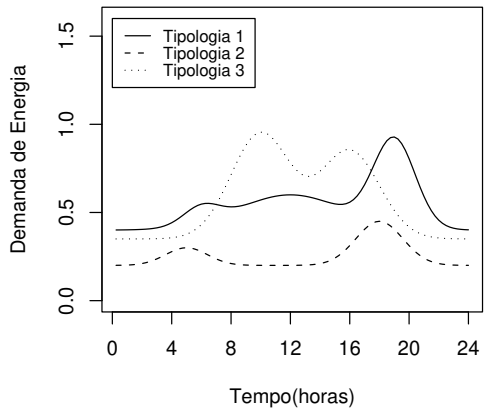


Figura 2.1.1: Tipologias hipotéticas usadas na simulação.

serão representadas por seqüências de 96 pontos eqüidistantes. O Modelo $N_p(\alpha, \delta)$ será usado para simular as observações individuais $y_{c,j,n_c}(t)$, $c = 1, 2, 3$ e $j = 1, 2, \ldots, J$, onde J é o número de replicações (dias). As tipologias hipotéticas mostradas na Figura 2.1.1 e calculadas nos pontos $t = 1, 2, \ldots, 96$ serão usadas como vetores de média α_c, $c = 1, 2, 3$. As matrizes de variância e covariância δ_c, $c = 1, 2, 3$, serão definidas de tal forma que a superfície formada pelos seus pontos seja continua no plano e suave. Para isso, definiremos as variâncias (diagonal da matriz) de tal forma que os pontos formem uma curva suave e, as correlações entre pontos separados por l espaços de tempo $(t, t + l)$, decresçam a medida que l aumenta. Para isso, definimos a variância como sendo trinta porcento da média,

$$\delta_c(t, t) = 0.30\alpha_c(t),$$

para $c = 1, 2, 3$. Para facilitar a notação, usamos,

$$\text{Correlação lag}l = Corr[\alpha_c(t), \alpha_c(t + l)],$$

para $t = 1, 2, \ldots, 96 - l$. Fixamos as correlações lag1 em $0,5$ e as correlações lag2 em $0,2$. As correlações lag$l$, $l = 3, \ldots, 96$, serão definidas como zero. A partir das variâncias