1 Análise Não-Paramétrica de Dados Funcionais: Uma Aplicação à Quimiometria

Autor: Marley Apolinario Saraiva
Orientador: Ronaldo Dias

Introdução

A análise de dados funcionais (ADF) trata o dado individual como uma função e não como um único valor em um ponto particular. Conceitualmente, estas funções são definidas como contínuas e na prática elas são geralmente observadas em pontos discretos. Em geral, os diversos dados funcionais serão independentes uns dos outros, porém não há suposições sobre a independência de diferentes valores dentro do mesmo dado funcional. Os mais comuns são funções do tempo, entretanto nada há de especial em ter o tempo como variável, isto não é uma imposição. Também não existe exigência para que os dados sejam suaves, mas freqüentemente a suavidade ou outra condição de regularidade será um aspecto chave na análise.

Definições

De acordo com Ferraty e Vieu (2006), podemos definir um dado (variável) funcional da seguinte maneira:

Definição 1.1. *Uma variável aleatória X é chamada de variável funcional se toma valores em um espaço dimensional infinito (ou espaço funcional). Uma observação x de X é chamada de dado funcional.*

Definição 1.2. *Um conjunto de dados funcional $x_1, ..., x_n$ é a observação de n variáveis funcionais $X_1, ..., X_n$ identicamente distribuídas.*

Em geral, métodos estatísticos tradicionais não apresentam bons resultados quando lidam com dados funcionais. Se considerarmos uma amostra discretizada de curvas, irão surgir dois problemas cruciais. O primeiro é a razão entre o tamanho da amostra e o número de variáveis (cada variável real corresponde a um ponto discretizado). O segundo é a possível existência de uma forte correlação entre as variáveis. Então surge a necessidade de que se desenvolvam métodos ou modelos que considerem a estrutura funcional deste tipo de dado.

Definição 1.3. *Seja $\mathbf{x}$ um vetor aleatório em $\mathbb{R}^p$ e seja ϕ uma função definida em $\mathbb{R}^p$ e que dependa da distribuição de $\mathbf{x}$. Um modelo para a estimação de ϕ consiste em introduzir algumas condições da forma $\phi \in \mathcal{C}$. O modelo é chamado de modelo paramétrico para a estimação de ϕ se $\mathcal{C}$ é indexado por um número finito de elementos de $\mathbb{R}$. Caso contrário, o modelo é chamado de modelo não paramétrico.*

Esta definição pode ser extendida para o contexto funcional da seguinte forma:

Definição 1.4. *Seja Z uma variável aleatória avaliada em algum espaço de dimensão infinita $\mathcal{F}$ e seja ϕ uma função definida em $\mathcal{F}$ e que dependa da distribuição de Z. Um modelo para a estimação de ϕ consiste em introduzir algumas condições da forma $\phi \in \mathcal{C}$. O modelo é chamado de modelo funcional paramétrico para a estimação de ϕ se $\mathcal{C}$ é indexado por um número finito de elementos de $\mathcal{F}$. Caso contrário, o modelo é chamado de modelo funcional não paramétrico.*

Na terminologia "Estatísticas Não-Paramétricas Funcionais" o adjetivo "não-paramétricas" se refere à forma do conjunto de condições, enquanto que "funcionais" se refere à natureza dos dados. Em outras palavras, aspectos não-paramétricos vêm da característica dimensional infinita do objeto a ser estimado e a designação "funcional" é devida à dimensão infinita dos dados. Esta é a razão de podermos identificar esta estrutura como no contexto de dimensão infinita dupla. De fato, ϕ pode ser visto como um operador linear e poderia-se usar a terminologia "modelo para estimação operacional" por analogia com a terminologia multivariada "modelo para estimação funcional".

Modelo Proposto

A área que se refere à aplicação de métodos estatísticos e matemáticos à problemas de origem química é chamada de quimiometria. Um problema comum é o de como utilizar espectroscopia para se determinar concentrações de vários constituintes em amostras químicas complexas. Considerando que os constituintes não absorvem luz em regiões separadas de freqüência, deve-se utilizar uma combinação de várias freqüências espectrais para se estimar as concentrações. O problema de como combinar a absorção em várias freqüências de forma ótima com o objetivo de aproximar um conjunto de medidas de concentrações é um problema de calibração multivariada (LAQQA, 2009).

Para resolver o problema de calibração multivariada, utilizamos a idéia da lei de Beer-Lambert (definição 1.6) e propomos o tratamento do conjunto de dados coletados utilizando análise não-paramétrica de dados funcionais, uma vez que os mesmos têm características funcionais intrínsecas, isto é, são curvas contínuas (espectros). Esta metodologia não apresenta os problemas teóricos com a dimensão dos dados como os métodos comumente utilizados. Além disso, devido à natureza funcional, acreditamos que modelos que levem em conta esta característica terão melhores resultados do que as técnicas mais utilizadas atualmente, que são de estatística multivariada, como por exemplo PLS (Mínimos Quadrados Parciais) e PCR (Regressão por Componentes Principais).

As duas definições a seguir são importantes para o entendimento do problema.

Definição 1.5. *A absorbância de um analito em um dado comprimento de onda ou freqüência é definida como*

$$x = -\log_{10}\left(\frac{I}{I_0}\right),$$

onde $x > 0$, I é a intensidade da luz transmitida, após a substância ser inserida no feixe de luz, e I_0 é a intensidade de luz incidente, antes da substância ser inserida no feixe de luz. Para espectroscopia de reflexão, $R = I/I_0$ é conhecido como reflectância ($0 < R < 1$).

Definição 1.6. *A lei de Beer-Lambert para m constituintes e k comprimentos de onda mais ruído é*

$$x_j = \sum_{l=1}^{m} y_l a_{lj} + \varepsilon_j, \tag{1.1}$$

para $j = 1, ..., k$, onde x_j é a absorbância da amostra no j-ésimo comprimento de onda, y_l é a concentração do l-ésimo constituinte, a_{lj} é a absorbância do l-ésimo constituinte puro no j-ésimo comprimento de onda e ε_j é o erro aleatório no j-ésimo comprimento de onda.

Mais detalhes sobre a lei de Beer-Lambert podem ser encontrados em Jorgensen e Goegebeur (2007).

Aqui propomos um modelo não-paramétrico funcional para a absorbância semelhante à lei de Beer-Lambert, que relaciona as absorbâncias observadas nas amostras com as concentrações dos constituintes, utilizando funções *spline* obtidas por bases B-*spline* . Os coeficientes das funções base são calculados utilizando o método de mínimos quadrados. Além disso, propomos um modelo para a estrutura de covariância dos dados, também levando em conta suas características funcionais. Observe que este tipo de metodologia difere da usual porque trata o dado de acordo com sua origem funcional e não como um problema multivariado.

Modelo Não-Paramétrico Funcional

Utilizando a idéia da lei de Beer-Lambert (definição 1.6), propomos o seguinte modelo de calibração

$$x_i(t) = \sum_{j=1}^{m} y_{ij}\alpha_j(t) + \epsilon_i(t), \tag{1.2}$$

onde y_{ij} é a concentração do j-ésimo constituinte na i-ésima amostra, t representa o comprimento de onda, $x_i(t)$ é a absorbância da i-ésima amostra no comprimento de onda t e $\epsilon_i(t)$ é o erro aleatório da i-ésima amostra no comprimento de onda t, com $i = 1, ..., n$. A função $\alpha_j(t)$ representa a absorbância do j-ésimo constituinte puro no comprimento de onda t.

As funções α serão aproximadas por funções *spline* cúbicas que serão suavizadas utilizando o critério de penalização da segunda derivada (detalhes podem ser encontrados em Souza (2008)). O coeficiente de suavização foi escolhido pelo critério de validação cruzada generalizada (CRAVEN; WAHBA, 1978/79). As funções *spline* serão calculadas por expansão em bases B-*spline* de ordem 4. Desta forma, 1.2 pode ser escrito como segue.

$$x_i(t) = \sum_{l=i}^{L}\sum_{j=1}^{m} \theta_{jl} y_{ij} B_l(t) + \epsilon_i(t). \tag{1.3}$$

Neste caso, o modelo não apresenta intercepto mas, no entanto, sabemos que não existem elementos químicos com absorbância exatamente igual a zero uma vez que $\frac{I}{I_0} < 1$, como na definição 1.6. Assim, propomos um modelo com intercepto, como mostrado a seguir.

$$x_i(t) = \sum_{l=1}^{L} \left[\theta_{0l} + \sum_{j=1}^{m} \theta_{jl} y_{ij} \right] B_l(t) + \epsilon_i(t) \tag{1.4}$$

Aqui, $B_l(t)$ corresponde à l-ésima base B-*spline* avaliada no ponto t e θ_{jl} é o coeficiente da l-ésima função base do j-ésimo constituinte.

Podemos representar o modelo na forma matricial da seguinte maneira $\boldsymbol{X(t) = D(t)\Theta + \varepsilon(t)}$, onde as matrizes contém os elementos mostrados a seguir.

$$\begin{pmatrix} x_1(t_1) \\ x_1(t_2) \\ \vdots \\ x_1(t_k) \\ x_2(t_1) \\ x_2(t_2) \\ \vdots \\ x_2(t_k) \\ \vdots \\ x_n(t_1) \\ \vdots \\ x_n(t_k) \end{pmatrix} = \begin{pmatrix} B_1(t_1) & y_{11}B_1(t_1) & \cdots & y_{m1}B_1(t_1) & \cdots & B_L(t_1) & y_{11}B_L(t_1) & \cdots & y_{m1}B_L(t_1) \\ B_1(t_2) & y_{11}B_1(t_2) & \cdots & y_{m1}B_1(t_2) & \cdots & B_L(t_2) & y_{11}B_L(t_2) & \cdots & y_{m1}B_L(t_2) \\ \vdots & \vdots & & \vdots & & \vdots & \vdots & & \vdots \\ B_1(t_k) & y_{11}B_1(t_k) & \cdots & y_{m1}B_1(t_k) & \cdots & B_L(t_k) & y_{11}B_L(t_k) & \cdots & y_{m1}B_L(t_k) \\ B_1(t_1) & y_{12}B_1(t_1) & \cdots & y_{m2}B_1(t_1) & \cdots & B_L(t_1) & y_{12}B_L(t_1) & \cdots & y_{m2}B_L(t_1) \\ B_1(t_2) & y_{12}B_1(t_2) & \cdots & y_{m2}B_1(t_2) & \cdots & B_L(t_2) & y_{12}B_L(t_2) & \cdots & y_{m2}B_L(t_2) \\ \vdots & \vdots & & \vdots & & \vdots & \vdots & & \vdots \\ B_1(t_k) & y_{12}B_1(t_k) & \cdots & y_{m2}B_1(t_k) & \cdots & B_L(t_k) & y_{12}B_L(t_k) & \cdots & y_{m2}B_L(t_k) \\ \vdots & \vdots & & \vdots & & \vdots & \vdots & & \vdots \\ B_1(t_1) & y_{1n}B_1(t_1) & \cdots & y_{mn}B_1(t_1) & \cdots & B_L(t_1) & y_{1n}B_L(t_1) & \cdots & y_{mn}B_L(t_1) \\ B_1(t_2) & y_{1n}B_1(t_2) & \cdots & y_{mn}B_1(t_2) & \cdots & B_L(t_2) & y_{1n}B_L(t_2) & \cdots & y_{mn}B_L(t_2) \\ \vdots & \vdots & & \vdots & & \vdots & \vdots & & \vdots \\ B_1(t_k) & y_{1n}B_1(t_k) & \cdots & y_{mn}B_1(t_k) & \cdots & B_L(t_k) & y_{1n}B_L(t_k) & \cdots & y_{mn}B_L(t_k) \end{pmatrix} \begin{pmatrix} \theta_{01} \\ \theta_{11} \\ \vdots \\ \theta_{m1} \\ \theta_{02} \\ \theta_{12} \\ \vdots \\ \theta_{m2} \\ \vdots \\ \theta_{0L} \\ \theta_{1L} \\ \vdots \\ \theta_{mL} \end{pmatrix} + \begin{pmatrix} \epsilon_1(t_1) \\ \epsilon_1(t_2) \\ \vdots \\ \epsilon_1(t_k) \\ \epsilon_2(t_1) \\ \epsilon_2(t_2) \\ \vdots \\ \epsilon_2(t_k) \\ \vdots \\ \epsilon_n(t_1) \\ \vdots \\ \epsilon_n(t_k) \end{pmatrix}$$

A matriz $\boldsymbol{\Theta}$ pode ser estimada utilizando o método de mínimos quadrados, ou seja, resolvendo o seguinte sistema $\hat{\boldsymbol{\Theta}} = (\boldsymbol{D(t)}^T\boldsymbol{D(t)})^{-1}\boldsymbol{D(t)}^T\boldsymbol{X(t)}$. Uma vez estimada a matriz $\boldsymbol{\Theta}$, é possível calcular as estimativas de $\boldsymbol{X(t)}$, da seguinte forma: $\hat{\boldsymbol{X}}\boldsymbol{(t) = D(t)\hat{\Theta}}$. A matriz curva média $\bar{\boldsymbol{X}}\boldsymbol{(t)}$ é obtida da seguinte forma: $\bar{\boldsymbol{X}}\boldsymbol{(t)} = \begin{pmatrix} \bar{x}(t_1) & \bar{x}(t_2) & \cdots & \bar{x}(t_k) \end{pmatrix}$, onde $\overline{x}(t) = \frac{1}{N}\sum_{i=1}^{N} x_i(t)$.

Função Covariância

A função covariância empírica para dados funcionais resume a dependência de registros em todo valor diferente do argumento e é calculado para todo t_1 e t_2 por

$$cov_X(t_1, t_2) = \frac{1}{N-1}\sum_{i=1}^{N}\{x_i(t_1) - \overline{x}(t_1)\}\{x_i(t_2) - \overline{x}(t_2)\}. \tag{1.5}$$

No sentido de obter uma melhor explicação da variabilidade dos dados, propomos a seguinte função de covariância:

$$cov_{\boldsymbol{x}}(t_i, t_j) = \eta(t_i)\eta(t_j)\exp(-\phi|t_j - t_i|), \tag{1.6}$$

onde a função η é contínua para todo t. Analogamente ao modelo 1.4, esta função será aproximada utilizando *splines* cúbicos, onde a suavização será obtida pelo critério de penalização da segunda derivada (detalhes em Souza (2008)) com parâmetro de suavização determinado pelo método de validação cruzada generalizada (CRAVEN; WAHBA, 1978/79). Utilizamos bases B-*spline* para o cálculo das funções *splines* . Desta forma, o modelo 1.6 pode ser escrito como

$$cov_{\boldsymbol{X}}(t_i, t_j) = \sum_{l_1=1}^{L} \sum_{l_2=1}^{L} \beta_{l_1} B_{l_1}(t_i) \beta_{l_2} B_{l_2}(t_j) \exp(-\phi |t_j - t_i|), \tag{1.7}$$

onde $B_l(t)$ representa o valor da l-ésima base B-*spline* avaliada no ponto t. Estas bases são as mesmas que já foram calculadas para o modelo 1.4. O valor de ϕ é fixado e é calculado da seguinte forma (SCHMIDT; CONCEIçãO; MOREIRA, 2008):

$$\phi = \frac{-2\log(0,05)}{\max\limits_{\forall t,s}(|t-s|)}.$$

Na forma matricial temos $\boldsymbol{\Sigma} = \boldsymbol{P\beta}$, onde os elementos das matrizes foram calculados como segue.

$$\begin{pmatrix} cov_{\boldsymbol{X}}(t_1,t_1) \\ cov_{\boldsymbol{X}}(t_2,t_1) \\ \vdots \\ cov_{\boldsymbol{X}}(t_k,t_1) \\ cov_{\boldsymbol{X}}(t_1,t_2) \\ cov_{\boldsymbol{X}}(t_2,t_2) \\ \vdots \\ cov_{\boldsymbol{X}}(t_k,t_2) \\ \vdots \\ cov_{\boldsymbol{X}}(t_1,t_k) \\ cov_{\boldsymbol{X}}(t_2,t_k) \\ \vdots \\ cov_{\boldsymbol{X}}(t_k,t_k) \end{pmatrix} = \begin{pmatrix} B_{11}B_{11}e_{11} & B_{11}B_{21}e_{11} & \cdots & B_{11}B_{L1}e_{11} & \cdots & B_{L1}B_{11}e_{11} & B_{L1}B_{21}e_{11} & \cdots & B_{L1}B_{L1}e_{11} \\ B_{12}B_{11}e_{21} & B_{12}B_{21}e_{21} & \cdots & B_{12}B_{L1}e_{21} & \cdots & B_{L2}B_{11}e_{21} & B_{L2}B_{21}e_{21} & \cdots & B_{L2}B_{L1}e_{21} \\ \vdots & \vdots & & \vdots & & \vdots & \vdots & & \vdots \\ B_{1k}B_{11}e_{k1} & B_{1k}B_{21}e_{k1} & \cdots & B_{1k}B_{L1}e_{k1} & \cdots & B_{Lk}B_{11}e_{k1} & B_{Lk}B_{21}e_{k1} & \cdots & B_{Lk}B_{L1}e_{k1} \\ B_{11}B_{12}e_{12} & B_{11}B_{22}e_{12} & \cdots & B_{11}B_{L2}e_{12} & \cdots & B_{L1}B_{12}e_{12} & B_{L1}B_{22}e_{12} & \cdots & B_{L1}B_{L2}e_{12} \\ B_{12}B_{12}e_{22} & B_{12}B_{22}e_{22} & \cdots & B_{12}B_{L2}e_{22} & \cdots & B_{L2}B_{12}e_{22} & B_{L2}B_{22}e_{22} & \cdots & B_{L2}B_{L2}e_{22} \\ \vdots & \vdots & & \vdots & & \vdots & \vdots & & \vdots \\ B_{1k}B_{12}e_{k2} & B_{1k}B_{22}e_{k2} & \cdots & B_{1k}B_{L2}e_{k2} & \cdots & B_{Lk}B_{12}e_{k2} & B_{Lk}B_{22}e_{k2} & \cdots & B_{Lk}B_{L2}e_{k2} \\ \vdots & \vdots & & \vdots & & \vdots & \vdots & & \vdots \\ B_{11}B_{1k}e_{1k} & B_{11}B_{2k}e_{1k} & \cdots & B_{11}B_{Lk}e_{1k} & \cdots & B_{L1}B_{1k}e_{1k} & B_{L1}B_{2k}e_{1k} & \cdots & B_{L1}B_{Lk}e_{1k} \\ B_{12}B_{1k}e_{2k} & B_{12}B_{2k}e_{2k} & \cdots & B_{12}B_{Lk}e_{2k} & \cdots & B_{L2}B_{1k}e_{2k} & B_{L2}B_{2k}e_{2k} & \cdots & B_{L2}B_{Lk}e_{2k} \\ \vdots & \vdots & & \vdots & & \vdots & \vdots & & \vdots \\ B_{1k}B_{1k}e_{kk} & B_{1k}B_{2k}e_{kk} & \cdots & B_{1k}B_{Lk}e_{kk} & \cdots & B_{Lk}B_{1k}e_{kk} & B_{Lk}B_{2k}e_{kk} & \cdots & B_{Lk}B_{Lk}e_{kk} \end{pmatrix} \begin{pmatrix} \beta_1\beta_1 \\ \beta_1\beta_2 \\ \vdots \\ \beta_1\beta_L \\ \beta_2\beta_1 \\ \beta_2\beta_2 \\ \vdots \\ \beta_2\beta_L \\ \vdots \\ \beta_L\beta_1 \\ \beta_L\beta_2 \\ \vdots \\ \beta_L\beta_L \end{pmatrix}$$

Para simplificar a notação, usamos $B_{ij} = B_i(t_j)$ e $e_{ij} = \exp(-\phi|t_j - t_i|)$.

A partir dos dados observados calculamos a matriz de covariância empírica considerando que a parte aleatória do modelo está nos erros. Desta forma tal matriz é igual à matriz de covariância dos resíduos. Esta matriz pode ser calculada segundo 1.5, e pode ser escrita da seguinte forma

$$\begin{aligned} \boldsymbol{\Sigma}^* &= \boldsymbol{\Sigma} + \boldsymbol{\omega} \\ \boldsymbol{\Sigma}^* &= \boldsymbol{P\beta} + \boldsymbol{\omega}, \end{aligned}$$

onde $\boldsymbol{\omega}$ é um erro aleatório cuja distribuição é normal com média 0 e matriz de variâncias $\sigma^2\boldsymbol{I}$.

Desta maneira, utilizamos os valores da matriz de covariância empírica para estimar os valores de $\boldsymbol{\beta}$ pelo método de mínimos quadrados, isto é, $\hat{\boldsymbol{\beta}} = (\boldsymbol{P^T P})^{-1}\boldsymbol{P^T\Sigma^*}$.

Assim, obtemos as estimativas de $\boldsymbol{\Sigma}$ da seguinte forma: $\hat{\boldsymbol{\Sigma}} = \boldsymbol{P}\hat{\boldsymbol{\beta}}$.

Para mostrar que esta função proposta é uma função de covariância válida, usamos o seguinte argumento:

Seja $W(t)$ um processo estocástico com função de covariância dada por $cov\,(W(s), W(t)) = \exp(-\phi|t-s|)$. Considere $Z(t) = \sum_{l=1}^{L} \beta_l B_l(t) W(t)$. A função de covariância de $Z(t)$ é dada por

$$\begin{aligned} cov\,(Z(s), Z(t)) &= cov\left(\sum_{l_1=1}^{L} \beta_{l_1} B_{l_1}(s) W(s), \sum_{l_2=1}^{L} \beta_{l_2} B_{l_2}(t) W(t)\right) \\ &= \sum_{l_1=1}^{L} \beta_{l_1} B_{l_1}(s) \sum_{l_2=1}^{L} \beta_{l_2} B_{l_2}(t) cov\,(W(s), W(t)) \\ &= \sum_{l_1=1}^{L} \sum_{l_2=1}^{L} \beta_{l_1} B_{l_1}(s) \beta_{l_2} B_{l_2}(t) \exp(-\phi|t-s|). \end{aligned}$$

Portanto, a função de covariância proposta em 1.7 é uma função de covariância válida.

Uma vez que temos um modelo funcional para a estrutura de covariância, podemos estimar também intervalos de confiança para a verdadeira função. Um intervalo de 95% é dado por: $\left[\bar{\boldsymbol{X}}(\boldsymbol{t}) - 2\boldsymbol{V}, \bar{\boldsymbol{X}}(\boldsymbol{t}) + 2\boldsymbol{V}\right]$, onde $\boldsymbol{V} = \sqrt{diag(\boldsymbol{\Sigma})}$.

Aplicações

O objetivo aqui é mostrar a aplicabilidade do modelo proposto. Para isto, é exposto um conjunto de dados funcionais, obtido em Jorgensen e Goegebeur (2007), e um conjunto de dados simulados. Assim, foi possível testar o modelo proposto em diferentes situações. Tal modelo estima uma função para a absorbância e uma função para a covariância e ambas foram suavizadas utilizando o critério de penalização da segunda derivada (detalhes em Souza (2008)) sendo que o parâmetro de suavização foi escolhido utilizando o critério de validação cruzada generalizada (CRAVEN; WAHBA, 1978/79).

Todos os ajustes foram feitos utilizando o software R Development Core Team (2009).

Conjunto de Dados Simulados

Este conjunto de dados simula um ambiente com 10 amostras de 3 analitos, cujos espectros foram observados em 30 comprimentos de onda. A Figura 1.1 mostra o conjunto de dados. Os espectros foram gerados pela função $x(t) = (\sin(2\pi t^3))^3 + \epsilon$, com $t \in [0, 1]$ e ϵ tem distribuição normal com média 0 e desvio-padrão $0,1$.

Utilizando o modelo proposto em 1.4 obtemos as curvas estimadas mostradas na Figura 1.2. A curva sólida em destaque é a estimação da curva média.

Utilizando a estimação da matriz de covariância obtida por 1.5, podemos construir intervalos de confiança para a função média, como mostra o gráfico da Figura 1.3.

Dados de Hidrocarbonetos Poliaromáticos (PAH)

Este conjunto de dados contém espectros EAS (Espectroscopia de Absorção Eletrônica). São 25 amostras químicas, sendo cada uma composta de 10 elementos químicos diferentes. Os espectros foram medidos em 27 comprimentos de onda diferentes, variando de 220nm a 350nm. A Figura 1.4 mostra o conjunto de dados.

Semelhantemente ao que foi feito com os dados simulados, obtemos as curvas estimadas mostradas na Figura 1.5. A curva sólida em destaque é a estimação da curva média, isto é, representa a função média das absorbâncias por comprimento de onda.

Podemos construir intervalos de confiança para a verdadeira função e neste caso, notamos que a amplitude do intervalo varia, mostrando que a variância é diferente em distintos comprimentos de onda, como mostra o gráfico da Figura 1.6.

Conclusões

Neste trabalho foi proposto um método de calibração funcional que não apresenta as principais dificuldades dos modelos de calibração multivariada aplicados à quimiometria. Foi proposto também um modelo para a função de covariância com o intuito de descrever a variabilidade dos erros envolvidos.

Aplicando este modelo a alguns conjuntos de dados simulados foi possível observar que a função estimada em cada caso é muito próxima da verdadeira e também que o intervalo de confiança proposto continha a verdadeira curva.

A aplicação a dados reais se mostrou igualmente satisfatória, uma vez que a função estimada e o intervalo de confiança conseguiram descrever as variações dos dados observados, sendo bastante adaptativos.

Assim, considerando as características funcionais intrínsecas do problema estudado obtivemos excelentes resultados, embora saibamos que ainda há muito a se fazer, como por exemplo, definir algum critério de comparação de modelos e desenvolver métodos preditivos. Além disso, este tipo de abordagem também permite que, caso exista, a informação a priori pode ser incorporada ao modelo. Neste caso, a obtenção da função de verossimilhança, cujos parâmetros são $\boldsymbol{\beta}$, $\boldsymbol{\Theta}$, $\boldsymbol{\phi}$, λ e k (λ é o parâmetro de suavização e k é o número de nós), se faz necessário e para tanto poderemos usar algoritmos baseados em Dias e Gamerman (2002).

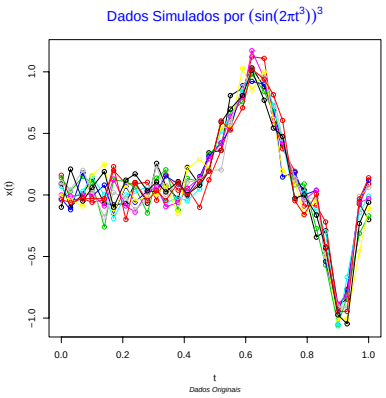

Figura 1.1: Conjunto de dados Simulados

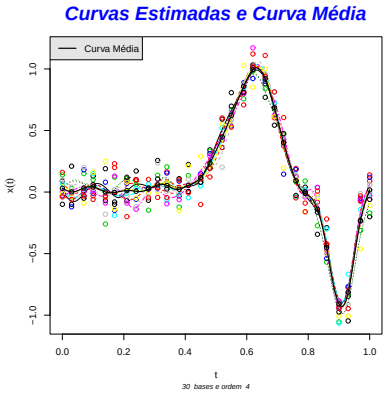

Figura 1.2: Funções estimadas para o conjunto de dados Simulados

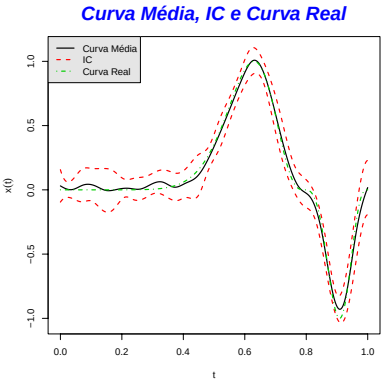

Figura 1.3: Curva média e intervalo de confiança para o conjunto de dados Simulados

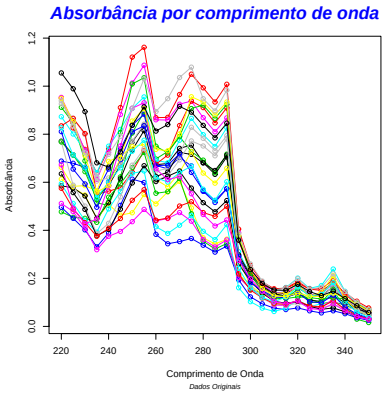

Figura 1.4: Conjunto de dados PAH

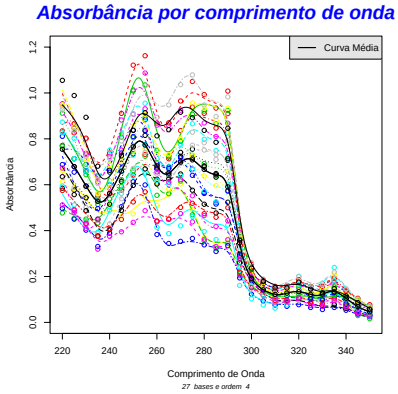

Figura 1.5: Funções estimadas para o conjunto de dados PAH

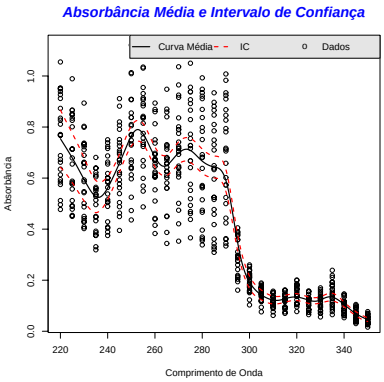

Figura 1.6: Curva média e intervalo de confiança para o conjunto de dados PAH

Referências

CRAVEN, P.; WAHBA, G. Smoothing noisy data with spline functions. estimating the correct degree of smoothing by the method of generalized cross-validation. *Numerische Mathematik*, v. 31, n. 4, p. 377–403, 1978/79.

DIAS, R.; GAMERMAN, D. A bayesian approach to hybrid splines nonparametric regression. *Journal of Statistical Computation and Simulation,*, v. 72, n. 4, p. 285–297, 2002.

FERRATY, F.; VIEU, P. *Nonparametric Functional Data Analysis*. 1. ed. New York: Springer-Velag Inc., 2006. (Springer Series in Statistics).

JORGENSEN, B.; GOEGEBEUR, Y. *Multivariate Data Analysis and Chemometrics*. http://statmaster.sdu.dk/courses/ST02: University of Southern Denmark, Department of Statistics, 2007.

LAQQA, L. D. Q. E. Q. A. 2009. Online: acessado em 11 de março de 2009. Disponível em: <http://laqqa.iqm.unicamp.br/Quimiometria.html>.

MULLEN, K. M.; STOKKUM, I. H. M. v. nnls: The lawson-hanson algorithm for non-negative least squares (nnls). *R package version 1.2*, 2008. Online: acessado em 10 de outubro de 2009. Disponível em: <http://cran.r-project.org/web/packages/nnls/nnls.pdf>.

R Development Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2009. ISBN 3-900051-07-0. Disponível em: <http://www.R-project.org>.

RAMSAY, J.; HOOKER, G.; GRAVES, S. *Functional Data Analysis with R and MATLAB*. 1. ed. New York: Springer New York, 2009. (Use R).

RAMSAY, J.; SILVERMAN, B. W. *Functional Data Analysis*. 1. ed. New York: Springer-Velag Inc., 1997. (Springer Series in Statistics).

RAMSAY, J.; SILVERMAN, B. W. *Applied Functional Data Analysis*. 1. ed. New York: Springer-Velag Inc., 2002. (Springer Series in Statistics).

SCHMIDT, A. M.; CONCEIçãO, M. d. F. a. G.; MOREIRA, G. A. Investigating the sensitivity of gaussian processes to the choice of their correlation function and prior specifications. *Journal of Statistical Computation and Simulation*, v. 78, n. 8, p. 681–699, 2008.

SOUZA, C. P. E. de. *Testes de hipóteses para dados funcionais baseados em distâncias: um estudo usando splines*. Dissertação (Mestrado) — IMECC-UNICAMP, Campinas-SP, 2008.

YAGLOM, A. M. *Correlation Theory of Stacionary and Related Random Funtions 1*. New York: Springer-Velag Inc., 1987.

YAGLOM, A. M. *Correlation Theory of Stacionary and Related Random Funtions 2*. New York: Springer-Velag Inc., 1987.