UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
ESCOLA DE ENGENHARIA
DEPARTAMENTO DE ENGENHARIA QUÍMICA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA QUÍMICA

Desenvolvimento de uma Ferramenta para Obtenção de Modelos Empíricos

DISSERTAÇÃO DE MESTRADO

Tiago Fiorenzano Finkler

PORTO ALEGRE
2003

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
ESCOLA DE ENGENHARIA
DEPARTAMENTO DE ENGENHARIA QUÍMICA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA QUÍMICA

Desenvolvimento de uma Ferramenta para Obtenção de Modelos Empíricos

Tiago Fiorenzano Finkler

Dissertação de Mestrado apresentada como requisito parcial para obtenção do título de Mestre em Engenharia

Área de concentração:

Pesquisa e Desenvolvimento de Processos

Orientador:
Prof. Dr. Nilo Sérgio Medeiros Cardozo

PORTO ALEGRE
2003

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

ESCOLA DE ENGENHARIA

DEPARTAMENTO DE ENGENHARIA QUÍMICA

PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA QUÍMICA

A Comissão Examinadora, abaixo assinada, aprova a Dissertação *Desenvolvimento de uma Ferramenta para Obtenção de Modelos Empíricos*, elaborada por Tiago Fiorenzano Finkler, como requisito parcial para obtenção do Grau de Mestre em Engenharia.

Comissão Examinadora:

Prof. Dr. Argimiro Resende Secchi

Prof. Dr. Jorge Otávio Trierweiler

Profa. Dra. Liliane Basso Barichello

"*O meu patriotismo não é exclusivo. Engloba tudo. Eu repudiaria o patriotismo que procurasse apoio na miséria ou na exploração de outras nações. O patriotismo que eu concebo não vale nada se não se conciliar sempre, sem exceções, com o maior bem e a paz de toda a humanidade."*

"Não há caminhos para a paz; a paz é o caminho."

Mahatma Gandhi

Agradecimentos

Em primeiro lugar, sempre, ao professor Ricardo e à professora Heloise, os dois grandes professores da minha vida, nunca esquecendo da minha querida irmã Tati. Muito obrigado pelo apoio, conforto e carinho que sempre me proporcionaram e por serem o sentido de todo o meu esforço.

Às grandes amizades que fiz no decorrer dos cursos de graduação e mestrado em engenharia, só resta oferecer a minha eterna gratidão pelo companheirismo que fez com que muitos dos momentos que vivi nos últimos sete anos fiquem registrados para sempre entre os melhores de minha vida.

Agradeço, também, a toda a minha família, a todos os meus amigos e ao pessoal da Braskem que me acompanhou durante o desenvolvimento deste projeto, especialmente ao engenheiro Gustavo Neumann. A todos os professores e funcionários da UFRGS, exímios profissionais aos quais devo a minha formação acadêmica, em particular ao professor Nilo, pela grande contribuição prestada na orientação deste trabalho.

Resumo

O objetivo deste trabalho é o desenvolvimento de uma ferramenta de regressão multivariável apropriada para abordar o problema da modelagem de propriedades relacionadas à qualidade final de produtos em processos industriais. No caso geral, dados provenientes de sistemas físicos e químicos típicos do meio industrial são caracterizados pela presença de relação não linear entre as variáveis, podendo, ainda, apresentar outros problemas que dificultam o processo de modelagem empírica, como baixa disponibilidade de observações experimentais, presença de ruído excessivo nas medidas e a presença de colinearidade entre os termos explicativos. Foi feita uma revisão de diferentes métodos de regressão multivariável tais como regressão linear múltipla (*MLR*), regressão não linear baseada em transformações das variáveis explicativas e métodos de redução de dimensionalidade (*PCA*, *PLS*, *QPLS* e *BTPLS*). Também foram propostas novas metodologias para a abordagem das questões da seleção de variáveis e estimação das incertezas dos modelos. Posteriormente, utilizando as metodologias revisadas e propostas, foi sugerida uma sistemática para o tratamento da questão da modelagem empírica de dados industriais, que constitui a base para a implementação da ferramenta desejada. A aplicabilidade da ferramenta desenvolvida foi ilustrada através de alguns estudos de caso retirados da literatura, onde modelos para a predição de propriedades relativas à qualidade de produtos produzidos em quatro tipos de processos industriais diferentes são obtidos.

Palavras chave: modelos empíricos, regressão multivariável, *PCA*, *PLS*, *QPLS*, *BTPLS*, seleção de variáveis, regressão stepwise, estimação de incertezas.

Abstract

The aim of this work is the development of a multivariate regression tool to approach the question of modeling properties related to the product quality in industrial processes. Modeling of industrial physical and chemical systems is usually characterized by the following difficulties: nonlinear relations between the variables, low availability of experimental observations, presence of excessive noise in the measurements and collinearity between the explanatory terms. A number of multivariate regression techniques used to take into account these difficulties were revised. These techniques include multiple linear regression (MLR), nonlinear regression based on explanatory variables transformation and dimension reduction methods (PCA, PLS, QPLS and BTPLS). New methodologies to carry out variable selection and uncertainty estimation in empirical modeling problems were also proposed. Using the revised and proposed methodologies, it was suggested a systematic approach to the process plant data empirical modeling problem. This approach constitutes the basis for the implementation of the desired tool. The applicability of the developed tool was illustrated through the analysis of some case studies from literature.

Key-word: empirical models, multivariate regression, PCA, PLS, QPLS, BTPLS, variable selection, stepwise regression, uncertainty estimation.

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1 Introdução

1.1. Motivação

No meio industrial, os processos são caracterizados por uma diversidade de variáveis que devem ser especificadas de modo a determinar as propriedades finais dos diferentes produtos. Infelizmente, as relações existentes entre as variáveis de processo e as propriedades do produto final nem sempre são exatamente conhecidas. A dificuldade encontrada na determinação destas relações é conseqüência da complexidade inerente a este contexto multivariável, onde a influência exercida por uma variável em determinada propriedade final é afetada pela especificação das demais. Na prática, geralmente, são conhecidas as "receitas", ou seja, as especificações de processo que conduzem aos diferentes produtos. O desenvolvimento de novos produtos (obtenção de novas "receitas") é realizado através de testes, onde as especificações do processo são alteradas até se atingir as propriedades finais desejadas. Normalmente, estes experimentos são orientados por profissionais que conhecem profundamente o processo e o produto, ou seja, uma equipe formada por engenheiros de processo e por engenheiros da área comercial. Entretanto, se estes profissionais forem questionados a respeito da relação existente entre as variáveis de processo e as propriedades do produto final, provavelmente serão verificadas algumas opiniões conflitantes. Isso é esperado, uma vez que estas opiniões, baseadas somente na soma de experiências vivenciadas por cada profissional, tendem a supervalorizar ocorrências isoladas que, eventualmente, podem não ter sido interpretadas de maneira correta. Esta descentralização do conhecimento pode, portanto, causar divergências que conduzam a experimentos mal sucedidos. Um teste mal sucedido, além de acarretar prejuízos financeiros, aumenta o tempo necessário para o desenvolvimento de um novo produto, fazendo crescer a fila de experimentos necessários para a obtenção da "receita" desejada.

Uma maneira eficiente de organizar e armazenar o conhecimento técnico de uma empresa é o desenvolvimento de modelos matemáticos, capazes de predizer as relações existentes entre as variáveis de um processo e as propriedades finais do produto. Bons modelos, além de serem úteis para muitas aplicações importantes como o controle avançado e a otimização do processo, são uma ferramenta poderosa para profissionais que orientam a

execução de testes para o desenvolvimento de novos produtos, permitindo que muitas dúvidas sejam esclarecidas ainda na fase de planejamento.

1.2. Objetivo

O objetivo do presente trabalho é o desenvolvimento de uma ferramenta para a obtenção de modelos empíricos que relacionem propriedades relativas à qualidade do produto com variáveis especificadas em processos de produção industrial. A ferramenta a ser desenvolvida também deverá ser capaz de identificar as variáveis do processo que são importantes para descrever o comportamento de uma dada propriedade final do produto e de fornecer estimativas para a precisão das predições dos modelos. Tanto a identificação como o desenvolvimento de metodologias apropriadas para a implementação da ferramenta desejada são focos de atenção deste trabalho.

1.3. Estrutura da Dissertação

Para atingir o objetivo deste trabalho, foram revisadas técnicas de regressão multivariável e propostas metodologias alternativas para o tratamento das questões da seleção de variáveis e da estimação das incertezas dos modelos. Posteriormente, com base nos métodos revisados e propostos, foi sugerida uma sistemática para a abordagem do problema de modelagem empírica, a qual constitui a base para a implementação da ferramenta desenvolvida. O conteúdo desta dissertação está disperso em mais cinco capítulos, além deste introdutório. Os tópicos abordados em cada um destes são descritos nos parágrafos a seguir.

O segundo capítulo apresenta alguns conceitos fundamentais necessários para o entendimento do trabalho e uma revisão bibliográfica a respeito de métodos de regressão multivariável. Primeiramente, são introduzidos os conceitos de modelo empírico e de modelo fenomenológico e, em seguida, é tratada a questão da estimação de parâmetros. Posteriormente, a questão da construção de modelos por regressão multivariável por mínimos quadrados é revisada, desde a construção do modelo até a análise estatística dos mesmos. No final do capítulo, são introduzidos os métodos de modelagem por redução de dimensionalidade lineares e suas extensões não lineares.

No terceiro capítulo, é tratada a questão da seleção de variáveis na construção de modelos empíricos. Após uma breve discussão a respeito das técnicas encontradas na literatura, na primeira seção, é apresentado o método *SROV* (*Stepwise Regression based on Orthogonalized Variables*), desenvolvido por Shacham e Brauner (2001). Na segunda seção, uma nova metodologia, o método *SRMP* (*Stepwise Regression based on Model Predictions*), é proposta. Na última seção do capítulo, o método proposto é ilustrado e comparado com o método *SROV* através de um estudo realizado com dados gerados artificialmente.

O quarto capítulo trata da questão da obtenção de aproximações para as incertezas associadas aos modelos através de técnicas de reamostragem. São revisados diferentes métodos de reamostragem baseados nas observações experimentais ou nos resíduos do modelo. Também é proposta uma nova metodologia de reamostragem, baseada no erro

experimental associado à medição das variáveis de entrada e saída. Posteriormente, são conduzidas duas simulações computacionais para a comparação dos métodos.

No quinto capítulo, as metodologias revisadas e propostas são organizadas sistematicamente, formando uma base para a implementação da ferramenta desejada. Ainda no quinto capítulo, a aplicabilidade da ferramenta desenvolvida é ilustrada através de alguns estudos de caso retirados da literatura e de um estudo com dados gerados artificialmente.

Por fim, no sexto e último capítulo, as principais conclusões são ressaltadas e algumas sugestões para a continuidade do mesmo são apresentadas.

Capítulo 2 Conceitos Fundamentais e Revisão Bibliográfica

2.1. Modelos Empíricos e Modelos Fenomenológicos

A obtenção de modelos que descrevam matematicamente a relação entre as variáveis de sistemas é uma questão de fundamental importância na indústria química, principalmente para áreas como controle, simulação e otimização de processos. Um modelo é uma representação aproximada para um sistema real e o processo de modelagem é um balanço entre precisão e simplicidade. Desejamos um modelo que forneça predições suficientemente precisas e que, ao mesmo tempo, seja o mais simples possível, de modo a minimizar o esforço computacional necessário para a obtenção da solução. Basicamente, podemos classificar os modelos matemáticos em dois grandes grupos: os modelos empíricos e os modelos fenomenológicos.

Nesta seção, são caracterizados estes dois tipos de modelos e, posteriormente, é apresentado o critério de classificação dos modelos em relação a forma pela qual os parâmetros aparecem em suas expressões. Nas seções seguintes, as questões da estimação de parâmetros e da análise estatística da regressão são abordadas. Na última seção, são apresentados os métodos de redução de dimensionalidade, uma importante alternativa para casos onde as variáveis explicativas estão mutuamente relacionadas.

2.1.1. Modelos Empíricos

Um modelo empírico é construído a partir de uma análise estatística de observações experimentais, utilizando-se técnicas de regressão. Estes modelos são utilizados em situações onde não há base teórica para alguma fundamentação a respeito da relação existente entre as variáveis do sistema. Nestes casos, deixa-se que os dados experimentais ditem a forma do modelo. Como exemplo de modelos empíricos, podemos citar a Equação de Antoine,

(Equação 2.1), que expressa a relação entre a pressão de vapor (P_v) de um líquido em função da temperatura (T).

$$\log_{10}(Pv) = A - \frac{B}{T + C} \tag{2.1}$$

Outro exemplo típico de modelo empírico que pode ser citado são as correlações que visam fornecer predições de propriedades finais de resinas poliméricas. A Eq. 2.2, por exemplo, estabelece uma relação empírica entre o índice de fluidez (*IF*) da resina e o seu peso molecular médio em massa ($\overline{M}_w$).

$$IF = a.\overline{M}_w{}^b \tag{2.2}$$

Podemos notar que, tanto na Eq. 2.1 como na Eq. 2.2, encontramos constantes, chamadas de parâmetros do modelo, que caracterizam o sistema que está sendo estudado. Como os parâmetros A, B e C presentes na Eq. 2.1 e os parâmetros a e b presentes na Eq. 2.2 são desconhecidos, eles devem ser estimados a partir de observações experimentais. Existem diferentes técnicas de regressão que podem ser utilizadas para estimar os parâmetros de um modelo. Entre elas estão a regressão linear, a regressão linear através de transformação de variáveis, a regressão não linear e a regressão a partir de estruturas latentes, por exemplo. Cada um destes métodos apresenta vantagens e desvantagens e, portanto, ao construirmos um modelo, devemos estar aptos a optar pela ferramenta mais apropriada para a situação em questão. Por hora, não vamos nos preocupar com os métodos de estimação de parâmetros, os quais serão estudados nas próximas seções desta revisão. Vamos apenas, utilizando dados de literatura, ilustrar simplificadamente as idéias básicas por trás dos modelos anteriormente apresentados, visando exemplificar os passos envolvidos na obtenção de modelos empíricos.

Observações experimentais da pressão de vapor do naftaleno em diferentes temperaturas podem ser encontradas em Stephenson et al. (1987). A modelagem da relação existente entre P_v e T através da Eq. de Antoine, consiste na obtenção das estimativas para as constantes A, B e C que melhor representem os dados experimentais disponíveis. O gráfico da Figura 2.1 apresenta o ajuste do modelo obtido com as estimativas para os parâmetros mostradas na legenda. Na Figura 2.2, apresentamos o modelo para a predição do *IF* de resinas de polipropileno. Os dados utilizados para a estimação das constantes a e b foram retirados de Latado et al. (2001).

Em ambos os casos estudados anteriormente, a forma funcional utilizada para descrever a dependência existente entre as variáveis foi sugerida a partir das próprias observações empíricas. Por este motivo, segundo a classificação proposta no início deste tópico, estes modelos são classificados como modelos empíricos.

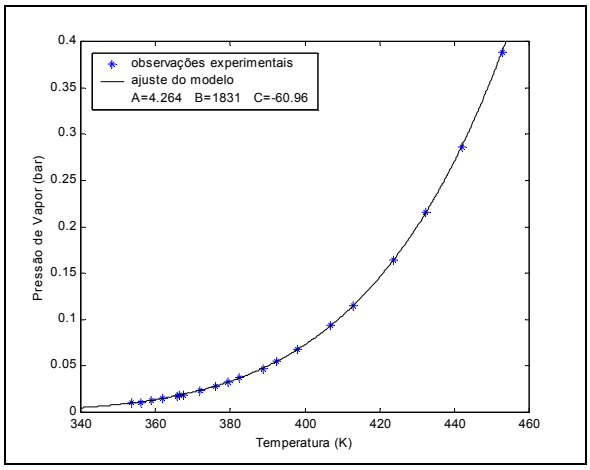

Figura 2.1: Modelo para a P_v.

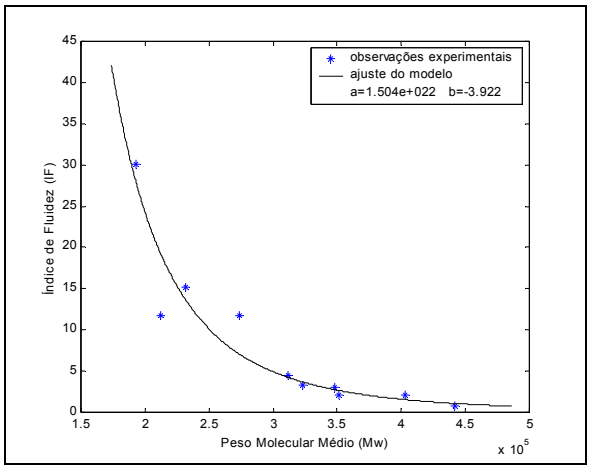

Figura 2.2: Modelo para *IF*.

2.1.2. Modelos Fenomenológicos e Semi-Empíricos

Diferentemente da modelagem empírica, ou puramente empírica, a modelagem fenomenológica é baseada no conhecimento dos processos físicos e químicos que estão por trás do sistema em estudo, tais como os princípios da termodinâmica e as leis da conservação da massa, da energia e da quantidade de movimento. Embora sejam baseados em fundamentações teóricas, muitos modelos fenomenológicos também recorrem a observações experimentais para a obtenção de parâmetros. Neste caso, os modelos são chamados de semi-empíricos. Um exemplo típico são os modelos de reatores químicos onde recorremos à análise de dados experimentais para determinarmos a dependência da taxa de reação com a concentração dos reagentes ou ainda para estimar os valores da energia de ativação e do fator de freqüência, presentes na equação de Arrhenius.

Para ilustrarmos este conceito, vamos considerar um reator batelada ideal isotérmico, onde ocorre a reação de primeira ordem, em fase líquida $A \rightarrow produtos$. Um balanço de massa para o sistema reacional conduz ao modelo:

$$\left(-r_A\right) = -\frac{dC_A}{dt} \text{ ou } k.C_A = \frac{dC_A}{dt} \tag{2.3}$$

onde $(-r_A)$ é a taxa de desaparecimento do reagente em mol/s, k é a velocidade específica de reação em m^3/s e C_A é a concentração de reagente em mol/L.

Nosso objetivo é a obtenção de um modelo que, a partir de uma dada concentração inicial C_{A0}, forneça a concentração de reagente no reator em função do tempo. Para isso, integramos a Eq. 2.3, obtendo:

$$C_A = C_{A_0} e^{-kt} \tag{2.4}$$

Até este ponto, temos um modelo que é fenomenológico, pois é baseado somente em conceitos de cinética química. Caso, para reação em questão, houvesse alguma teoria que

permitisse calcular o valor da constante cinética apenas em função das características das moléculas envolvidas, o modelo seria puramente fenomenológico.

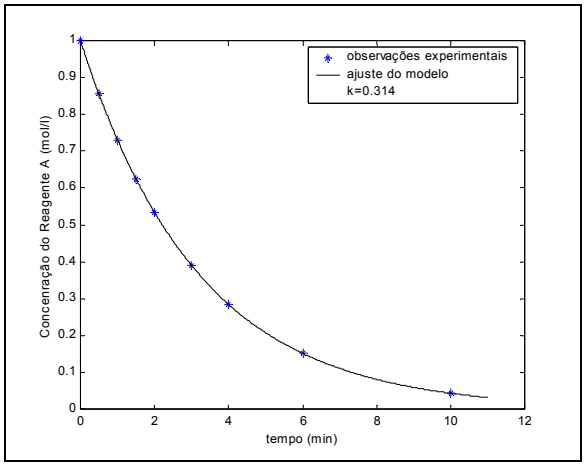

Figura 2.3: Modelo para a reação em batelada.

No entanto, como a velocidade específica da reação não é conhecida, ela deve ser estimada medindo-se a concentração do reagente em diferentes instantes da reação em uma batelada experimental. Assim, podemos utilizar estas observações para plotar os valores de C_A em função do tempo e a constante k pode então ser estimada determinando-se o valor que melhor ajusta as predições da Eq. 2.4 às observações experimentais. Fogler (1992) apresenta um exemplo onde a constante cinética para a reação de formação de etileno glicol a partir de óxido de etileno é determinada. As observações experimentais da concentração de óxido de etileno em função do tempo, assim como o valor de k estimado são mostrados na Figura 2.3.

Neste caso, como a constante cinética foi determinada empiricamente, o modelo apresentado pela Eq 2.4 passa a ser um modelo semi-empírico.

2.1.3. Classificação de modelos com relação a linearidade

A classificação do modelo matemático depende de como os parâmetros aparecem nas equações. A forma geral para modelos lineares pode ser escrita como:

$$y(x_1, x_2, ..., x_n) = \beta_1 f_1(x_1, x_2, ..., x_n) + \beta_2 f_2(x_1, x_2, ..., x_n) + ... + \beta_p f_p(x_1, x_2, ..., x_n)$$

onde $f_i(x_1, x_2, ..., x_n)$ são formas funcionais conhecidas, β_i são os parâmetros do modelo, y é o vetor das variáveis dependentes e x_i são as variáveis explicativas. Por exemplo:

$y(x) = \beta_1 + \beta_2 x$ linear em β e x

$y(x) = \beta_1 + \beta_2 x + \beta_3 x^2$ linear em β e não linear em x

A forma geral para modelos não lineares pode ser escrita como:

$$y(x_1, x_2, ..., x_n) = f(x_1, x_2, ..., x_n; \beta_1, \beta_2, ..., \beta_p)$$

Por exemplo:

$y(x_1, x_2) = \beta_1 + \beta_2 \beta_3 x_1 + \beta_3 x_2$ não linear em β e linear em x

$y(x) = \beta_1 \exp(\beta_2 x)$ não linear em β e x

2.2. Estimação de Parâmetros

Nesta seção, é abordada a questão da estimação de parâmetros em processos de modelagem. Após uma breve introdução, é feita uma revisão a respeito do método da máxima verossimilhança, uma técnica de estimação de parâmetros fundamentada na teoria probabilística que permite que observações empíricas sejam utilizadas para inferir a respeito da relação existente entre diferentes variáveis aleatórias. Posteriormente, é demonstrado que, sob a validade das devidas hipóteses, as estimativas de máxima verossimilhança coincidem com as estimativas do popular método dos mínimos quadrados, que é largamente utilizado para a obtenção de modelos empíricos. Na próxima seção, é feita uma revisão completa da análise estatística de um problema de regressão multivariável utilizando o método dos mínimos quadrados. Será detalhada a dedução da expressão para a obtenção das estimativas dos parâmetros para o caso onde as diferentes variáveis relacionam-se linearmente com a variável de resposta. Nas seções seguintes, serão apresentadas técnicas de transformação de variáveis, que são úteis para a estimação de parâmetros em situações onde a relação entre as diferentes variáveis do sistema é não linear e técnicas de redução de dimensionalidade, que permitem que estimativas para os parâmetros sejam obtidas em situações onde as entradas do modelo estão mutuamente correlacionadas.

Como pode ser verificado, nos exemplos anteriormente discutidos, a estimação de parâmetros é fundamental no processo de construção de modelos empíricos e semi-empíricos. Para melhor explicar esta afirmação, vamos imaginar que, para um polímero fictício, a relação apresentada na Eq. 2.2 fosse capaz de explicar completamente o comportamento do índice de fluidez e que, ainda, determinações experimentais para o *IF* e o M_w pudessem ser obtidas livres de erro experimental. Neste caso, "determinações exatas" para os parâmetros *a* e *b* poderiam ser calculadas a partir de dois pontos experimentais. Isso poderia ser feito pela simples substituição dos valores de *IF* e M_w de cada uma destas duas observações na Eq. 2.2, o que forneceria um sistema de duas equações, cujas duas únicas incógnitas seriam os parâmetros *a* e *b*.

Entretanto, na prática, a "determinação" destes parâmetros resolvendo o sistema de 2 equações não é adequada. A Tabela 2.1 apresenta quatro "determinações" para os parâmetros *a* e *b* do modelo de *IF* utilizando, para cada caso, um par de resinas distinto, extraído dos dados experimentais de Latado et al. (2001). Como pode ser observado, os resultados das determinações diferem drasticamente.

Tabela 2.1: Variabilidade nas determinações dos parâmetros no modelo de *IF*.

	a	*b*
1	6,22E+24	-4,41
2	2,29E+15	-2,70
3	3,75E+20	-3,31
4	5,91E+20	-3,70

Um dos principais fatores que contribuem para a variabilidade das determinações é o fato de que a medição das variáveis está sempre sujeita a incertezas. Ao realizarmos a "determinação" dos parâmetros resolvendo o sistema de equações mencionado anteriormente, as incertezas experimentais se propagam aos resultados. Especialmente no caso das medidas de *IF* e M_w, há uma grande quantidade de ruído nos dados. O ruído diminui a quantidade de informação útil contida nas observações experimentais reduzindo assim a confiabilidade das mesmas. Quanto maior a quantidade de ruído presente nos dados, mais observações experimentais são necessárias para que se obtenha um modelo com determinada precisão.

É interessante verificar que se o mesmo procedimento for realizado com a Eq. de Antoine, utilizando-se os dados de Stephenson et al. (1987), as "determinações" dos parâmetros através da resolução do sistema, agora com 3 equações e 3 incógnitas, apresentam excelente repetibilidade, conforme pode ser observado na Tabela 2.2.

Tabela 2.2: Variabilidade nas determinações dos parâmetros na Eq. de Antoine.

	A	*B*	*C*
1	4,274	1832	61,58
2	4,280	1832	62,04
3	4,271	1832	61,44
4	4,284	1832	62,23

Como a precisão das medidas de P_v e T é alta, levando-se em consideração as faixas de valores estudadas, a quantidade de informação contida nas medidas individuais é grande, de modo que as mesmas são altamente confiáveis. Sendo assim, podemos obter modelos precisos a partir de um número relativamente pequeno de amostras.

O processo de modelagem consiste, basicamente, na extração da informação contida nas observações experimentais, sintetizando-a na forma de parâmetros. Estatisticamente falando, as estimativas para os parâmetros de um modelo são aquelas que maximizam a probabilidade de ocorrência das observações experimentais disponíveis. Formuladas as devidas hipóteses, podemos, nos fazendo valer de observações experimentais, obter tais estimativas através do método da máxima verossimilhança, que é apresentado a seguir.

2.2.1. Método da Máxima Verossimilhança

Para formular o problema de máxima verossimilhança associado a estimação dos parâmetros de um modelo matemático que quantifique a relação existente entre as entradas e saídas de um sistema qualquer vamos, primeiramente, assumir que a saída está relacionada com as entradas de uma maneira tal que o valor verdadeiro da variável de resposta, denotado pela variável *v*, depende somente dos valores verdadeiros das *k* variáveis explicativas

consideradas, denotados pelas variáveis $u_1, u_2, ..., u_k$. A relação existente entre as entradas e a saída é dada pela função genérica *f*:

$$v = f(u_1, u_2, ..., u_k; \beta_1, \beta_2, ..., \beta_L) \tag{2.5}$$

Os parâmetros do modelo $\beta_1, \beta_2, ..., \beta_L$ são desconhecidos e devem ser estimados a partir de observações experimentais do sistema disponíveis. No caso geral, dispomos de *n* observações experimentais. Para cada uma delas, podemos realizar um experimento que nos fornece a medida *y* para a variável de resposta ou de saída e outros *k* experimentos que nos fornecem as medidas $x_1, x_2, ..., x_k$ para as variáveis explicativas ou de entrada. Cada um dos experimentos pode ser repetido um número *r* de vezes, que pode ser diferente para cada uma das variáveis, dependendo da confiabilidade das respectivas medidas.

Do ponto de vista estatístico, as melhores estimativas para os parâmetros do modelo são aquelas que maximizam a probabilidade de ocorrência das observações experimentais disponíveis. Para isso, precisamos construir a função de verossimilhança, que coloca a probabilidade *P* de encontrarmos as observações experimentais disponíveis em função dos parâmetros a serem estimados. Chamando de *p* a probabilidade de ocorrência das medidas individuais, a função de verossimilhança genérica é dada por:

$$P = \prod_{i=1}^{n} \prod_{l=1}^{r} \left(p_{y\,il}\, p_{x1\,il}\, p_{x2\,il} \cdots p_{xk\,il} \right)$$

$$\ln(P) = \sum_{i=1}^{n} \sum_{l=1}^{r} \ln(p_{y\,il}\, p_{x1\,il}\, p_{x2\,il} \cdots p_{xk\,il}) \tag{2.6}$$

Para que a função de verossimilhança possa ser avaliada, é necessário que alguma hipótese a respeito da distribuição seguida pelas medidas experimentais seja assumida. Sendo assim, assumindo que, para cada uma das $i = 1, 2, ..., n$ observações, as $l = 1, 2, ..., r$ réplicas dos experimentos $x_{i1l}, x_{i2l}, ..., x_{ikl}$ e y_{il} seguem uma distribuição normal em torno de u_i e v_i, com variâncias respectivamente iguais a $\sigma_{x1}, \sigma_{x2}, ..., \sigma_{xk}$ e σ_y, a função de verossimilhança torna-se:

$$P = \prod_{i=1}^{n} \prod_{j=1}^{r} \left[\left(\frac{1}{\sqrt{2\pi}\sigma_y} e^{-\frac{(y_{ij} - v_i)^2}{2\sigma_y^2}} \right) \left(\frac{1}{\sqrt{2\pi}\sigma_{x1}} e^{-\frac{(x1_{ij} - u1_i)^2}{2\sigma_{x1}^2}} \right) \cdots \left(\frac{1}{\sqrt{2\pi}\sigma_{k2}} e^{-\frac{(xk_{ij} - uk_i)^2}{2\sigma_k^2}} \right) \right] \tag{2.7}$$

Os parâmetros do modelo são inseridos na função de verossimilhança substituindo-se, para cada uma das *n* observações, o termo v_i, correspondente ao valor verdadeiro da variável de resposta, pela expressão do modelo (Eq. 2.5). Antes de prosseguir, vamos substituir os valores verdadeiros $u_{1i}, u_{2i}, ..., u_{ki}$ e $v_i = f(u_{1i}, u_{2i}, ..., u_{ki}; \beta_1, \beta_2, ..., \beta_P)$ na Eq. 2.7 pelas suas respectivas estimativas $\bar{x}_{1i}, \bar{x}_{2i}, ..., \bar{x}_{ki}$ e $\hat{y}_i = f(\bar{x}_{1i}, \bar{x}_{2i}, ..., \bar{x}_{ki}; b_1, b_2, ..., b_P)$:

$$P=\prod_{i=1}^{n}\prod_{j=1}^{r}\left[\left(\frac{1}{\sqrt{2\pi}\sigma_y}e^{-\frac{\left(y_{ij}-\hat{y}_i\right)^2}{2\sigma_y^2}}\right)\left(\frac{1}{\sqrt{2\pi}\sigma_{x1}}e^{-\frac{\left(x1_{ij}-\bar{x}1_i\right)^2}{2\sigma_{x1}^2}}\right)\ldots\left(\frac{1}{\sqrt{2\pi}\sigma_{k2}}e^{-\frac{\left(xk_{ij}-\bar{x}k_i\right)^2}{2\sigma_k^2}}\right)\right] \quad (2.8)$$

Sendo assim, obtemos a função de verossimilhança que fornece a probabilidade de ocorrência de um dado conjunto de observações experimentais em função das estimativas para os parâmetros do modelo e para os valores verdadeiros das variáveis explicativas. As melhores estimativas $b_1, b_2, ..., b_L$ para os parâmetros do modelo e $\bar{x}_{1i}, \bar{x}_{2i}, ..., \bar{x}_{ki}$ para os valores verdadeiros das variáveis explicativas em cada uma das observações são aquelas que maximizam a Eq. 2.8. É fácil demonstrar que este problema de otimização é equivalente à minimização de:

$$S=\sum_{i=1}^{n}\sum_{l=1}^{r}\left[\frac{1}{\sigma_{yi}^2}\left(y_{il}-\hat{y}_i\right)^2+\frac{1}{\sigma_{x1i}^2}\left(x_{1il}-\bar{x}_{1i}\right)^2+\ldots+\frac{1}{\sigma_{xki}^2}\left(x_{kil}-\bar{x}_{ki}\right)^2\right] \quad (2.9)$$

O ponto de mínimo da Eq. 2.9 deve ser obtido numericamente, utilizando um método tipo Newton, por exemplo. A Figura 2.4 apresenta o fluxograma de cálculo que permite a implementação de uma função que toma como entradas os parâmetros do modelo e os valores verdadeiros das variáveis explicativas em cada observação, retornando o somatório *S* a ser minimizado (Eq. 2.9).

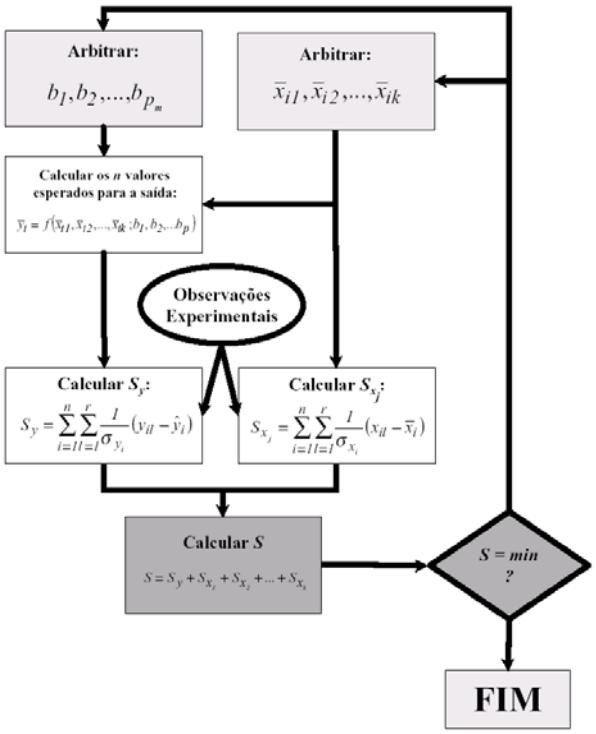

Figura 2.4: Rotina genérica para o método da máxima verossimilhança.

Embora o método ma máxima verossimilhança permita a consideração de erros de medida nas variáveis explicativas de maneira consistente, a minimização da Eq. 2.9 pode vir a

ser uma tarefa complicada. Como, além dos parâmetros do modelo, temos que estimar ainda os valores verdadeiros das variáveis explicativas para cada uma das *n* observações, o número de parâmetros desconhecidos é relativamente grande, o que dificulta a obtenção da solução.

Se assumirmos que, para cada uma das n observações, as medidas $x_{1i}, x_{2i}, ..., x_{ki}$ das variáveis explicativas podem ser obtidas com erro desprezível, os termos $(x_{1i} - \bar{x}_{1i})^2, ..., (x_{ki} - \bar{x}_{ki})^2$ da Eq. 2.9 tornam-se nulos e podem ser eliminados:

$$S = \sum_{i=1}^{n} \sum_{j=1}^{r} \left[\frac{1}{\sigma_{yi}^2} (y_{ij} - \hat{y}_i)^2 \right] \tag{2.10}$$

Nestas condições, a maximização da probabilidade de ocorrência das observações disponíveis torna-se equivalente a minimização da soma ponderada dos quadrados dos desvios das predições do modelo em relação às medidas experimentais *y*. Ou seja, um problema de mínimos quadrados ponderado pela recíproca da variância experimental. Assumindo que σ_y é constante, recaímos em um problema de mínimos quadrados ordinários clássico:

$$S = \sum_{i=1}^{n} \sum_{j=1}^{r} \left[(y_{ij} - \hat{v}_i)^2 \right] \tag{2.11}$$

A obtenção da solução de mínimos quadrados é relativamente simples. Se o problema em questão for linear nos parâmetros, ela pode ser obtida analiticamente. Para casos mais gerais temos de recorrer a métodos de otimização. A Figura 2.5, ilustra a idéia básica deste método para o caso onde a função *f*, mostrada na Eq. 2.5, é uma linha reta ($y = b_1 x + b_0$). De posse das estimativas b_0 *e* b_1, podemos traçar uma linha reta no plano *xy*. Cada ponto experimental neste plano estará situado a uma distância vertical e_i da linha reta. As estimativas b_0 *e* b_1 são determinadas de modo a minimizar a soma dos quadrados destas distâncias, $||e||^2$, que equivale a variabilidade residual do modelo.

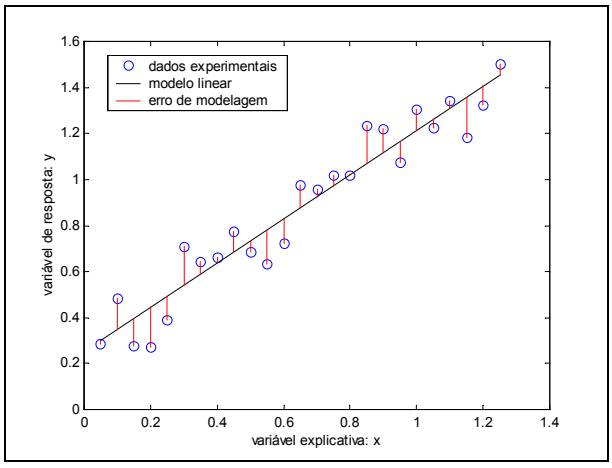

Figura 2.5: Ilustração do procedimento de mínimos quadrados.

Devido às significativas simplificações que traz às computações, o emprego do método dos mínimos quadrados tornou-se bastante popular e, com freqüência, a técnica é utilizada em situações onde as hipóteses assumidas na sua formulação não são válidas. Nestes casos, a motivação para o emprego da técnica é a minimização da soma dos quadrados dos desvios individuais e_i, ou seja, os parâmetros são estimados de modo que a diferença entre as predições do modelo e os valores medidos para a variável de resposta seja a menor possível.

2.3. Regressão Linear Múltipla por Mínimos Quadrados

Quando o problema de modelagem é linear em relação aos parâmetros do modelo, o problema de estimação por mínimos quadrados apresenta solução analítica. Nesta seção, os passos a serem seguidos para a obtenção desta solução são detalhados.

O ponto de partida para a formulação do problema são as observações experimentais do sistema. Na prática, para cada uma das n observações, trabalhamos com as medidas $x_{i1}, x_{i2}, ..., x_{ik}$ das variáveis explicativas e com a medida y_i para a variável de resposta. Utilizamos estas medidas para obter as estimativas $b_0, b_1, b_2, ..., b_k$ para os parâmetros do modelo. Para efeitos de simplificação, estamos supondo que não há réplicas para as medidas das variáveis em cada observação, de modo que, segundo a notação estabelecida, o conjunto de dados experimentais é constituído de apenas de n pontos (um por observação). Deste modo, o modelo da Eq. 2.5 é escrito da seguinte forma:

$$\hat{y}_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + ... + b_k x_{ik} \tag{2.12}$$

Na Eq. 2.12, o subscrito i denota as diferentes observações experimentais. Este modelo pode ser generalizado na forma matricial:

$$\hat{\boldsymbol{y}} = \boldsymbol{Xb} \tag{2.13}$$

onde:

$$\hat{\boldsymbol{y}} = [\hat{y}_1, \hat{y}_2, ... \hat{y}_n]^T \tag{2.14}$$

é o vetor contendo as predições do modelo para o valor da variável de resposta em cada uma das n observações experimentais,

$$\boldsymbol{b} = [b_0, b_1, b_2, ... b_k]^T \tag{2.15}$$

é o vetor cujos coeficientes representam as estimativas para cada um dos $k+1$ coeficientes do modelo linear e, por fim,

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1k} \\ 1 & x_{12} & x_{22} & \dots & x_{k2} \\ 1 & x_{13} & x_{23} & \dots & x_{k3} \\ . & . & . & . & . \\ 1 & x_{1k} & x_{2k} & \dots & x_{kk} \end{bmatrix} \tag{2.16}$$

é a matriz das entradas, cujas colunas apresentam as medidas para as variáveis explicativas em cada observação experimental. Como estamos assumindo que as medidas das variáveis explicativas são obtidas com erro nulo, os valores de $x_{i1}, x_{i2}, ..., x_{ik}$ equivalem aos valores verdadeiros das entradas $u_1, u_2, ..., u_k$. As predições do modelo $\hat{y}_i$ são estimativas para os valores verdadeiros, v_i, da variável de resposta. Devido às incertezas presentes nas medidas y_i (e/ou a outros fatores como a não consideração de variáveis explicativas importantes, por exemplo) os valores preditos pelo modelo diferem dos valores verdadeiros da variável de resposta v_i por uma quantia e_i, o erro de predição. Deste modo, podemos escrever:

$$\boldsymbol{v} = \hat{\boldsymbol{y}} + \boldsymbol{e} \tag{2.17}$$

onde

$$\boldsymbol{e} = [e_1, e_2, ...e_n]^T \tag{2.18}$$

é o vetor cujos elementos representam a diferença entre a predição do modelo e o valor verdadeiro da variável de resposta. O método dos mínimos quadrados busca a determinação do vetor dos coeficientes $\boldsymbol{b}$ de modo que a soma dos quadrados dos elementos e_i seja minimizada. Sendo assim, a estimação dos parâmetros consiste na minimização da variabilidade residual do modelo (Eq. 2.19), que equivale a soma do quadrado dos desvios individuais. No ponto de vista matricial, a solução de mínimos quadrados é aquela que minimiza a norma quadrática do vetor $\boldsymbol{e}$:

$$S_y^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \|\boldsymbol{e}\|^2$$

$$S_y^2 = \sum_{i=1}^{n} [y_i - (b_0 + b_1 x_{i1} + b_2 x_{i2} + ... + b_k x_{ik})]^2 \tag{2.19}$$

O ponto de mínimo da Eq. 2.19 é aquele onde as derivadas em relação a cada um dos parâmetros do modelo ($b_0, b_1, b_2, ..., b_k$) são nulas. Sendo assim, devemos diferenciar a função $S_y^2(b_0, b_1, b_2, ..., b_k)$ em relação a cada um dos parâmetros e igualar a expressão obtida a zero. Desta forma, iremos obter um sistema com *k+1* equações e *k+1* variáveis, cuja solução fornece as estimativas para os parâmetros que minimizam a soma quadrática dos erros:

$$\begin{aligned}
&\frac{\partial S}{\partial b_0} = -2\sum_{i=1}^{n}\left(y_i - b_0 - b_1 x_{i1} - b_2 x_{i2} - \ldots - b_k x_{ik}\right) = 0 \\
&\frac{\partial S}{\partial b_1} = -2\sum_{i=1}^{n}\left(y_i - b_0 - b_1 x_{i1} - b_2 x_{i2} - \ldots - b_k x_{ik}\right).x_{i1} = 0 \\
&\frac{\partial S}{\partial b_2} = -2\sum_{i=1}^{n}\left(y_i - b_0 - b_1 x_{i1} - b_2 x_{i2} - \ldots - b_k x_{ik}\right).x_{21} = 0 \\
&\vdots \\
&\frac{\partial S}{\partial b_k} = -2\sum_{i=1}^{n}\left(y_i - b_0 - b_1 x_{i1} - b_2 x_{i2} - \ldots - b_k x_{ik}\right).x_{k1} = 0
\end{aligned} \tag{2.20}$$

Reordenando os somatórios, obtemos o seguinte sistema de equações lineares:

$$\begin{aligned}
&nb_0 + b_1\sum_{i=1}^{n} x_{i1} + b_2\sum_{i=1}^{n} x_{i2} + \ldots + b_k\sum_{i=1}^{n} x_{ik} = \sum_{i=1}^{n} y_i \\
&b_0\sum_{i=1}^{n} x_{i1} + b_1\sum_{i=1}^{n} x_{i1}x_{i1} + \ldots + b_k\sum_{i=1}^{n} x_{i1}x_{ik} = \sum_{i=1}^{n} x_{i1}y_i \\
&b_0\sum_{i=1}^{n} x_{i2} + b_1\sum_{i=1}^{n} x_{i2}x_{i1} + \ldots + b_k\sum_{i=1}^{n} x_{ik}x_{i2} = \sum_{i=1}^{n} x_2 y_i \\
&\vdots \\
&b_0\sum_{i=1}^{n} x_{ik} + b_1\sum_{i=1}^{n} x_{ik}x_{i1} + \ldots + b_k\sum_{i=1}^{n} x_{ik}x_{ik} = \sum_{i=1}^{n} x_{ik}y_i
\end{aligned} \tag{2.21}$$

que pode ser colocado na forma matricial:

$$\begin{bmatrix}
n & \sum_{i=1}^{n} x_{i1} & \sum_{i=1}^{n} x_{i2} & \ldots & \sum_{i=1}^{n} x_{ik} \\
\sum_{i=1}^{n} x_{i1} & \sum_{i=1}^{n} x_{i1}x_{i1} & \sum_{i=1}^{n} x_{i1}x_{i2} & \ldots & \sum_{i=1}^{n} x_{i1}x_{ik} \\
\sum_{i=1}^{n} x_{i2} & \sum_{i=1}^{n} x_{i2}x_{i1} & \sum_{i=1}^{n} x_{i2}x_{i2} & \ldots & \sum_{i=1}^{n} x_{i2}x_{ik} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
\sum_{i=1}^{n} x_{ik} & \sum_{i=1}^{n} x_{ik}x_{i1} & \sum_{i=1}^{n} x_{ik}x_{i1} & \ldots & \sum_{i=1}^{n} x_{ik}x_{ik}
\end{bmatrix}
\begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ \vdots \\ b_k \end{bmatrix} =
\begin{bmatrix}
\sum_{i=1}^{n} y_i \\
\sum_{i=1}^{n} x_{i1}y_i \\
\sum_{i=1}^{n} x_{i2}y_i \\
\vdots \\
\sum_{i=1}^{n} x_{ik}y_i
\end{bmatrix} \tag{2.22}$$

ou seja:

$$\left(\boldsymbol{X}^T\boldsymbol{X}\right)\boldsymbol{b} = \boldsymbol{(X}^T\boldsymbol{y)} \tag{2.23}$$

se multiplicarmos os dois lados da equação acima por $\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}$ obtemos:

$$\boldsymbol{b} = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{(X}^T\boldsymbol{y)} \tag{2.24}$$

que é a solução de clássica para o problema de mínimos quadrados. Conforme demonstrado na seção anterior, se assumirmos que as réplicas das medidas experimentais se distribuem normalmente em torno do valor verdadeiro das variáveis e que as variáveis explicativas são

determinadas com erro desprezível, o método dos mínimos quadrados é um caso especial do método da máxima verossimilhança. Isso significa que, sob a validade destas hipóteses, a estimativa de $\boldsymbol{b}$ obtida pela minimização da Eq. 2.19 é ótima do ponto de vista estatístico uma vez que torna a ocorrência dos dados experimentais disponíveis tão provável quanto possível.

2.4. Análise Estatística da Regressão Linear Múltipla

Na prática, um modelo é utilizado para predizer o comportamento de uma variável de resposta. Como o erro presente nas medidas experimentais se propaga para as estimativas dos parâmetros, as predições fornecidas pelo modelo sempre estão associadas a algum grau de incerteza. Portanto, após a etapa de estimação de parâmetros, é importante que seja conduzida uma análise estatística da regressão, de modo que possam ser avaliadas as seguintes questões:

- o intervalo de predição do modelo (análise dos erros de predição);
- a adequabilidade do modelo (uso do teste F e covariâncias);
- a significância estatística dos parâmetros (uso do teste t);
- os intervalos de confiança dos parâmetros (uso das variâncias).

Para que estas questões possam ser avaliadas, é necessário que a matriz de covariância das predições $\boldsymbol{S}_y^{'2}$ seja conhecida. Obedecendo a notação seguida na seção anterior, $\boldsymbol{S}_y^{'2}$ é dada pelo produto vetorial:

$$\boldsymbol{S}_y^2 = (\hat{\boldsymbol{y}} - \boldsymbol{v})(\hat{\boldsymbol{y}} - \boldsymbol{v})^T \tag{2.25}$$

Substituindo o vetor das predições do modelo $\hat{\boldsymbol{y}}$ por $\boldsymbol{Xb}$ e o vetor dos valores verdadeiros da variável de resposta $\boldsymbol{v}$ por $\boldsymbol{X\beta}$:

$$\boldsymbol{S}_y^2 = (\boldsymbol{Xb} - \boldsymbol{X\beta})(\boldsymbol{Xb} - \boldsymbol{X\beta})^T$$

$$\boldsymbol{S}_y^2 = [\boldsymbol{X}(\boldsymbol{b} - \boldsymbol{\beta})][\boldsymbol{X}(\boldsymbol{b} - \boldsymbol{\beta})]^T$$

$$\boldsymbol{S}_y^2 = \boldsymbol{X}(\boldsymbol{b} - \boldsymbol{\beta})(\boldsymbol{b} - \boldsymbol{\beta})^T \boldsymbol{X}^T \tag{2.26}$$

onde o produto vetorial $\boldsymbol{(b-\beta)(b-\beta)'}$ é a matriz de covariância dos parâmetros, $\boldsymbol{S}_b^2$:

$$\boldsymbol{S}_b^2 = \boldsymbol{(b-\beta)(b-\beta)}^T \tag{2.27}$$

Substituindo o vetor da estimativa dos parâmetros $\boldsymbol{b}$ pela solução de mínimos quadrados para o problema de regressão linear múltipla, obtém-se:

$$\boldsymbol{S}_b^2 = \left[(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} - (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{v}\right]\left[(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} - (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{v}\right]^T$$

$$\boldsymbol{S}_{\boldsymbol{b}}^{\boldsymbol{2}} = \left[\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-\boldsymbol{1}}\boldsymbol{X}^T\left(\boldsymbol{y}-\boldsymbol{v}\right)\right]\left[\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-\boldsymbol{1}}\boldsymbol{X}^T\left(\boldsymbol{y}-\boldsymbol{v}\right)\right]^T$$

$$\boldsymbol{S}_{\boldsymbol{b}}^{\boldsymbol{2}} = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-\boldsymbol{1}}\boldsymbol{X}^T\left(\boldsymbol{y}-\boldsymbol{v}\right)\left(\boldsymbol{y}-\boldsymbol{v}\right)^T\boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-\boldsymbol{1}^T}$$

$$\boldsymbol{S}_{\boldsymbol{b}}^{\boldsymbol{2}} = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-\boldsymbol{1}}\boldsymbol{X}^T\boldsymbol{\sigma}_{\boldsymbol{y}}^{\boldsymbol{2}}\boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-\boldsymbol{1}^T} \tag{2.28}$$

onde $\boldsymbol{\sigma}_{\boldsymbol{y}}^{\boldsymbol{2}}(n,n)$ é a matriz de covariância dos erros experimentais. Se os erros experimentais não estão correlacionados e apresentam uma variância σ_y^2 constante, a matriz $\boldsymbol{\sigma}_{\boldsymbol{y}}^{\boldsymbol{2}}$ é equivalente a $\sigma_y^2\boldsymbol{I}$ e, então, pode-se escrever:

$$\boldsymbol{S}_{\boldsymbol{b}}^{\boldsymbol{2}} = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-\boldsymbol{1}}\boldsymbol{X}^T\boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-\boldsymbol{1}^T}\sigma_y^2$$

$$\boldsymbol{S}_{\boldsymbol{b}}^{\boldsymbol{2}} = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-\boldsymbol{1}}\sigma_y^2 \tag{2.29}$$

Finalmente, esta expressão para $\boldsymbol{S}_{\boldsymbol{b}}^{\boldsymbol{2}}$ pode ser substituída na Eq. 2.26, fornecendo a expressão para a matriz de covariância das predições:

$$\boldsymbol{S}_{\boldsymbol{y}}^{\boldsymbol{2}} = \boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-\boldsymbol{1}}\boldsymbol{X}^T\sigma_y^2 \tag{2.30}$$

De posse da matriz $\boldsymbol{S}_{\boldsymbol{y}}^{\boldsymbol{2}}$, as questões lançadas no início desta seção podem ser avalidas. Para uma revisão de como os testes estatísticos devem ser aplicados, recomenda-se a leitura de Secchi (1997).

2.5. Modelos Não Lineares (Transformação de Variáveis)

Nas seções anteriores, foi feita uma revisão a respeito do modelo de regressão linear múltipla. Este modelo considera que uma dada variável de resposta se relaciona linearmente com todas as entradas presentes no modelo. Em muitas situações práticas, a relação existente entre as diferentes variáveis do sistema pode apresentar curvaturas consideráveis, o que torna a utilização do modelo linear inapropriada. Com o intuito de considerar efeitos de não-linearidade na construção de modelos empíricos multivariáveis, Box e Tidwell (1962) investigaram o uso de transformações nas variáveis explicativas. Primeiramente, os autores apresentaram um procedimento genérico que permite que a variável de resposta seja considerada como uma função das variáveis de entrada transformadas. Tanto a função que relaciona a resposta com as variáveis transformadas, como as transformações a serem realizadas em cada entrada são assumidas como arbitrárias. Em particular, eles propuseram a utilização de uma combinação linear de potências das variáveis explicativas originais:

$$\boldsymbol{y} = b_0 + b_1\boldsymbol{x}_1^{a_1} + b_2\boldsymbol{x}_2^{a_2} + \ldots + b_k\boldsymbol{x}_k^{a_k} \tag{2.31}$$

Este modelo é, na realidade, uma extensão do modelo linear trabalhado anteriormente, onde adicionamos um expoente real em cada uma das variáveis de entrada. A inclusão destes

parâmetros adicionais faz com que não seja possível a obtenção de uma solução analítica para o problema. Os autores então propõem um procedimento iterativo do tipo Newton-Raphson para a obtenção das estimativas dos parâmetros a partir das observações experimentais.

O procedimento iterativo inicia assumindo-se uma aproximação inicial $a_1^0, a_2^0, ...a_k^0$ para os valores dos expoentes $a_1, a_2, ...a_k$ e expandindo-se a expressão para $\boldsymbol{y}$,dada pela Eq. 2.31, em série de Taylor em torno destes valores. Ignorando-se os termos de ordem superior a um, obtemos uma expressão que aproxima $\boldsymbol{y}$ como uma função linear de $a_1, a_2, ...a_k$.

$$\boldsymbol{y} = b_0 + \sum_{j=1}^{k} b_j \boldsymbol{x}_j^{a_j^0} + \sum_{j=1}^{k} (a_j - a_j^0) \frac{\partial \boldsymbol{y}}{\partial \boldsymbol{aj}}\bigg|_{a_j = a_j^0} \tag{2.32}$$

Uma escolha conveniente para os valores iniciais de $a_1, a_2, ...a_k$ é assumirmos que todos os expoentes são iguais à unidade, como se não estivesse sendo realizada nenhuma transformação, ou seja:

$$\boldsymbol{y} = b_0 + b_1 \boldsymbol{x}_1 + b_2 \boldsymbol{x}_2 + \boldsymbol{...} + b_k \boldsymbol{x}_k \tag{2.33}$$

Deste modo, a Eq. 2.32 torna-se:

$$\boldsymbol{y} = b_0 + \sum_{j=1}^{k} b_j \boldsymbol{x}_j + \sum_{j=1}^{k} (a_j - 1) b_j \boldsymbol{x}_j \boldsymbol{ln}(\boldsymbol{x}_j) \tag{2.34}$$

Como os valores b_j no último termo da Eq. 2.34 são desconhecidos, é necessário que, de algum modo, sejam obtidas estimativas para estes valores. Os valores b_j podem ser estimados convenientemente a partir da Eq. 2.33, através do método dos mínimos quadrados.

De posse das estimativas para b_j, os produtos $b_j \boldsymbol{x}_j \boldsymbol{ln}(\boldsymbol{x}_j)$*'s* podem ser encarados como novos conjuntos de "variáveis explicativas" e as estimativas para $(a_j - 1)$ podem ser obtidas por regressão linear múltipla. Posteriormente, as variáveis explicativas originais $\boldsymbol{x_1, x_2, ..., x_k}$ podem ser substituídas por $\boldsymbol{x_1^{a_1}, x_2^{a_2}, ..., x_k^{a_k}}$ na Eq. 2.31 e todo o ciclo pode ser repetido, até que seja atingida a convergência.

2.6. Métodos de Redução de Dimensionalidade

Outra questão importante a ser considerada na construção de modelos empíricos é o problema da colinearidade, que consiste na presença de interdependência mútua entre as variáveis explicativas. Se as colunas de $\boldsymbol{X}$ forem linearmente dependentes, enfrentaremos problemas na inversão da matriz $(\boldsymbol{X}^T \boldsymbol{X})$ e, conseqüentemente, na determinação do vetor $\boldsymbol{b}$. Nestas situações, são de grande valor os métodos de redução de dimensionalidade que permitem que o problema da colinearidade seja encarado utilizando-se todas as colunas de $\boldsymbol{X}$ na composição do modelo. Neste trabalho são tratados os métodos da análise dos componentes pricipais (*PCA - principal component regression*) e dos mínimos quadrados parciais (*PLS - partial least squares*). A análise dos componentes principais é equivalente a decomposição da matriz de dados em seus valores singulares. A teoria do *PCA* está bem

desenvolvida e explicações aprofundadas podem ser encontradas em livros textos de análise multivariável como, por exemplo, Mardia et al. (2000) e Höskuldsson (1996). O método *PLS* pode ser encarado como uma extensão do método *PCA*. Geladi e Kowalski (1986) apresentam um excelente tutorial para esta metodologia. Inicialmente, o método *PLS* foi muito discutido em termos do seu algoritmo. Entretanto, alguns trabalhos têm sido desenvolvidos para entender a sua estrutura em um nível mais fundamental. Importantes contribuições nesta área foram feitas por Höskuldson (1988 e 1996) e Helland (1988).

A idéia básica por traz dos métodos de redução de dimensionalidade consiste em realizar a regressão sobre uma projeção da matriz original $\boldsymbol{X}$ em um subespaço de dimensão reduzida, que procura eliminar a informação redundante e o ruído presente nos dados. A diferença básica entre as técnicas de redução de dimensionalidade está na maneira como esta decomposição é realizada. No caso do *PCA*, a matriz $\boldsymbol{X}$ é decomposta em seus componentes principais. No caso do *PLS*, a matriz $\boldsymbol{X}$ é decomposta buscando-se as direções que melhor descrevem a variável de resposta.

Nesta seção, primeiramente, são revisados os métodos *PCA* e *PLS* lineares. Posteriormente, é apresentada a extensão não linear para o algoritmo *PLS* proposta por Wold et al. (1989) e seu caso particular, o *QPLS* (*quadratic partial least squares*). Por fim, apresenta-se o algoritmo *BTPLS* (*Box-Tidwell based partial least squares*), proposto por Li et al. (2001), que utiliza um procedimento de transformação de variáveis flexível, capaz de se ajustar a uma ampla variedade de curvas.

2.6.1. Análise de Componentes Principais (PCA)

Nesta seção, será apresentado o procedimento de decomposição em estruturas latentes baseado na análise dos componentes principais (*PCA - Principal Component Analysis*). Na verdade, a técnica *PCA* da origem aos demais métodos de redução de dimensionalidade que serão apresentados posteriormente. Basicamente, *PCA* é um método de escrever a matriz $\boldsymbol{X(n,k)}$ como uma soma de a matrizes, todas com posto unitário:

$$\boldsymbol{X} = \boldsymbol{X_1} + \boldsymbol{X_2} + \ldots + \boldsymbol{X_a} \tag{2.35}$$

A motivação para isso, no que se refere a problemas de regressão, está na possibilidade de, realizando esta decomposição na matriz das entradas, separarmos a informação útil da informação redundante. Desta forma, podemos trabalhar com novas variáveis, que são combinações lineares independentes das variáveis originais, eliminando assim o problema da colinearidade.

Calculamos as matrizes $\boldsymbol{X_1}, \boldsymbol{X_2}, \ldots, \boldsymbol{X_a}$ realizando o produto externo de dois vetores, conforme mostram a Eq. 2.36 e a Figura 2.6:

$$\boldsymbol{X} = \boldsymbol{t_1}\boldsymbol{p_1}^T + \boldsymbol{t_2}\boldsymbol{p_2}^T + \ldots + \boldsymbol{t_a}\boldsymbol{p_a}^T \tag{2.36}$$

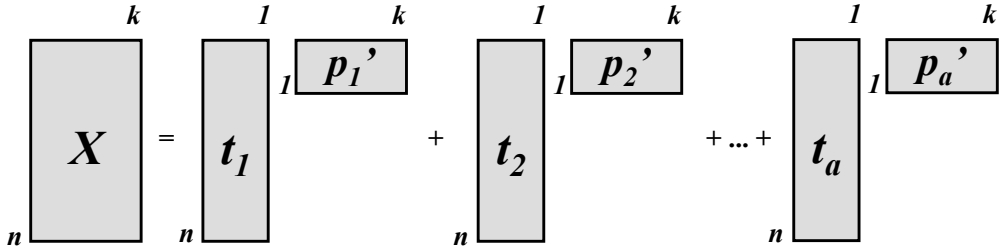

Figura 2.6: Representação gráfica da decomposição *PCA* em produtos de vetoriais.

A decomposição acima também pode ser escrita na forma do produto entre as matrizes $\boldsymbol{T}'(n,a)$ e $\boldsymbol{P}^T(a,k)$, onde $\boldsymbol{t_1},\boldsymbol{t_2},...,\boldsymbol{t_a}$ constituem as colunas de $\boldsymbol{T}$ e $\boldsymbol{p}_1^T,\boldsymbol{p}_2^T,...,\boldsymbol{p}_a^T$ constituem as linhas de $\boldsymbol{P}'$:

$$\boldsymbol{X} = \boldsymbol{T}\boldsymbol{P}^T \tag{2.37}$$

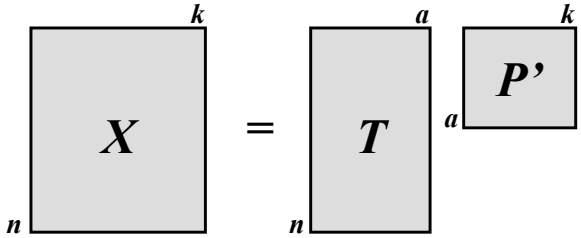

Figura 2.7: Representação gráfica da decomposição *PCA* em produto matricial.

Cada um dos vetores $\boldsymbol{t_1},\boldsymbol{t_2},...,\boldsymbol{t_a}$ na Eq. 2.36 é composto de *n* elementos, resultantes de combinações lineares das colunas de $\boldsymbol{X}$ ao passo que os vetores $\boldsymbol{p}_1^T,\boldsymbol{p}_2^T,...,\boldsymbol{p}_a^T$ são compostos de *k* elementos, resultantes de combinações lineares das linhas de $\boldsymbol{X}$. Os vetores $\boldsymbol{t_1},\boldsymbol{t_2},...,\boldsymbol{t_a}$ (*score vectors*) e $\boldsymbol{p}_1^T,\boldsymbol{p}_2^T,...,\boldsymbol{p}_a^T$ (*loading vectors*) são extraídos da matriz original $\boldsymbol{X}$ através de vetores pesos, que devem apresentar comprimento unitário:

$$\boldsymbol{t} = \boldsymbol{X}\boldsymbol{w} \tag{2.38}$$

$$\boldsymbol{p}^T = \boldsymbol{X}^T\boldsymbol{v} \tag{2.39}$$

Na prática, os vetores $\boldsymbol{t_1},\boldsymbol{t_2},...,\boldsymbol{t_a}$ e $\boldsymbol{p}_1^T,\boldsymbol{p}_2^T,...,\boldsymbol{p}_a^T$ não são determinados todos simultaneamente. Determinamos os pares de vetores pesos $(\boldsymbol{w_1},\boldsymbol{v_1}),(\boldsymbol{w_2},\boldsymbol{v_2}),...,(\boldsymbol{w_a},\boldsymbol{v_a})$ um a um. A partir do primeiro par (w_1,v_1) extraímos a direção $\boldsymbol{t_1}$ da matriz de dados original ($\boldsymbol{X_0}$=$\boldsymbol{X}$). Podemos então ortogonalizar a matriz $\boldsymbol{X_0}$ em relação ao que foi extraído:

$$\boldsymbol{X_1} = \boldsymbol{X_0} - \boldsymbol{t_1}\boldsymbol{p}^T \tag{2.40}$$

A matriz residual $\boldsymbol{X_1}$, obtida a partir da Eq. 2.40, contém a informação de $\boldsymbol{X_0}$ que é ortogonal a $\boldsymbol{t_1}$. Isso nos garante que, se determinarmos um novo par de vetores pesos (w_2,v_2), poderemos extrair de $\boldsymbol{X_1}$ uma nova direção $\boldsymbol{t_2}$, que será linearmente independente de $\boldsymbol{t_1}$. Podemos então ortogonalizar a matriz $\boldsymbol{X_1}$ em relação à direção $\boldsymbol{t_2}$ e repetir o procedimento, continuando até extrairmos *a* direções, que serão todas independentes entre si.

O ponto chave do método *PCA* está na escolha dos $\boldsymbol{w}$'s e $\boldsymbol{v}$'s, uma vez que todas as computações são efetuadas a partir destes valores. Para ilustrar como os pesos são determinados, vamos nos fixar na extração do primeiro componente principal, $\boldsymbol{t_1}$. Conforme mostra a Figura 2.8, a escolha do vetor peso $\boldsymbol{w_1}$ irá determinar a combinação linear das colunas de $\boldsymbol{X_0}$ que irá gerar $\boldsymbol{t_1}$, enquanto a escolha do vetor peso $\boldsymbol{v}$ irá determina a combinação linear das linhas de $\boldsymbol{X_0}$ que irá gerar $\boldsymbol{p_1'}$.

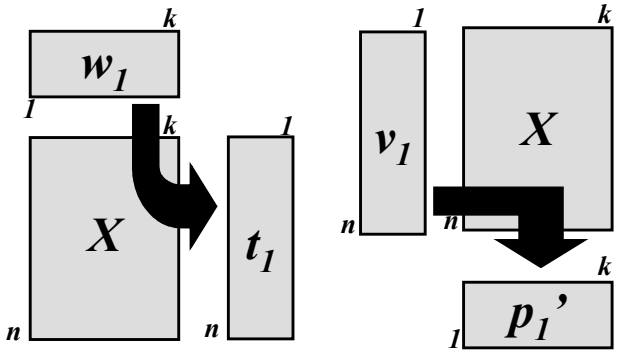

Figura 2.8: Extração dos vetores peso $\boldsymbol{w}$ e $\boldsymbol{v}$ a partir da matriz $\boldsymbol{X}$.

O primeiro componente principal, vetor $\boldsymbol{t_1}$, concentra parte da informação originalmente inserida nas colunas de $\boldsymbol{X_0}$. Quando extraímos a primeira direção $\boldsymbol{t_1}$ do espaço coluna da matriz $\boldsymbol{X_0}$, desejamos que ela concentre a maior quantidade de informação possível. E é nisso que nos baseamos para determinar o critério de escolha para $\boldsymbol{w_1}$. A quantidade de informação contida na direção $\boldsymbol{t_1}$, assim como nas colunas de $\boldsymbol{X_0}$, é medida pela respectiva variância. Como estamos lidando com vetores de média nula a variância é equivalente a norma quadrática dos mesmos. Devemos então, para garantir que o volume de informação extraído na primeira direção seja o maior possível, escolher o vetor peso $\boldsymbol{w_1}$ que maximize a norma quadrática do vetor $\boldsymbol{t_1}$.

Como o vetor $\boldsymbol{w_1}$ deve apresentar comprimento unitário, estamos diante de um problema de otimização com restrição. A restrição em questão pode ser inserida na função objetivo através da técnica dos multiplicadores de Lagrange. Nossa meta é determinar o vetor de $\boldsymbol{w_1}$, de comprimento unitário, tal que $\boldsymbol{t_1 = X_0 w_1}$ apresente comprimento quadrático máximo. Isso equivale a maximização de $\boldsymbol{t_1}^{'}\boldsymbol{t_1} = \boldsymbol{w_1}^{'}\boldsymbol{X_0}^{'}\boldsymbol{X_0 w_1}$. Portanto, podemos contruir a seguinte função objetivo a ser maximizada:

$$S = \boldsymbol{w_1}^T \boldsymbol{X_0}^T \boldsymbol{X_0 w_1} - \lambda\left(\boldsymbol{w_1}^T \boldsymbol{w_1} - 1\right) \tag{2.41}$$

Para encontrar o ponto de máximo, vamos então diferenciar S em relação a $\boldsymbol{w_1}$:

$$\frac{\partial S}{\partial \boldsymbol{w_1}} = 2\boldsymbol{X_0}^T \boldsymbol{X_0 w_1} - 2\lambda \boldsymbol{w_1} \tag{2.42}$$

Igualando a derivada de S em relação a $\boldsymbol{w_1}$ a zero, obtemos:

$$\boldsymbol{X_0}^T \boldsymbol{X_0 w_1} = \lambda \boldsymbol{w_1} \tag{2.43}$$

Sendo assim, a extração do primeiro componente principal da matriz $\boldsymbol{X_0}$ é equivalente ao cálculo dos autovetores $\boldsymbol{w}$*'s* e dos autovalores λ*'s* da matriz $\boldsymbol{X_0}^T\boldsymbol{X_0}$. É importante verificarmos que a Eq. 2.43 tem múltiplas soluções. Como $\lambda=\boldsymbol{t}^T\boldsymbol{t}$, a escolha do autovetor $\boldsymbol{w}=\boldsymbol{w_1}$ associado ao maior autovalor, $\boldsymbol{\lambda}=\boldsymbol{\lambda_1}$, garante que a direção $\boldsymbol{t_1} = \boldsymbol{X_0 w_1}$ é aquela que apresenta variância máxima. Os demais autovetores, soluções da Eq. 2.43, correspondem às demais direções e a serem extraídas, do mesmo modo que os respectivos autovalores correspondem às variâncias das mesmas.

Para extrair as direções $t_1, t_2, ..., t_a$ da matriz de dados original, iremos utilizar o algoritmo *NIPLAS* (*Nonlinear Iterative Partial Least Squares*). Pode ser demonstrado que, na convergência, os resultados obtidos por este algoritmo são os mesmos que os calculados a partir dos autovetores. Uma explicação mais detalhada de como *NIPALS* funciona pode se encontrada em Geladi (1986). Basicamente, parte-se da matriz de dados original $\boldsymbol{X}$ e, através de um procedimento iterativo, os pares $\boldsymbol{t}$*'s* e $\boldsymbol{p}$*'s* são determinados um a um. Um sumário com os passos básicos do algoritmo *NIPALS* é apresentado na Tabela 2.3.

Tabela 2.3: Algoritmo *NIPALS* para *PCA*.

Passo	Sumário do Passo	Computação
0	Normalizar e centrar $\boldsymbol{X}$.	
1	Tomar a primeira coluna de $\boldsymbol{X}$ como aproximação inicial para $\boldsymbol{t_h}$.	
2	Calcular $\boldsymbol{p_h}$	$\boldsymbol{p_h}^T = \boldsymbol{t_h}^T\boldsymbol{X}/\boldsymbol{t_h}^T\boldsymbol{t_h}$
3	Normalizar $\boldsymbol{p_h}$	$\boldsymbol{p_h} = \boldsymbol{p_h}/\|\boldsymbol{p_h}\|$
4	Calcular $\boldsymbol{t_h}$	$\boldsymbol{t_h} = \boldsymbol{X}\boldsymbol{p_h}/\boldsymbol{p_h}^T\boldsymbol{p_h}$
5	Comparar o valor de $\boldsymbol{t_h}$ obtido em 4 com o utilizado em 2. Se são iguais avançar para. Senão voltar para 2.	
6	Calcular a matriz residual, ortogonalizando $\boldsymbol{X}$ em relação ao que foi extraído.	$\boldsymbol{F} = \boldsymbol{X} - \boldsymbol{t_h}\boldsymbol{p_h}^T$
7	Se direções adicionais forem necessárias, substituir $\boldsymbol{X}$ por $\boldsymbol{F}$ e voltar para o passo 1.	

Cada um dos vetores $\boldsymbol{t_1}, \boldsymbol{t_2}, ..., \boldsymbol{t_a}$ obtidos pelo algoritmo *NIPALS*, concentra uma parcela da informação útil contida nas colunas da matriz $\boldsymbol{X}$ original. No contexto da construção de um modelo de regressão, estas direções apresentam a vantagem de serem independentes entre si. Isto sugere que o modelo de regressão possa ser construído utilizando como variáveis explicativas os vetores $\boldsymbol{t_1}, \boldsymbol{t_2}, ..., \boldsymbol{t_a}$, ou a matriz $\boldsymbol{T(n,a)}$, o que eliminaria os problemas enfrentados na inversão da matriz $\boldsymbol{X'X}$, decorrentes da colinearidade entre as entradas. A regressão realizada a partir dos componentes principais é chamada de *PCR* (*Principal Component Regression*).

Então, em um modelo de regressão linear padrão,

$\hat{y} = Xb$, podemos substituir a matriz de dados original, $\boldsymbol{X}$, pela matriz dos componentes principais, $\boldsymbol{T}$. Neste caso, as predições do novo modelo passariam a ser computadas por:

$$\hat{\boldsymbol{y}} = \boldsymbol{T}\hat{\boldsymbol{\alpha}} \tag{2.44}$$

onde $\hat{\alpha}$ é a estimativa para o verdadeiro vetor de coeficientes do modelo α, obtida a partir de um determinado conjunto de observações experimentais segundo a Eq. 2.45:

$$\hat{\boldsymbol{\alpha}} = \left(\boldsymbol{T}^T\boldsymbol{T}\right)^{-1}\boldsymbol{T}^T\boldsymbol{y} \tag{2.45}$$

Geralmente, é de interesse que se tenha uma solução em termos das variáveis originais do problema e não em termos das variáveis transformadas. A estimativa $\hat{\boldsymbol{b}}$ para o vetor dos coeficientes do modelo *PCR* em termos das variáveis originais pode ser obtida da seguinte maneira:

$$\begin{aligned} \boldsymbol{X}\hat{\boldsymbol{b}} &= \boldsymbol{T}\hat{\boldsymbol{\alpha}} = \hat{\boldsymbol{y}} \\ \boldsymbol{X}\hat{\boldsymbol{b}} &= \boldsymbol{X}\boldsymbol{P}^T\hat{\boldsymbol{\alpha}} \\ \hat{\boldsymbol{b}} &= \boldsymbol{P}^T\hat{\boldsymbol{\alpha}} \end{aligned} \tag{2.46}$$

Foram demonstrados os passos a serem seguidos para obtenção do vetor dos coeficientes do modelo linear pelo do método *PCR*. Este método permite que a regressão seja realizada mesmo na presença de colinearidade entre as variáveis de entrada do modelo e também permite a redução de ruído nos dados através da habilidade de descartar os componentes inferiores da decomposição. Entretanto, esta metodologia extrai as direções da matriz original segundo um critério que se baseia somente na informação presente em $\boldsymbol{X}$, Portanto, há o risco de que informação útil seja confundida com ruído e descartada com os componentes de menor importância. Este risco pode ser minimizado se o critério utilizado na decomposição levar em conta também a informação presente na matriz $\boldsymbol{Y}$. É esta a idéia básica do método *PLS* (*Partial Least Squares* ou *Projection to Latent Structures*) que será detalhado na próxima seção.

2.6.2. Mínimos Quadrados Parciais (PLS)

Originalmente, o método dos mínimos quadrados parciais (*PLS – Partial Least Squares*) foi construído com base nas propriedades do algoritmo *NIPALS*, apresentado no tópico anterior. O algoritmo de decomposição é estendido de modo que as matrizes $\boldsymbol{X(n,k)}$ e $\boldsymbol{Y(n,m)}$ sejam decompostas simultaneamente em:

$$\boldsymbol{X} = \boldsymbol{T}\boldsymbol{P}^T + \boldsymbol{E} = \boldsymbol{t}_1\boldsymbol{p}_1^T + \boldsymbol{t}_2\boldsymbol{p}_2^T + \ldots + \boldsymbol{t}_a\boldsymbol{p}_a^T + \boldsymbol{E} \tag{2.47}$$

$$\boldsymbol{Y} + \boldsymbol{F} = \boldsymbol{U}\boldsymbol{Q}^T = \boldsymbol{u}_1\boldsymbol{q}_1^T + \boldsymbol{u}_2\boldsymbol{q}_2^T + \ldots + \boldsymbol{u}_a\boldsymbol{q}_a^T + \boldsymbol{F} \tag{2.48}$$

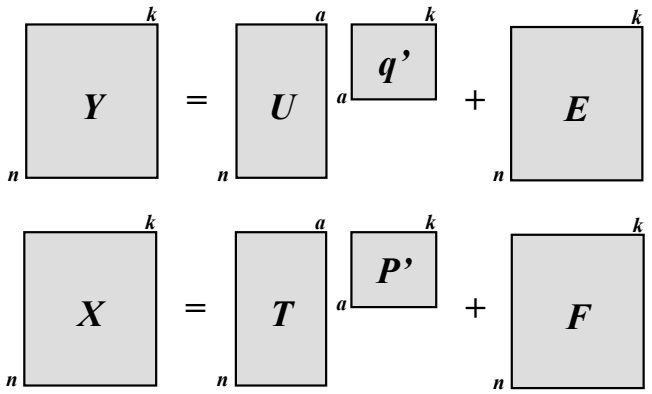

Figura 2.9: Representação gráfica das decomposições do método *PLS*.

Do mesmo modo que em uma análise do tipo *PCA*, a matriz $\boldsymbol{T}$ concentra a informação útil originalmente inserida nas colunas de $\boldsymbol{X}$. Similarmente, a matriz $\boldsymbol{U}$ concentra a informação útil originalmente contida nas colunas de $\boldsymbol{Y}$. Neste contexto, informação útil passa a ser aquela que nos permite construir o melhor modelo para a relação existente entre $\boldsymbol{X}$ e $\boldsymbol{Y}$. Na Tabela 2.4, é apresentado o algoritmo *NIPALS* modificado para conduzir a decomposição *PLS*. Este algoritmo seleciona o vetor peso $\boldsymbol{w}$ e o vetor peso $\boldsymbol{q}$ de modo que as direções $\boldsymbol{t}$ e $\boldsymbol{u}$ extraídas apresentem a melhor relação possível do ponto de vista de um modelo linear $\boldsymbol{u} = b\boldsymbol{t}$. No decorrer do algoritmo, é realizado o mapeamento da relação linear existente entre as direções $\boldsymbol{u_h}$ e $\boldsymbol{t_h}$ extraídas ($h = 1, 2, ..., a$) através do cálculo dos coeficientes $b_1, b_2, ..., b_a$.. Além dos coeficientes, os vetores $\boldsymbol{w_h}$, $\boldsymbol{p_h}$ e $\boldsymbol{q_h}$ são necessários para possibilitar a obtenção de futuras predições para o bloco as variáveis do bloco $\boldsymbol{Y}$ a partir de medidas das variáveis do bloco $\boldsymbol{X}$. Tais predições podem ser obtidas pelo seguinte algoritmo de retro-substituição:

1. Tratar a nova matriz $\boldsymbol{X}(n_2,k)$ do mesmo modo que o conjunto de treino foi tratado, subtraindo-se as médias previamente calculadas e multiplicando-se pelas constantes de escalonamento.

2. Ajustar os valores iniciais para a matriz das predições, $\hat{\boldsymbol{Y}}(n_2, m)$ como sendo nulos.

3. Para cada dimensão do modelo ($h=1,2,...a$) realizar os passos 4 a 7.

4. Calcular $\boldsymbol{t_h} = \boldsymbol{X}\boldsymbol{w_h}$

5. Calcular $\hat{\boldsymbol{u}}_h = b_h\boldsymbol{t_h}$

6. Atualizar os valores preditos somando-se a matriz $\hat{\boldsymbol{u}}_h\boldsymbol{q_h}^T$ à matriz $\hat{\boldsymbol{Y}}(n_2, m)$.

7. Formar os resíduos para a matriz das entradas: $\boldsymbol{X} = \boldsymbol{X}\text{-}\boldsymbol{t_h}\boldsymbol{p_h}^T$

Este procedimento permite que a informação contida em um conjunto de observações, denominado conjunto de treino, seja utilizada para futuras predições das saídas a partir do conhecimento das entradas. Entretanto, no caso linear, é possível que seja obtido um modelo explícito, em termos das variáveis originais da matriz $\boldsymbol{X}$. A dedução da expressão para a computação dos coeficientes do modelo linear em termos das variáveis originais em cada

etapa da decomposição é relativamente simples e pode ser encontrada em Kvalheim e Karstang (1989). Por uma questão de objetividade, a dedução de tal expressão não será demonstrada neste trabalho.

Tabela 2.4: Algoritmo *NIPALS* para *PLS*.

Passo	Sumário do Passo	Computação
0	Normalizar e centrar $\boldsymbol{X}$.	
1	Tomar a primeira coluna de $\boldsymbol{X}$ como chute inicial para $\boldsymbol{t_h}$ e a primeira coluna de $\boldsymbol{Y}$ como chute inicial para $\boldsymbol{u_h}$	
2	Calcular $\boldsymbol{w_h}$	$\boldsymbol{w_h}^T = \boldsymbol{u_h}^T \boldsymbol{X} / \boldsymbol{u_h}^T \boldsymbol{u_h}$
3	Normalizar $\boldsymbol{w_h}$	$\boldsymbol{w_h} = \boldsymbol{w_h} / \lVert \boldsymbol{w_h} \rVert$
4	Calcular $\boldsymbol{t_h}$	$\boldsymbol{t_h} = \boldsymbol{X w_h} / \boldsymbol{w_h}^T \boldsymbol{w_h}$
5	Calcular $\boldsymbol{q_h}$	$\boldsymbol{q_h}^T = \boldsymbol{t_h}^T \boldsymbol{Y} / \boldsymbol{t_h}^T \boldsymbol{t_h}$
6	Normalizar $\boldsymbol{q_h}$	$\boldsymbol{q_h} = \boldsymbol{q_h} / \lVert \boldsymbol{q_h} \rVert$
7	Comparar o valor de $\boldsymbol{t_h}$ obtido em 4 com o utilizado em 2. Se são iguais avançar para 6. Senão voltar para 2.	
8	Calcular $\boldsymbol{p_h}$	$\boldsymbol{p_h}^T = \boldsymbol{t_h}^T \boldsymbol{X} / \boldsymbol{t_h}^T \boldsymbol{t_h}$
9	Normalizar $\boldsymbol{p_h}$	$\boldsymbol{p_h} = \boldsymbol{p_h} / \lVert \boldsymbol{p_h} \rVert$
10	Mapear a relação entre $\boldsymbol{t_h}$ e $\boldsymbol{u_h}$ utilizando o modelo linear.	$b_h = \left(\boldsymbol{t_h}^T \boldsymbol{u_h} \right) / \left(\boldsymbol{t_h}^T \boldsymbol{t_h} \right)$
8	Calcular a matriz residual para as entradas.	$\boldsymbol{F} = \boldsymbol{X} - \boldsymbol{t_h} \boldsymbol{p_h}^T$
9	Calcular a matriz residual para as saídas.	$\boldsymbol{E} = \boldsymbol{Y} - b_h \boldsymbol{t_h} \boldsymbol{q_h}^T$
10	Se direções adicionais forem necessárias, substituir $\boldsymbol{X}$ por $\boldsymbol{F}$ e voltar para o passo 1.	

2.6.3. Algoritmo PLS não linear (QPLS)

Os métodos de redução de dimensionalidade apresentados até agora são capazes de superar problemas relacionados à colinearidade e também possibilitam a filtragem de ruído nas medidas. Entretanto, ao lidarmos com sistemas físicos e químicos complexos, o método *PLS* linear nem sempre é adequado para modelar a estrutura subjacente, que pode ser altamente não linear. Ao aplicarmos o método *PLS* linear para problemas não lineares, corremos o risco de descartar informação preditiva junto com as variáveis latentes de menor variabilidade. Isso ocorre porque a variabilidade dos blocos que não pode ser capturada pelo mapeamento linear acaba sendo confundida com os resíduos.

Conseqüentemente, extensões não lineares para o método *PLS* têm sido propostas. Segundo Wold et al. (1989), Gnanadesikan (1977) estudou a decomposição *PCA* de uma matriz $\boldsymbol{X}$ expandida com termos de segunda ordem e, posteriormente, alguns trabalhos foram desenvolvidos para estender as conclusões de Gnanadesikan para o contexto de modelagem por *PLS*. Para um número limitado de variáveis explicativas, esta abordagem pode ser adequada para modelar curvaturas nas relações. Entretanto, à medida que o número de variáveis independentes cresce, o número de termos de segunda ordem pode se tornar muito alto, o que dificulta as computações e a interpretação dos resultados, tornando a abordagem inapropriada. Em Höskuldson (1996) é feita uma discussão a respeito das dificuldades encontradas quando se utiliza esta abordagem para estender o modelo *PLS* para situações onde as saídas e as entradas apresentam relação não linear.

Para evitar os problemas causados pelo aumento exagerado das dimensões da matriz $\boldsymbol{X}$ em situações onde o número de variáveis explicativas é alto, Wold et al. (1989) desenvolveram um algoritmo de regressão não linear para o método *PLS* que mantém a estrutura do método original, incluindo a ortogonalidade das variáveis latentes explicativas $\boldsymbol{t}$*'s*. Ao invés de expandir o espaço da matriz $\boldsymbol{X}$ com termos de segunda ordem, a relação existente entre as direções extraídas do bloco $\boldsymbol{Y}$, $\boldsymbol{u}$*'s*, e as direções $\boldsymbol{X}$, $\boldsymbol{t}$*'s*, é mapeada por uma função não linear genérica:

$$\hat{\boldsymbol{u}}_h = f_h(\boldsymbol{t}_h) \qquad h = 1,2,...,a \tag{2.49}$$

O algoritmo desenvolvido pelos autores possibilita a utilização de qualquer função *f* para o mapeamento da relação não linear existente entre as direções extraídas, desde que a mesma seja contínua e diferenciável em relação aos pesos $\boldsymbol{w}$.

Esta abordagem requer algumas modificações no procedimento de determinação das estruturas latentes, $\boldsymbol{t}$*'s* e $\boldsymbol{u}$*'s*. O algoritmo original determina as direções que fornecem o melhor mapeamento da relação existente entre as direções do ponto de vista de um modelo linear. A utilização de uma função não linear deverá apresentar como ótimas outras combinações lineares das colunas de $\boldsymbol{X}$ e $\boldsymbol{Y}$. Para levar isto em conta, os autores propuseram um procedimento do tipo Newton-Raphson para a determinação dos pesos $\boldsymbol{w}$. Ou seja, partindo-se de uma aproximação inicial para as direções, obtido por *PLS* linear, o mapeamento não linear é expandido em série de Taylor e resolvido para os incrementos $\boldsymbol{\Delta w}$. Posteriormente, Baffi et al. (1999), apresentaram algumas contribuições teóricas para este procedimento de atualização dos vetores pesos. Na descrição de Baffi et al., expandindo f_h em série de Taylor e ignorando os termos de ordem superior a um, obtemos a aproximação dada pela Eq. 2.50.

$$\hat{\boldsymbol{u}}_h \approx f_h(\boldsymbol{X}, \boldsymbol{w}_0) + \sum_{j=1}^{k} \left. \frac{\partial \boldsymbol{f}_h}{\partial \boldsymbol{w}_{hj}} \right|_{w=w_h^0} \boldsymbol{\Delta w} \tag{2.50}$$

Tabela 2.5: Algoritmo *PLS* não linear.

Passo	Sumário do Passo	Computações
0	Normalizar $\boldsymbol{X}$ e $\boldsymbol{Y}$.	
1	Tomar a primeira coluna de $\boldsymbol{Y}$ como chute inicial para $\boldsymbol{u_h}$.	
2	Regressão das colunas de $\boldsymbol{X}$ em $\boldsymbol{u_h}$	$\boldsymbol{w^T = u^T X / u^T u}$
3	Normalizar $\boldsymbol{w_h}$	$\boldsymbol{w = w / \|w\|}$
4	Calcular $\boldsymbol{t_h}$	$\boldsymbol{t = Xw / w^T w}$
5	Realizar a regressão não linear.	$\boldsymbol{c \xleftarrow{fit} [u_h = f(t_h) + e]}$
6	Obter a predição $\boldsymbol{r_h}$ para $\boldsymbol{u_h}$	$\boldsymbol{r_h = f(t_h ; c_h)}$
7	Regressão das colunas de $\boldsymbol{Y}$ em $\boldsymbol{r_h}$	$\boldsymbol{q_h^T = r_h^T Y / r_h^T r_h}$
8	Normalizar $\boldsymbol{q_h}$	$\boldsymbol{q = q / \|q\|}$
9	Calcular $\boldsymbol{u_h}$	$\boldsymbol{u_h = Y q_h / q_h^T q_h}$
10	Atualizar $\boldsymbol{w_h}$ como pelo procedimento Newton-Raphson apresentado anteriormente.	
11	Normalizar $\boldsymbol{w_h}$	$\boldsymbol{w_h = w_h / \|w_h\|}$
12	Calcular o novo vetor $\boldsymbol{t_h}$	$\boldsymbol{t_h = X w_h / w_h^T w_h}$
13	Verificar a convergência em $\boldsymbol{t_h}$. Se a convergência foi alcançada, avançar para 14. Senão voltar para 5.	
14	Realizar a regressão não linear.	$\boldsymbol{c \xleftarrow{fit} [u_h = f(t_h) + e]}$
15	Obter a nova predição $\boldsymbol{r_h}$ para $\boldsymbol{u_h}$	$\boldsymbol{r_h = f(t_h ; c_h)}$
16	Calcular $\boldsymbol{p_h}$	$\boldsymbol{p_h = t_h X / t_h^T t_h}$
17	Calcular a matriz residual para as entradas	$\boldsymbol{F = X - t_h p_h^T}$
18	Calcular a matriz residual para as saídas	$\boldsymbol{E = Y - r_h q_h^T}$
19	Se novas dimensões forem necessárias, $\boldsymbol{X}$ e $\boldsymbol{Y}$ devem ser substituídos por $\boldsymbol{F}$ e $\boldsymbol{E}$ e os passos 1-19 são repetidos.	

A Eq. 2.50 pode ser rearranjada da seguinte forma:

$$\boldsymbol{e} = \hat{\boldsymbol{u}}_h - f_h(\boldsymbol{X}, \boldsymbol{w}_0) \approx \sum_{j=1}^{k} \left. \frac{\partial \boldsymbol{f}_h}{\partial \boldsymbol{w}_{hj}} \right|_{\boldsymbol{w}=\boldsymbol{w}_h^0} \boldsymbol{\Delta w} \tag{2.51}$$

O vetor $\boldsymbol{e}$ na Eq. 2.51 pode ser calculado facilmente. Do mesmo modo, podemos avaliar as derivadas da função *f* com relação aos pesos nos pontos. Portanto, a única incógnita na Eq. 2.51 é o vetor dos incrementos $\boldsymbol{\Delta w}$. Uma aproximação para $\boldsymbol{\Delta w}$ pode ser obtida pelo método dos mínimos quadrados. Uma vez obtida a aproximação, os incrementos podem ser adicionados aos respectivos pesos para fornecer a atualização dos seus valores. Este procedimento de atualização dos pesos é então incorporado ao algoritmo *NIPALS*, fornecendo o algoritmo *PLS* não linear apresentado na Tabela 2.5

Particularizando o caso genérico, Wold et al. (1989) desenvolveram o algoritmo *QPLS* (*Quadratic Partial Least Squares*), que utiliza como função de mapeamento um polinômio de segundo grau:

$$\hat{\boldsymbol{u}}_{\boldsymbol{h}} = c_{0_h} + c_{1_h} \boldsymbol{t}_{\boldsymbol{h}} + c_{2_h} \boldsymbol{t}_{\boldsymbol{h}}^{\boldsymbol{2}} \qquad h = 1,2,...,a \tag{2.52}$$

Quando utilizamos o método *PLS* não linear, não é possível obter uma expressão de $\boldsymbol{y}$ em termos das variáveis originais da matriz $\boldsymbol{X}$. As predições devem ser obtidas a partir de um algoritmo de retro-substituição, similar ao apresentado a seção 2.6.2, substituindo-se o vetor $\hat{\boldsymbol{u}}_{\boldsymbol{h}} = b_h \boldsymbol{t}_h$ por $\hat{\boldsymbol{u}}_{\boldsymbol{h}} = c_{0_h} + c_{1_h} \boldsymbol{t}_{\boldsymbol{h}} + c_{2_h} \boldsymbol{t}_{\boldsymbol{h}}^{\boldsymbol{2}}$.

2.6.4. PLS baseado em transformação Box-Tidwell (BTPLS)

O algoritmo de regressão apresentado na Tabela 2.5 é um ponto de partida para a utilização da metodologia *PLS* em problemas não lineares. Este algoritmo foi desenvolvido por Wold et al. (1989) e, posteriormente, aprimorado por Baffi et al. (1999). Nestes trabalhos, a função *f* utilizada para mapear a relação não linear existente entre $\boldsymbol{u}_{\boldsymbol{h}}$ e $\boldsymbol{t}_{\boldsymbol{h}}$ foi um polinômio de segundo grau. Obviamente, este tipo de mapeamento pode deixar a desejar em muitas situações. A escolha da função de mapeamento *f* é uma questão crítica, uma vez que, para um dado conjunto de dados, os diversos pares $\boldsymbol{u}_{\boldsymbol{h}}$ e $\boldsymbol{t}_{\boldsymbol{h}}$ podem apresentar relações consideravelmente diferentes. Isso significa que, no caso geral, a função utilizada para o mapeamento da relação entre os blocos $\boldsymbol{X}$ e $\boldsymbol{Y}$ deve ser modificada à medida que o algoritmo avança e as direções são extraídas.

Considerando esta questão, Li et al. (2001) propuseram o algoritmo *BTPLS* (*Box-Tidwell Based PLS*), que baseia-se no algoritmo apresentado na Tabela 2.5 e utiliza transformações de variáveis do tipo Box-Tidwell para o mapeamento da relação não linear existente entre as direções $\boldsymbol{t}_{\boldsymbol{h}}$ e $\boldsymbol{u}_{\boldsymbol{h}}$. A idéia básica por trás do procedimento de transformação proposto é a obtenção de uma família de modelos de regressão bastante flexíveis, de modo que diferentes tipos de relações entre as direções $\boldsymbol{u}_{\boldsymbol{h}}$ e $\boldsymbol{t}_{\boldsymbol{h}}$ possam ser mapeadas sem a necessidade da substituição da função *f* do algoritmo não linear.

Como o algoritmo de decomposição é exatamente igual ao descrito na seção anterior, vamos, nesta revisão, apresentar apenas a função não linear resultante da transformação de variáveis proposta e o procedimento sugerido para a estimação dos parâmetros da mesma. Basicamente, a transformação de variável proposta pelos autores, apresentada na Eq. 2.53, é uma extensão do método de Box e Tidwell (1962), portando ela foi chamada de transformação Box-Tidwell modificada.

$$\hat{\boldsymbol{u}}_{\boldsymbol{h}} = b_{0h} + b_{1h}\left[\mathrm{sgn}(\boldsymbol{t}_{\boldsymbol{h}})\right]^{\delta_h}\left|\boldsymbol{t}_{\boldsymbol{h}}\right|^{\alpha_h} \text{ para } \delta_h = 0 \text{ ou } \delta_h = 1 \tag{2.53}$$

Seguindo o procedimento genérico para a estimação dos parâmetros de modelos de regressão baseados em transformações das variáveis explicativas proposto por Box e Tidwell (1962), b_{0h}, b_{1h}, δ_h e α_h podem ser obtidas através da resolução dos seguintes problemas de otimização:

$$\hat{b}_{1h} = \underset{b_{0h}, b_{1h}, \delta_h}{\arg\min} \left\{ \sum_{i=1}^{n} \left[\boldsymbol{u}_{\boldsymbol{hi}} - \left\{ b_{0h} + b_{1h}\left[\mathrm{sgn}(\boldsymbol{t}_{\boldsymbol{hi}})\right]^{\delta_h}\left|\boldsymbol{t}_{\boldsymbol{hi}}\right| \right\} \right]^2 \right\} \tag{2.54}$$

$$\hat{\gamma} = \underset{b_{0h}, b_{1h}, \delta_h, \gamma}{\arg\min} \left\{ \sum_{i=1}^{n} \left[\boldsymbol{u}_{hi} - \left\{ b_{0h} + b_{1h}\left[\mathrm{sgn}(\boldsymbol{t}_{hi})\right]^{\delta_h}\left|\boldsymbol{t}_{hi}\right| + \gamma b_{1h}\left[\mathrm{sgn}(\boldsymbol{t}_{hi})\right]^{\delta_h}\left|\boldsymbol{t}_{hi}\right| \ln(\left|\boldsymbol{t}_{hi}\right|) \right\} \right]^2 \right\} \tag{2.55}$$

$$\alpha_h = \left[\frac{\hat{\gamma}}{2\hat{b}_{1h}} + 1 \right]^2 \tag{2.56}$$

$$\left[b_{0h}, b_{1h}, \delta_h\right] = \underset{b_{0h}, b_{1h}, \delta_h}{\arg\min} \left\{ \sum_{i=1}^{n} \left[\boldsymbol{u}_{hi} - \left\{ b_{0h} + b_{1h}\left[\mathrm{sgn}(\boldsymbol{t}_{hi})\right]^{\delta_h}\left|\boldsymbol{t}_{hi}\right|^{\alpha_h} \right\} \right] \right\}^2 \tag{2.57}$$

Estas otimizações são essencialmente quadráticas e podem ser resolvidas por mínimos quadrados, decompondo-as em dois subproblemas, um baseado em $\delta_h=0$ e outro baseado em $\delta_h=1$. Segundo os autores, a utilização dos resultados da primeira iteração já são satisfatórios.

Os gráficos da Figura 2.10 ilustram a flexibilidade dos modelos obtidos através das transformações Box-Tidwell modificada. No seu trabalho, Li et al., 2001 também propuseram a utilização do modelo da Eq. 2.58 ao invés do apresentado na Eq. 2.53:

$$\hat{\boldsymbol{u}}_h = b_{0h} + b_{1h}\boldsymbol{t}_h + b_{2h}\left[\mathrm{sgn}(\boldsymbol{t}_h)\right]^{\delta_h}\left|\boldsymbol{t}_h\right|^{\alpha_h} \text{ para } \delta_h = 0 \text{ ou } \delta_h = 1 \tag{2.58}$$

Às custas de um parâmetro extra, o modelo definido na Eq. 2.58 apresenta maior flexibilidade para mapear a relação existente entre as estruturas latentes extraídas das matrizes $\boldsymbol{X}$ e $\boldsymbol{Y}$ durante as diferentes etapas da decomposição realizada pelo algoritmo *PLS* não linear. Chamamos de *BTPLS(I)* o algoritmo implementado com a Eq. 2.53 e de *BTPLS(II)* o implementado com a Eq. 2.58.

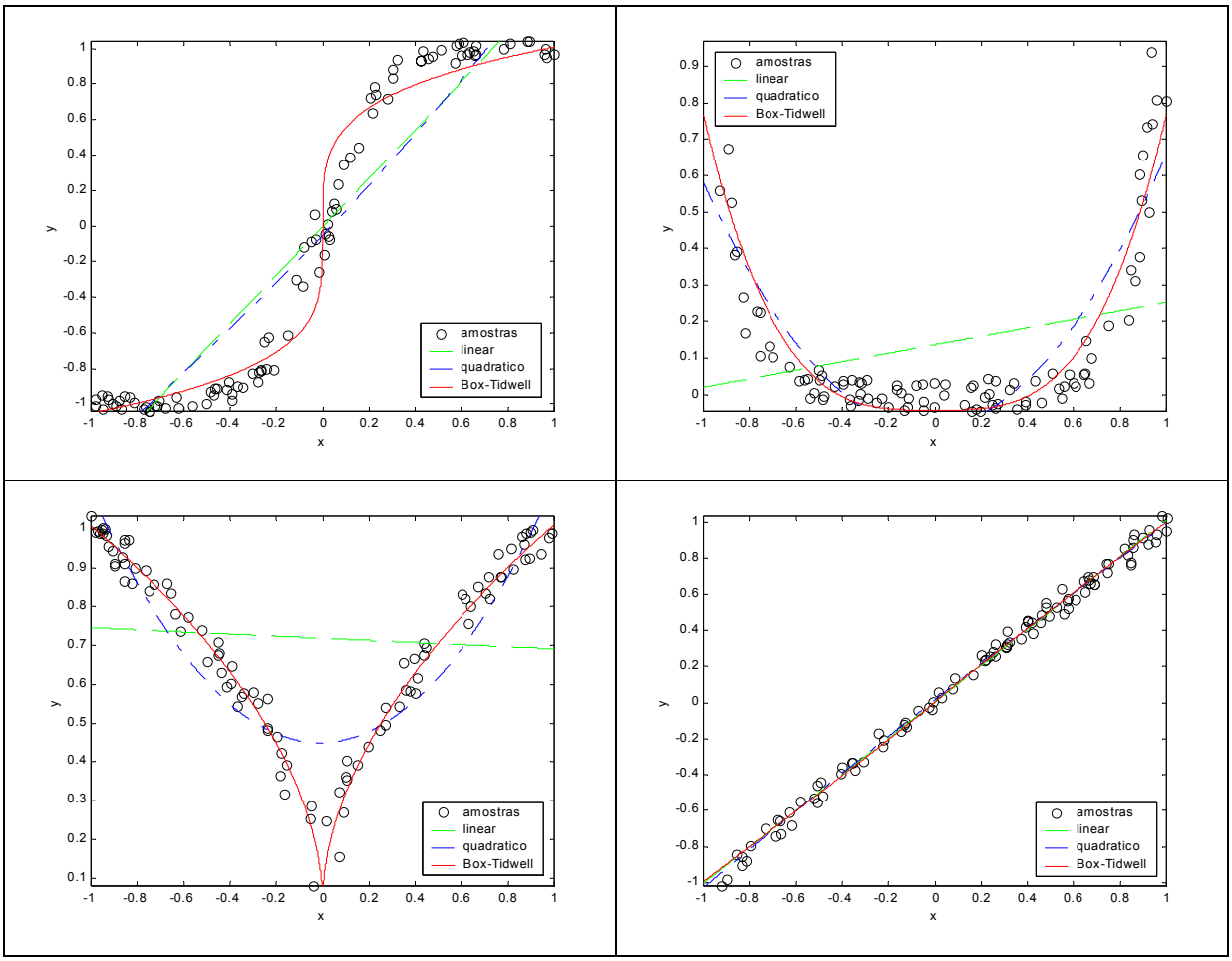

Figura 2.10: Ilustração dos modelos baseados na transformação Box-Tidwell modificada.

Em relação ao *BTPLS(I)*, a implementação do *BTPLS(II)* requer modificações nos procedimentos de estimação dos parâmetros e na atualização dos vetores pesos no algoritmo *PLS* não linear. Estas alterações não serão desenvolvidas nesta revisão. Dos dois métodos *BTPLS* apresentados, *BTPLS(I)* é preferível quando a simplicidade do modelo é mais importante que a precisão ou ainda quando o número de amostras é baixo em relação ao número de variáveis, causando risco de *overfit*. Nos outros casos, *BTPLS(II)* é a melhor alternativa.

Capítulo 3 Seleção de Variáveis em Regressão Multivariável

Nos capítulos anteriores, foram tratadas questões referentes à construção de modelos empíricos através de regressão multivariável, onde recorremos às medidas experimentais $\boldsymbol{X}$ e $\boldsymbol{y}$ para modelar a relação existente entre as entradas e a saída de um sistema. Uma questão de fundamental importância que surge quando modelos são construídos desta maneira é que, no caso geral, não sabemos de antemão se todas as variáveis explicativas, ou seja se todas as colunas presentes na matriz das entradas, devem ser consideradas. A utilização de muitos termos explicativos pode conduzir a modelos instáveis, caracterizados por mudanças drásticas nos parâmetros estimados frente à adição ou remoção de poucos pontos experimentais, ou ainda pelo fornecimento de resultados absurdos em extrapolações. Por outro lado, ao descartamos termos explicativos importantes, corremos o risco de obter um modelo impreciso, caracterizado por uma alta variância nas predições. Por isso, em muitas situações práticas, o problema de modelagem empírica está associado à questão da seleção de variáveis.

Normalmente, a questão da seleção de variáveis é abordada por métodos do tipo *stepwise*, que realizam o procedimento de modelagem em etapas. Em cada etapa, são realizados testes estatísticos para classificar as variáveis e uma única variável explicativa é selecionada para compor o modelo. Na etapa seguinte, todas as outras variáveis são novamente testadas e aquela que se mostrar mais adequada para descrever a variável de resposta é então escolhida. Este procedimento é repetido até que se julgue que nenhuma das variáveis restantes é adequada para compor o modelo. Estes métodos são muito populares e uma descrição mais detalhada a respeito dos mesmos pode ser encontrada em livros texto de estatística aplicada como, por exemplo, Werkema e Aguiar (1996). No capítulo 11.1 do livro de Höskuldsson (1996), é apresentado um procedimento *stepwise* mais elaborado, que trabalha com variáveis ortogonalizadas, o que o torna mais eficaz para lidar com situações onde as entradas apresentam correlação mútua. Shacham e Brauner (1999 b) incorporaram ao procedimento de regressão *stepwise* a utilização de indicadores que permitem avaliar a influência que a precisão experimental exerce nos resultados, possibilitando a identificação das causas que limitam a inclusão de novas variáveis ao modelo.

Os procedimentos *stepwise* mencionados anteriormente se baseiam no modelo de regressão linear múltipla (*MLR*) por mínimos quadrados. Quando trabalhamos com o modelo *MLR*, a questão da seleção de variáveis se torna importante não apenas pela identificação de variáveis não importantes, mas também pela identificação de variáveis linearmente dependentes, que desestabilizam a solução. No entanto, quando trabalhamos com métodos de redução de dimensionalidade, a questão da colinearidade não é um problema e, portanto, a abordagem do problema de seleção de variáveis por estes meios pode ser inadequada. Höskuldsson (1996) apresenta algumas alternativas para a seleção de variáveis na construção de modelos lineares de dimensão reduzida.

Neste capítulo, será feita uma breve revisão do procedimento, desenvolvido por Shacham e Brauner (1999 b) e, posteriormente, será proposto um novo método para a seleção de variáveis na construção de modelos empíricos baseado na capacidade preditiva do modelo. O método proposto pode ser aplicado a qualquer técnica de regressão, especialmente aos métodos de redução de dimensionalidade, lineares e não lineares, revisados no Capítulo 3 (*PCR*, *PLS*, *QPLS* e *BTPLS*). Na última seção, é conduzida uma comparação entre estes dois procedimentos.

3.1. Procedimento SROV

Nesta seção, é feita uma breve descrição a respeito do método *SROV*, desenvolvido por Shacham e Brauner (1999-B). O procedimento será apresentado brevemente, sintetizando as principais idéias introduzidas pelos autores, maiores detalhes a respeito desta técnica podem ser encontradas no trabalho original ou, ainda, em Shacham e Brauner (2003).

Basicamente, o procedimento é constituído de etapas sucessivas sendo que, em cada etapa uma das variáveis explicativas é escolhida para entrar no modelo. As variáveis explicativas que já foram incluídas no modelo (nas etapas anteriores) são chamadas de variáveis básicas enquanto as variáveis que ainda não foram selecionadas são chamadas de variáveis não básicas. Em cada etapa, as variáveis não básicas e a variável de resposta são primeiramente atualizadas, subtraindo-se a informação que é colinear às variáveis básicas. Esta atualização gera variáveis não básicas que são ortogonais em relação às variáveis do conjunto básico. Inicialmente, dispomos da matriz $\boldsymbol{X}(n,k)$, cujas colunas contém as n medidas para cada uma das k variáveis explicativas, e o vetor das saídas $\boldsymbol{y}(n,1)$, que contém as n medidas da variável de resposta. Na partida do algoritmo, o conjunto não-básico, contendo todas as variáveis explicativas, está cheio e o conjunto básico está vazio. O algoritmo então centra todas as variáveis, subtraindo do vetor $\boldsymbol{y}$ e das k colunas da matriz $\boldsymbol{X}$ as respectivas médias. Posteriormente, são calculados os valores para o coeficiente de correlação r_j entre cada uma das variáveis explicativas e a variável de resposta utilizando a Eq. 3.59. Também são calculados, para todas as variáveis do conjunto não-básico, os valores dos indicadores *TNR* (*truncation to noise ratio*) e *CNR* (*correlation to noise ratio*), utilizando-se, respectivamente, as equações 3.60 e 3.61.

O coeficiente de correlação r_j pode assumir valores entre 0 e 1, sendo que, quanto mais alto é o seu valor, mais forte é a associação linear existente entre $\boldsymbol{x_j}$ e a variável de resposta $\boldsymbol{y}$. O cálculo dos indicadores *TNR* e *CNR* se baseia na hipótese de que as variáveis de entrada $\boldsymbol{x_j}$ e

a variável de resposta $\boldsymbol{y}$ são medidas com erro experimental que, neste texto, será denotado, respectivamente, por $\boldsymbol{\delta x_j}$ e $\boldsymbol{\varepsilon}$. O indicador TNR_j é um valor representativo da validade da informação contida em uma variável explicativa $\boldsymbol{x_j}$, que consiste na divisão da variância das medidas desta variável pela variância do erro contido nas mesmas. O indicador CNR_j é um valor representativo da validade da informação contida em r_j, que consiste na divisão do produto $\boldsymbol{y' x_j}$ (covariância) pelo erro contido no mesmo, obtido a apartir da fórmula da propagação do erro. Sendo assim, o requisito mínimo para que a variável seja selecionada é que os indicarores apresentem valor maior que a unidade.

$$r_j = \frac{y^T x_j}{\left| y \right\| x_j \right|} \tag{3.59}$$

$$TNR_j = \left[\frac{\boldsymbol{x_j^T x_j}}{\boldsymbol{\delta x_j^T \delta x_j}} \right]^{\frac{1}{2}} = \frac{\left\| \boldsymbol{x_j} \right\|}{\left\| \boldsymbol{\delta x_j} \right\|} \tag{3.60}$$

$$CNR_j = \frac{\left| \boldsymbol{y^T x_j} \right|}{\sum_{i=1}^{n} \left(\left| x_{ij} \varepsilon_i \right| + \left| y \delta x_{ij} \right| \right)} \tag{3.61}$$

Satisfeitos os requisitos de $TNR > 1$ e $CNR > 1$, a variável que apresentar maior coeficiente de correlação com a resposta é então selecionada para ingressar no modelo, passando a fazer parte do conjunto básico. Chamando a variável escolhida de $\boldsymbol{x_p}$, a estimativa do parâmetro b_p correspondente é obtida por:

$$\boldsymbol{b_p} = \frac{\boldsymbol{y^T x_p}}{\boldsymbol{x_p^T x_p}} \tag{3.62}$$

Após selecionada a variável a ser adicionada à base, os valores da variável de resposta e das variáveis não-básicas devem ser atualizados segundo as equações 3.63 e 3.64:

$$\boldsymbol{y^{k+1}} = \boldsymbol{y^k} - \boldsymbol{b_p x_p} \tag{3.63}$$

$$\boldsymbol{x_j^{k+1}} = \boldsymbol{x_j^k} \quad \frac{\boldsymbol{x_j^{kT} x_p}}{\boldsymbol{x_j^{kT} x_j^k}} \ \boldsymbol{x_p} \tag{3.64}$$

O vetor $\boldsymbol{y^{k+1}}$ representa a variabilidade residual, que não pode ser explicada pelas variáveis incluídas na base até o estágio k. A variável $\boldsymbol{x^{k+1}}$ corresponde a $\boldsymbol{x^k}$ descontado da parcela linearmente dependente de $\boldsymbol{x_p}$. Fazemos isso porque, como $\boldsymbol{x_p}$ já foi incluído na base, qualquer informação colinear a esta variável é inútil para descrever a variabilidade residual $\boldsymbol{y^{k+1}}$. Antes de avançarmos para o próximo estágio, é conveniente que a significância estatística do coeficiente b_p seja verificada através de um teste t. Definindo-

se um nível de significância α, o intervalo de confiança db_p para o coeficiente b_p pode ser calculado pela Eq. 3.65:

$$db_p = t(\nu,\alpha)\sqrt{s^2\left(\boldsymbol{x_p}.\boldsymbol{x_p}\right)} \tag{3.65}$$

onde t é a distribuição de *student* com ν graus de liberdade e s é o erro padrão da estimativa, que, neste caso, pode ser aproximado por $\left(\boldsymbol{y}^{\boldsymbol{k+1}}\boldsymbol{y}^{\boldsymbol{k+1}}\right)/\nu$.

A significância estatística do parâmetro é comprovada se $db_p/|b_p|$ for menor que a unidade. Então, o algoritmo deve passar para o próximo estágio. Este procedimento é repetido até que, para todas as variáveis contidas no conjunto não básico, *CNR* ou *TNR* apresentem valor inferior a unidade. Neste ponto, atingimos o melhor modelo de regressão que pode ser obtido a partir dos dados disponíveis. Na publicação onde o procedimento *SROV* foi originalmente apresentado, Shacham e Brauner (1999 b) sugeriram que os indicadores do método *SROV* para as variáveis não incluídas na base fossem utilizados como ferramenta de diagnóstico, indentificando ações que podem ser tomadas para a obtenção de melhores modelos. Três casos típicos foram identificados pelos autores:

> Todas as variáveis fora da base apresentam $CNR_j < 1$. Neste caso, a inclusão de novos termos explicativos na base é impedido pelo nível do ruído. O modelo poderia ser melhorado pela aquisição de dados mais precisos para $\boldsymbol{y}$ e $\boldsymbol{X}$.
>
> Para algumas variáveis fora da base temos $CNR>1$ mas $TNR<1$. A redução acelerada do valor do indicador *TNR* frente a do *CNR* aponta para a presença de colinearidade entre as entradas. Neste caso, o aumento do intervalo de valores nos quais as variáveis explicativas foram determinadas ou da precisão de suas medidas poderia conduzir a melhores modelos pelo efeito de atenuação da colinearidade.
>
> Existem variáveis fora da base para as quais $CNR>1$ e $TNR>1$, mas o procedimento foi encerrado devido à falta de significância estatística do parâmetro b_p estimado. Neste caso, a precisão dos dados não é o fator crítico, uma vez que o nível de ruído ainda não foi alcançado. Esta situação indica problemas com a estrutura do modelo, como, por exemplo, a utilização de uma forma funcional inadequada ou a não consideração de variáveis explicativas importantes.

Conforme demonstrado nos três casos anteriores, os indicadores do método *SROV* são úteis para diagnosticar limitações no processo de construção de um modelo de regressão. Entretanto, a utilização destes indicadores requer, imprescindivelmente, disponibilidade de estimativas do nível de ruído presente nas variáveis. No primeiro estágio, o algoritmo usa as estimativas de $\boldsymbol{\varepsilon}$ e $\boldsymbol{\delta x_j}$ fornecidas pelo usuário. Nos estágios subseqüentes, são utilizadas atualizações destas estimativas obtidas através de perturbações numéricas. Para isso, o algoritmo executa duas regressões em paralelo, usando dois conjuntos de dados, o conjunto original e um conjunto perturbado. Deste modo, as estimativas para o nível de ruído presente nas variáveis transformadas podem ser obtidas, em qualquer iteração, através das diferenças entre os valores dos dois conjuntos de dados.

Por fim, cabe ressaltar que, as variáveis ortogonalizadas não são aquelas que são medidas ou manipuladas na prática, a utilização do modelo obtido em termos das mesmas pode não ser conveniente. Por este motivo, após a determinação das variáveis básicas, os parâmetros devem ser recalculados em termos das variáveis originais.

3.2. Procedimento Proposto

No procedimento *SROV*, apresentado na seção anterior, o problema da colinearidade entre as entradas é contornado impedindo que variáveis não básicas fortemente relacionadas às variáveis básicas ingressem no modelo. Como, no fim das contas, o procedimento *SROV* gera um modelo de regressão linear múltipla (*MLR*), obtido pelo método dos mínimos quadrados, esta abordagem se faz necessária para garantir a estabilidade dos parâmetros estimados. Entretanto, é importante ressaltar que variáveis não básicas descartadas por apresentarem uma alta correlação com as variáveis básicas podem conter informação útil para descrever o comportamento da variável de resposta. Sendo assim, existe o risco de que esta informação útil seja perdida quando impedimos que estas variáveis entrem no modelo. Por esse motivo, essa abordagem pode ser desvantajosa em situações onde temos muitas variáveis explicativas mutuamente relacionadas, pois, nestes casos, a quantidade de informação útil descartada pode passar a ser considerável. Nestas situações, é conveniente que o processo de modelagem seja conduzido por métodos de redução de dimensionalidade, como os métodos do tipo *PLS*, por exemplo.

Como já foi mencionado, estes métodos realizam uma decomposição da matriz de dados original, identificando variáveis latentes (combinações lineares das entradas originais) que concentram a maior parte da informação útil presente nos dados. Mesmo assim, a utilização de variáveis explicativas que não apresentem nenhuma relação com a variável de resposta pode acabar prejudicando a identificação das variáveis latentes e, por isso, uma ferramenta de seleção de variáveis também se faz necessária em processos de modelagem empírica através de métodos de redução de dimensionalidade. Nesta seção, é proposto um novo método de seleção de variáveis, o método *SRMP* (s*tepwise regression based on model predictions*), o qual permite que sejam escolhidas as variáveis explicativas na geração de modelos empíricos a partir de qualquer técnica de regressão, inclusive as técnicas de redução de dimensionalidade (lineares e não lineares) apresentadas no Capítulo 2.

Basicamente, o método *SRMP*, como qualquer procedimento *stepwise*, constrói o modelo de forma gradual, analisando a importância de cada uma das variáveis explicativas individualmente. Para apresentar o procedimento proposto, vamos, novamente, considerar a situação onde dispomos da matriz $\boldsymbol{X}(n,k)$, cujas colunas contém as n medidas para cada uma das k "canditadas" a variáveis explicativas, e o vetor das saídas $\boldsymbol{y}(n,1)$, que contém as n medidas da variável de resposta em questão. Do mesmo modo que no método *SROV*, as variáveis explicativas em questão são divididas em dois conjuntos: o conjunto básico e o conjunto não básico. Inicialmente, o conjunto básico, que contém as variáveis que irão compor o modelo final, está vazio enquanto o conjunto não básico, que contém as variáveis que não irão compor o modelo final, está cheio. Então, o procedimento é realizado em etapas sendo que, em cada etapa, uma variável passa do conjunto não básico para o conjunto básico. O critério para selecionar as variáveis nas diferentes etapas do procedimento é o efeito

exercido pelas mesmas na capacidade preditiva do modelo final. Na primeira etapa do procedimento, é gerado um modelo para cada uma das k variáveis explicativas presentes no conjunto não básico e a variável que fornecer o modelo com melhor capacidade preditiva é selecionada para ingressar na base. Na segunda etapa, são gerados k-1 novos modelos, combinando-se cada uma das variáveis presentes no conjunto não básico com a variável selecionada na etapa anterior e, novamente, a variável que fornecer o modelo com melhor capacidade preditiva é selecionada para ingressar na base. Note que, ao contrário do método *SROV*, os dados não devem ser ajustados (ortogonalizados) em relação às variáveis selecionadas para entrar na base. O procedimento é repetido até que todas as variáveis restantes no conjunto não básico forneçam um modelo com capacidade preditiva inferior a do modelo obtido na etapa anterior. A Figura 3.1 apresenta um fluxograma ilustrativo, que representa esquematicamente o procedimento *SRMP*.

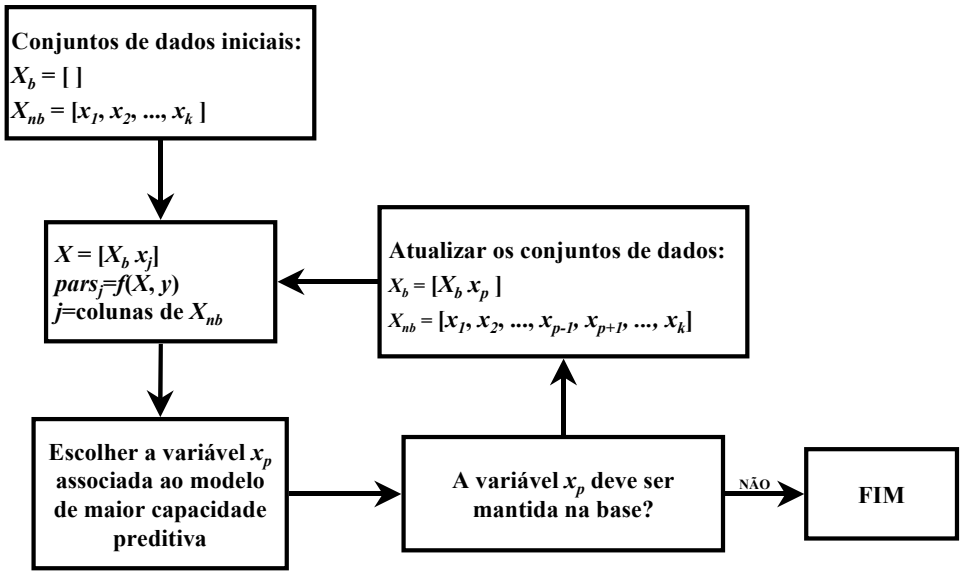

Figura 3.1: Fluxograma esquemático do procedimento *SRMP*.

Como podemos notar, além da técnica de regressão a ser utilizada na construção dos modelos, existem dois fatores que devem ser especificados na implementação do procedimento proposto. Estes dois fatores são a especificação de um índice para a medida da capacidade preditiva dos modelos obtidos em cada etapa do procedimento e a especificação de um critério para determinar em que momento a adição de variáveis à base deve ser encerrada. Na realidade, estes fatores podem ser especificados de diferentes maneiras. Isso significa que o algoritmo apresentado na Figura 3.1 é uma versão genérica do método proposto e que, portanto, versões específicas do procedimento *SRMP* podem ser obtidas através da modificação destes fatores.

Neste trabalho, os fatores em questão foram especificados de uma maneira especialmente útil quando o procedimento *SRMP* é utilizado para selecionar variáveis em modelos construídos através de técnicas de redução de dimensionalidade. A capacidade preditiva está relacionada com a predição do valor da variável de resposta para amostras diferentes das utilizadas na construção do modelo e pode ser medida por um teste de validação cruzada. O ponto de partida do teste que iremos utilizar é a separação das observações experimentais disponíveis em dois conjuntos, o conjunto de treino e o conjunto

de teste. A separação é feita escolhendo-se aleatoriamente 20% dos dados para compor o conjunto de teste. O modelo é então construído utilizando-se apenas as observações experimentais do conjunto de treino e as predições para o valor da variável de resposta referentes às amostras do conjunto de teste são computadas. Considerando um caso onde o conjunto de teste tenha n_2 observações, a soma dos quadrados dos desvios das predições do modelo para as observações do conjunto de teste em relação aos valores experimentais (*PRESS – predictive sum of squares*):

$$PRESS = \frac{\sum_{i=1}^{n_2} (y_i - \hat{y}_i)^2}{n_2} \tag{3.66}$$

é uma medida da capacidade preditiva do modelo, uma vez que as observações experimentais presentes no conjunto de teste não participaram da construção do mesmo.

A princípio, o valor da *PRESS* poderia ser usado diretamente como índice para classificar os modelos em cada etapa do procedimento *SRMP* quanto à capacidade preditiva. Se este fosse o caso, o critério de classificação, obviamente, seria: quanto menor o valor da *PRESS*, melhor a capacidade preditiva do modelo. Entretanto, como a computação da *PRESS* é uma variável aleatória, a análise de um valor isolado pode não ser representativa e, portanto, é mais conveniente trabalharmos com o resultado médio de, digamos, 100 computações desta soma, o que torna o índice avaliado mais confiável.

Para justificar o critério especificado para a determinação do momento em que a adição de variáveis à base deve ser encerrada, é apresentada a Figura 3.2, que ilustra o comportamento típico da *PRESS* em função do número de variáveis adicionadas à base pelo procedimento *SRMP* em casos onde existem muitas variáveis explicativas relacionadas e os modelos são construídos através de métodos de redução de dimensionalidade. Como pode ser observado, a *PRESS* do modelo final diminui rapidamente nas primeiras etapas do procedimento, isso acontece por que, nestes casos, é comum o fato de algumas poucas variáveis explicativas serem capazes de explicar a maior parte do comportamento da variável de resposta. Também é notável que, nas etapas seguintes, o valor da *PRESS* torna-se aproximadamente constante, formando um patamar na curva da Figura 3.2. A formação de tal patamar é devida ao fato de as variáveis adicionadas nestas etapas estarem fortemente relacionadas às variáveis incluídas na base nas etapas anteriores do procedimento. Por este motivo, o efeito que tais variáveis exercem na capacidade preditiva do modelo final não pode ser nitidamente visualizado. Tipicamente, os valores do patamar tendem a diminuir lentamente, mas eles também podem apresentar um comportamento levemente oscilatório, devido ao caráter randômico da computação da *PRESS* dos modelos pelo teste de validação cruzada. Por isso, se não for verificado um aumento significativo no valor da *PRESS*, não há motivos para que a variável seja impedida de ingressar à base. Por outro lado, se ao final de uma determinada etapa, for verificado que a *PRESS* teve o seu valor significativamente aumentado em relação a etapa anterior, não há motivos para continuar adicionando variáveis à base. Portanto, é muito importante que a *PRESS* do modelo obtido na etapa atual do procedimento *SRMP* seja comparada com a *PRESS* do modelo obtido na etapa anterior.

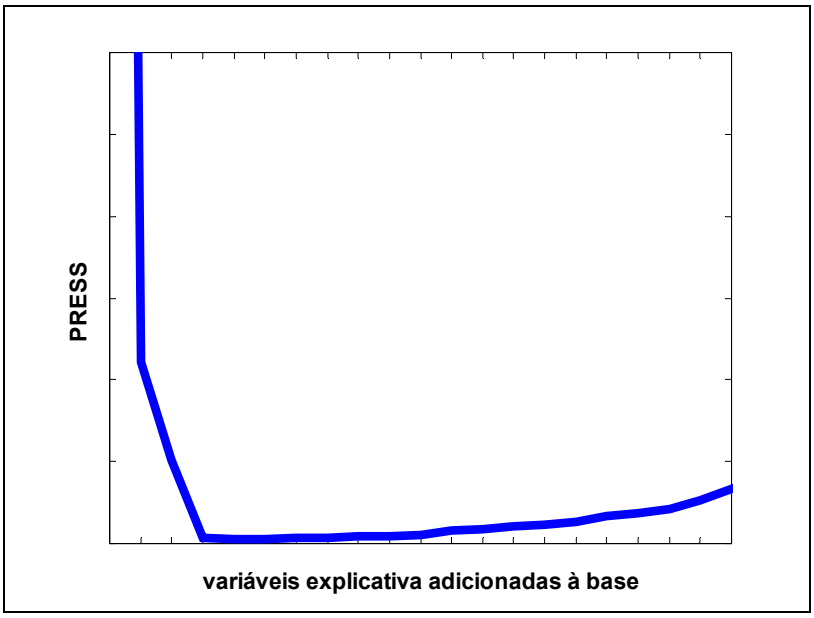

Figura 3.2: Comportamento típico da *PRESS*.

Como já foi mencionado, o teste de validação cruzada é repetido 100 vezes para a computação da *PRESS* média. Este número é razoavelmente grande e permite que a comparação entre os valores da *PRESS* dos modelos obtidos em duas etapas sucessivas do procedimento *SRMP* seja feita através uma análise estatística. Vamos então representar por m_i o valor referente à média das 100 computações da *PRESS* para o modelo obtido quando a variável $\boldsymbol{x_p}$ é selecionada para ingressar na base na etapa *i* do procedimento *SRMP*. Se as computações da *PRESS* na etapa *i* pertencerem a mesma "população" das computações da *PRESS* na etapa *i-1*, a diferença $D = m_i - m_{i-1}$ será uma variável aleatória que deverá apresentar valor esperado igual a zero. É fácil demonstrar que, se s_i é o desvio padrão das computações utilizadas no cálculo de m_i, a variável *D* apresenta variância $S^2 = (s_i^2 + s_{i-1}^2)/100$. Assumindo que *D* segue uma distribuição normal, a comparação desejada pode ser feita por um teste *t*. Neste caso, a significância da hipótese de *D* ser diferente de zero é dada pela expressão:

$$SIG = tcdf(D/S, 100) \tag{3.67}$$

ou seja, a função de distribuição *t* com 100 graus de liberdade integrada de menos infinito até *D/S*. Neste trabalho, vamos assumir que se *SIG* for maior que 0.99 a *PRESS* do modelo obtido na etapa atual do procedimento *SRMP* é significativamente maior do que a *PRESS* do modelo obtido na etapa anterior e que, portanto, a adição de variáveis à base deve ser encerrada.

3.3. Comparação entre os Procedimentos

Nesta seção, os métodos de seleção de variáveis *SROV* e *SRMP* serão comparados através de uma simulação computacional, utilizando um conjunto de dados artificial. O conjunto de dados gerado possui dez variáveis explicativas, $\boldsymbol{x_1}$, $\boldsymbol{x_2}$, ..., $\boldsymbol{x_{10}}$, e uma variável de resposta, $\boldsymbol{y}$. A variável $\boldsymbol{y}$ foi gerada como a soma de dois vetores independentes, $\boldsymbol{t_1}$ e $\boldsymbol{t_2}$, compostos por elementos aleatoriamente distribuídos entre zero e um. As variáveis $\boldsymbol{x_1}$, $\boldsymbol{x_2}$, $\boldsymbol{x_3}$ e $\boldsymbol{x_4}$ foram geradas como combinações lineares de $\boldsymbol{t_1}$ e $\boldsymbol{t_2}$, ao passo que as variáveis $\boldsymbol{x_5}$, $\boldsymbol{x_6}$, ..., $\boldsymbol{x_{10}}$

são constituídas por valores aleatoriamente distribuídos entre zero e um. A Tabela 3.1 apresenta as 25 observações "experimentais" que foram consideradas na simulação. Um ruído de magnitude equivalente a 0.1% foi adicionado aos dados para simular a presença de um pequeno erro experimental.

Vamos então, utilizando o conjunto de dados da Tabela 3.1, comparar o desempenho dos procedimentos *SROV* e *SRMP*. Obviamente, devido ao modo como os dados foram gerados, toda a informação necessária para a descrição do comportamento da variável de resposta está distribuída entre as variáveis $\boldsymbol{x_1}$, $\boldsymbol{x_2}$, $\boldsymbol{x_3}$ e $\boldsymbol{x_4}$. Os métodos serão comparados quanto a capacidade de identificar a importância de cada uma destas variáveis e também quanto a capacidade preditiva do modelo final fornecido.

Tabela 3.1: Conjunto de dados utilizado na comparação dos métodos *SROV* e *SRMP*.

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	y
1.260	0.394	0.651	0.481	0.275	0.067	0.235	0.179	0.701	0.690	1.634
0.714	0.263	0.393	0.288	0.439	0.676	0.369	0.627	0.680	0.499	0.911
0.910	0.442	0.562	0.409	0.406	0.274	0.807	0.167	0.985	0.553	1.106
0.796	0.274	0.427	0.314	0.227	0.402	0.138	0.638	0.579	0.595	1.023
1.190	0.441	0.655	0.481	0.351	0.670	0.061	0.962	0.108	0.506	1.513
0.764	0.457	0.520	0.377	0.158	0.816	0.186	0.803	0.873	0.627	0.889
0.483	0.193	0.274	0.201	0.832	0.811	0.679	0.717	0.982	0.676	0.606
0.537	0.313	0.361	0.262	0.269	0.074	0.989	0.308	0.857	0.299	0.627
1.387	0.544	0.781	0.573	0.065	0.888	0.466	0.691	0.367	0.855	1.745
0.587	0.237	0.334	0.245	0.625	0.514	0.574	0.862	0.240	0.053	0.737
0.735	0.127	0.321	0.240	0.564	0.586	0.254	0.210	0.763	0.002	1.005
0.691	0.242	0.372	0.275	0.877	0.917	0.515	0.569	0.414	0.444	0.887
1.147	0.526	0.689	0.504	0.275	0.084	0.962	0.903	0.622	0.499	1.406
0.739	0.200	0.364	0.270	0.747	0.949	0.476	0.913	0.676	0.446	0.972
0.531	0.220	0.305	0.224	0.900	0.427	0.647	0.493	0.714	0.101	0.662
0.977	0.380	0.549	0.403	0.802	0.179	0.030	0.623	0.352	0.610	1.234
0.403	0.049	0.165	0.125	0.247	0.991	0.005	0.530	0.491	0.951	0.559
0.551	0.110	0.249	0.186	0.337	0.539	0.914	0.280	0.979	0.336	0.746
1.217	0.478	0.686	0.503	0.833	0.862	0.744	0.670	0.142	0.271	1.533
0.280	0.039	0.116	0.088	0.468	0.055	0.681	0.694	0.017	0.023	0.385
0.986	0.490	0.615	0.448	0.632	0.817	0.640	0.151	0.694	0.497	1.194
0.538	0.312	0.361	0.262	0.315	0.886	0.668	0.036	0.014	0.560	0.630
0.902	0.189	0.415	0.308	0.296	0.792	0.330	0.136	0.293	0.217	1.216
0.646	0.151	0.304	0.227	0.070	0.937	0.912	0.708	0.196	0.853	0.861
0.724	0.322	0.429	0.314	0.273	0.907	0.091	0.020	0.235	0.373	0.895

Vamos iniciar analisando o desempenho do método *SROV*. Conforme foi detalhado na Seção 3.1, este método seleciona as variáveis explicativas com base nos valores dos coeficientes de correlação *r*, dos índices *TNR* e *CNR* e da razão $db/|b|$. Em cada etapa, satisfeitos os critérios $TNR>1$, $CNR>1$ e $db/|b|<1$, a variável de entrada que apresentar maior correlação com a saída é selecionada. A Tabela 3.2 apresenta o sumário dos resultados obtidos quando o método *SROV* é utilizado para modelar a relação existente entre a variável $\boldsymbol{y}$ e as variáveis $\boldsymbol{x_1}$, $\boldsymbol{x_2}$, ..., $\boldsymbol{x_{10}}$.

Tabela 3.2: Sumário dos resultados da construção do modelo pelo método *SROV*.

variável	etapa 1				etapa 2				etapa 3			
	r	TNR	CNR	dB	r	TNR	CNR	dB	r	TNR	CNR	dB
x_1	0.99	373	249	0.05	-	-	-	-	-	-	-	-
x_2	0.74	457	202	0.38	1.00	189	36.0	0.02	-	-	-	-
x_3	0.92	397	245	0.18	1.00	87.6	31.7	0.02	0.43	1.05	0.35	0.94
x_4	0.93	345	232	0.17	1.00	75.6	30.5	0.02	0.31	0.85	0.20	1.37
x_5	0.14	538	38.3	2.98	0.01	554	0.24	77.9	0.12	564	0.20	3.78
x_6	0.04	627	10.8	11.62	0.10	606	3.64	4.32	0.09	549	0.12	4.89
x_7	0.20	506	52.2	2.03	0.29	443	10.2	1.45	0.50	329	0.75	0.76
x_8	0.09	511	24.8	4.80	0.00	471	0.09	174	0.15	367	0.21	3.01
x_9	0.09	544	27.9	4.49	0.15	528	5.63	2.82	0.05	437	0.08	8.05
x_{10}	0.25	436	65.44	1.61	0.10	416	3.83	4.20	0.22	419	0.32	1.99

Na primeira etapa do procedimento, a variável que apresentou maior correlação com a resposta foi a variável $\boldsymbol{x_1}$ e, como os critérios requeridos foram todos satisfeitos, esta variável foi adicionada à base. Na segunda etapa, a variável $\boldsymbol{x_2}$ foi a que apresentou maior correlação com $\boldsymbol{y}$. Na terceira etapa, para todas as variáveis explicativas restantes, os índices CNR_j e db_j/b_j demonstram nitidamente que o nível de ruído foi alcançado e que as correlações verificadas não são significativas.

Os valores preditos pelo modelo obtido na segunda etapas do procedimento *SROV* são plotados contra os valores "experimentais" na Figura 3.3Também são apresentados os valores do coeficiente de correlação e da *PRESS* para o modelo final. Do mesmo modo que foi definido na seção anterior, o valor do somatório *PRESS* é calculado com base em um conjunto de teste composto 20% das observações (aqui também utilizamos a média de 100 computações como valor representativo da soma).

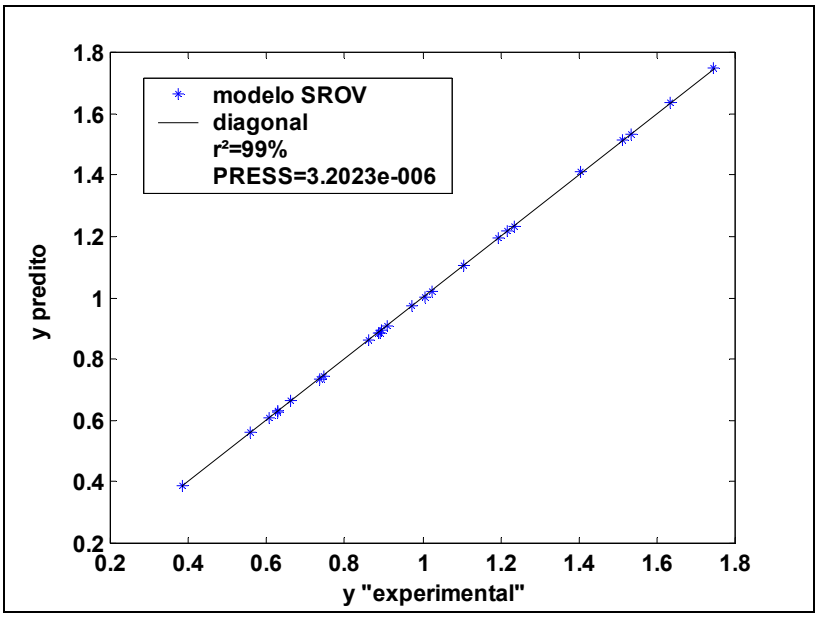

Figura 3.3: Avaliação do modelo obtido na segunda etapa do método *SROV*.

A Tabela 3.3 apresenta o sumário dos resultados da construção do modelo pelo método *SRMP*. Na primeira etapa, a variável x_1 foi a que forneceu o modelo com melhor capacidade preditiva (*PRESS* = 0.00199) e, portanto, foi adicionada à base. Na segunda etapa, a variável x_2 apresentou o modelo com melhor capacidade preditiva e, consequentemente, foi adicionada à base. Na terceira etapa, a variável x_3 foi incluída no modelo e assim por diante. Na etapa 5, quando a variável x_5 foi incorporada ao conjunto básico, a capacidade preditiva do modelo foi consideravelmente prejudicada. Por isso, o procedimento deve ser encerrado e o modelo final deve ser composto apenas pelas variáveis x_1, x_2, x_3 e x_4.

Tabela 3.3: Sumário dos resultados da construção do modelo pelo método *SRMP*.

variável	1	2	3	4	5	9	6	7	10	8
PRESS X 10^4	19.9	0.03	0.03	0.03	49.4	106	163	189	235	247
significância [%]	0.00	0.00	0.4	89.8	100	100	99.7	86.5	97.1	68.0

Na Figura 3.4 podemos visualizar o comportamento da soma de quadrados preditiva (*PRESS*) a medida que as variáveis vão sendo incluídas na base. Podemos observar que este exemplo retrata o caso típico previsto na Figura 3.2, onde a *PRESS* diminui rapidamente nas primeiras etapas do procedimento, atinge um valor aproximadamente estável nas etapas intermediárias e passa a crescer nas etapas finais.

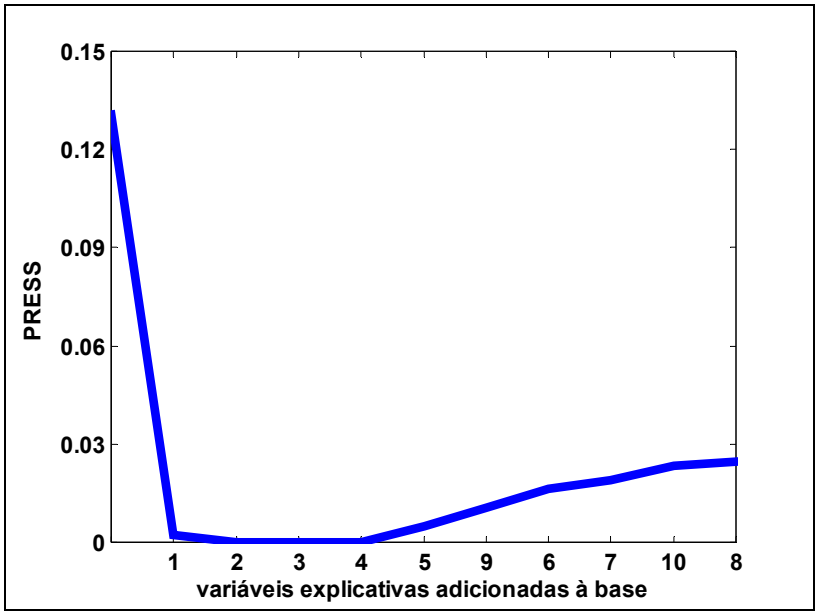

Figura 3.4 Valor da *PRESS* em função das variáveis presentes no modelo.

Os valores preditos pelo modelo obtido na quarta etapa do procedimento *SRMP* são plotados contra os valores "experimentais" na Figura 3.2. Também são apresentados os valores do coeficiente de correlação e da soma de quadrados preditiva (*PRESS*) para o modelo final obtido. Como pode ser observado, comparando os resultados com os obtidos pelo método *SROV*, o método proposto é capaz de identificar corretamente as quatro variáveis importantes para a descrição do comportamento da variável de resposta. Quanto ao ajuste do modelo aos dados "experimentais", não foram verificadas mudanças consideráveis.

Entretanto, no que diz respeito a capacidade preditiva, foi notado um decréscimo significativo, de aproximadamente 15%, na *PRESS*. Ao que tudo indica, esta melhora na capacidade preditiva está associada à utilização do método *PLS*. Conforme mencionado anteriormente, toda a informação necessária para a descrição do comportamento da variável de resposta está distribuída entre as variáveis $\boldsymbol{x_1}$, $\boldsymbol{x_2}$, $\boldsymbol{x_3}$ e $\boldsymbol{x_4}$. Como o modelo *PLS* é capaz de utilizar a informação presente nestas quatro variáveis de entrada, ele é capaz de fornecer predições mais precisas para o valor de $\boldsymbol{y}$ do que um modelo do tipo *MLR*, baseado apenas em $\boldsymbol{x_1}$ e $\boldsymbol{x_2}$.

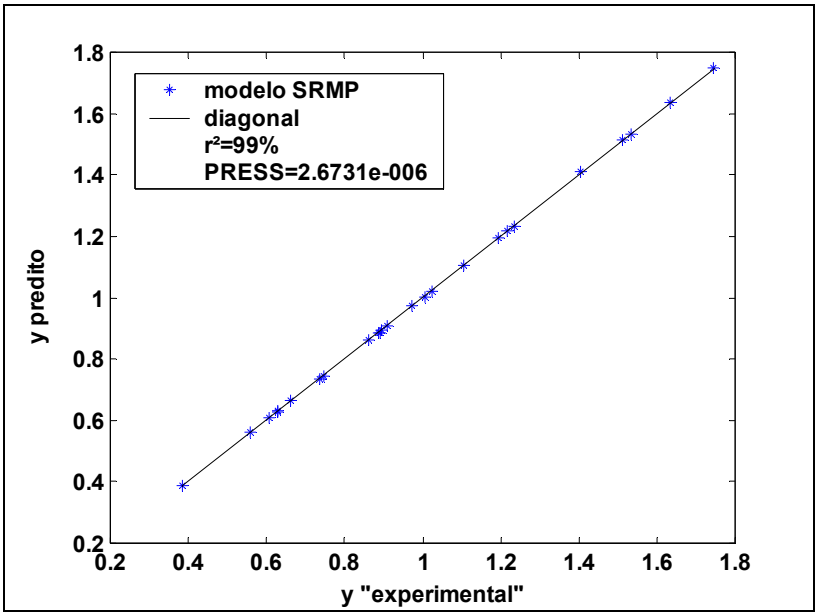

Figura 3.5: Avaliação do modelo obtido na quarta etapa do método *SRMP*.

É interessante notarmos que, segundo Shacham e Brauner (1999 b), o diagnóstico fornecido pelo método *SROV* deve ser capaz de identificar as situações onde a adição de variáveis à base foi impedida pela presença de colinearidade entre as entradas. Sendo assim, sempre que tal situação for verificada, a utilização do método *SRMP* é uma alternativa a ser considerada pois, por permitir que os parâmetros sejam estimados por métodos de redução de dimensionalidade, o mesmo pode conduzir a modelos com maior capacidade preditiva. Relembrando, os autores do método *SROV* sugerem que, se algumas variáveis estão sendo impedidas de entrar na base devido aos efeitos da colinearidade, estas deverão apresentar $TNR<1$ e $CNR>1$. Porém, como podemos observar na Tabela 3.2, embora seja exatamente este o caso, esta situação não foi verificada na comparação conduzida nesta seção, onde tanto o índice *TNR* como o índice *CNR* apresentaram valores inferiores ou muito próximos à unidade para as variáveis $\boldsymbol{x_3}$ e $\boldsymbol{x_4}$. Provavelmente, isso aconteceu porque, como a maior parte do comportamento da variável de resposta foi explicada nas etapas um e dois do procedimento, a variabilidade residual de $\boldsymbol{y}$ se tornou pequena de mais para que o índice *CNR* a diferenciasse do ruído. Entretanto, o fato de algumas variáveis não incluídas na base serem fortemente relacionadas com a selecionada é revelado de forma nítida se analisarmos apenas o comportamento do índice *TNR*. Esta afirmativa pode ser verificada pela análise da Figura 3.6, que mostra o comportamento deste índice para algumas das variáveis descartadas pelo procedimento *SROV* ($\boldsymbol{x_3}$, $\boldsymbol{x_4}$, $\boldsymbol{x_6}$ e $\boldsymbol{x_7}$,).

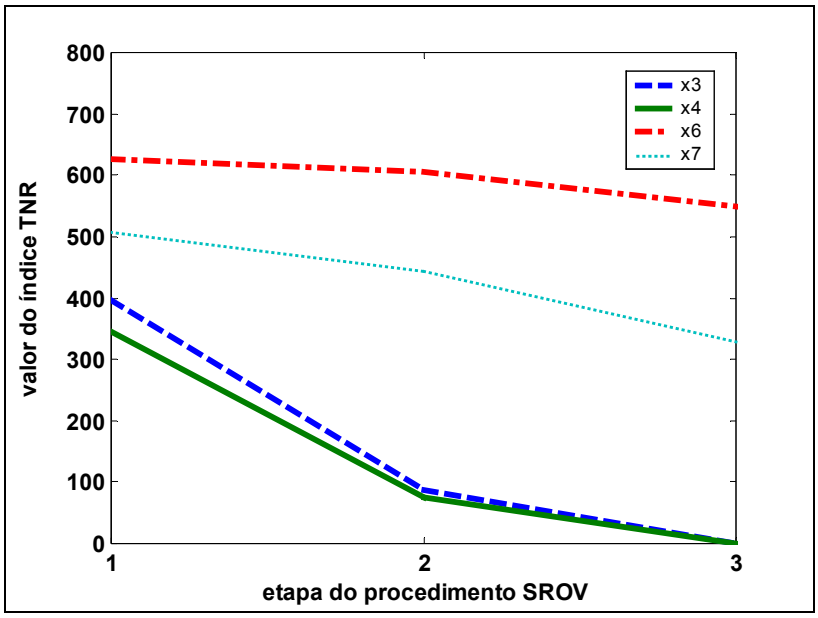

Figura 3.6: Análise do índice *TNR* para das variáveis descartadas pelo método *SROV*.

Como podemos observar na Figura 3.6 ou na Tabela 3.2, quando as variáveis $\boldsymbol{x_1}$ ou $\boldsymbol{x_2}$ são adicionadas à base, o valor do índice *TNR* para as variáveis $\boldsymbol{x_3}$ e $\boldsymbol{x_4}$ diminui pelo menos uma ordem de grandeza, ao passo que o valor do índice *TNR* para qualquer uma das outras variáveis não diminui mais do que 30%. A princípio, como o ruído presente nas variáveis selecionadas para ingressar no modelo se propaga às demais pelo procedimento de ortogonalização, é esperado que o valor do índice *TNR* sofra reduções a medida que o procedimento *SROV* avança. Entretanto, reduções drásticas do índice *TNR* como as verificadas para as variáveis $\boldsymbol{x_3}$ e $\boldsymbol{x_4}$ são típicas de casos onde temos problemas de forte correlação entre as entradas. Portanto, sempre que o procedimento *SROV* for utilizado para selecionar as variáveis a serem utilizadas em um modelo, é conveniente que o comportamento dos índices *TNR* seja monitorado. Se for verificado que, para algumas variáveis, o valor deste índice sofre mudanças drásticas quando o procedimento avança de uma etapa para a outra, a utilização do método *SRMP* deve ser considerada como uma alternativa concreta para a obtenção de modelos com maior capacidade preditiva.

Por fim, cabe ressaltar que o procedimento *SROV*, além de ser sensivelmente mais rápido quando o número de variáveis a serem avaliadas é alto, também apresenta a vantagem de fornecer um diagnóstico da regressão. Desta forma, sempre que soubermos de antemão que o modelo não apresenta muitas entradas fortemente correlacionadas e que o método *MLR* é adequado para a estimação dos parâmetros, este procedimento é uma boa alternativa para a questão da seleção de variáveis. Em casos mais complicados, onde o modelo obtido pela técnica *MLR* não é adequado, o procedimento *SRMP* parece ser mais indicado. Neste capítulo, foi demonstrado que o procedimento proposto é vantajoso em casos onde as entradas são altamente relacionadas e o método *PLS* linear é utilizado na estimação dos parâmetros. Mas, como já foi mencionado anteriormente, o método *SRMP* pode ser implementado utilizando-se qualquer técnica de regressão multivariável. Se forem utilizados métodos diferentes dos avaliados neste estudo, a especificação dos critérios talvez devam ser revistas, o que fica de sugestão para trabalhos futuros.

Capítulo 4 Estimação de Incertezas em Regressão Multivariável

Em situações reais, a quantificação de variáveis está sempre sujeita às incertezas dos erros experimentais. Quando um procedimento de regressão é realizado, as incertezas presentes na medidas das variáveis envolvidas se propagam aos parâmetros do modelo. Uma conseqüência direta deste fato é que as predições do modelo também estarão associadas à incertezas. Na realidade, quando utilizarmos o modelo para predizer o valor da variável de resposta de uma única amostra, cujos valores das variáveis explicativas são dadas pelas colunas do vetor $\boldsymbol{x_i}$, existirão duas fontes de incertezas: as incertezas presentes nas medidas da amostra $\boldsymbol{x_i}$ (erros futuros) e as incertezas presentes nos parâmetros do modelo (erros passados). Para que a confiabilidade das predições possa ser avaliada, ambas as fontes de erros devem ser quantitativamente determinadas, o que, estatisticamente, corresponde à computação da variância σ^2 ou do desvio padrão σ das mesmas.

Os erros associados a uma determinação experimental isolada do vetor $\boldsymbol{x_i}$ normalmente são avaliados com base em informações a respeito dos métodos ou instrumentos utilizados na quantificação das variáveis explicativas. Se tais informações não estiverem disponíveis, os mesmos podem ser quantificados computando-se as variâncias de réplicas da determinação experimental das variáveis. Analogamente, as incertezas presentes nos parâmetros do modelo podem ser avaliadas através de repetições do procedimento de regressão. O procedimento de regressão pode ser repetido R vezes, resultando em R réplicas das determinações experimentais dos valores das variáveis de entrada e saída $\boldsymbol{X^1y^1}$, $\boldsymbol{X^2y^2}$, ..., $\boldsymbol{X^Ry^R}$ para as n amostras. As R estimativas $\boldsymbol{b^1}$, $\boldsymbol{b^2}$, ..., $\boldsymbol{b^R}$ para o vetor dos parâmetros do modelo obtidas a partir das réplicas experimentais apresentam variância dada pela Eq. 4.68.

$$\boldsymbol{S^2(b)} = \frac{1}{R-1}\sum_{r=1}^{R}\left(\boldsymbol{b}^r - \boldsymbol{\beta}\right)\left(\boldsymbol{b}^r - \boldsymbol{\beta}\right)^T \tag{4.68}$$

onde $\boldsymbol{\beta}$ é o vetor com os valores esperados para os coeficientes do modelo. Naturalmente, o desvio padrão $S(b_j)$ de cada um dos j coeficientes é dado pela raiz quadrada dos elementos da diagonal da matriz $\boldsymbol{S^2(b)}$:

$$S(b_j) = \left[\boldsymbol{S}^2(\boldsymbol{b})_{jj}\right]^{1/2} \tag{4.69}$$

Na prática, como a realização de experimentos está associada a custos, a estimativa de incerteza nos coeficientes de regressão a partir de repetições do procedimento de modelagem pode não ser adequada. Portanto, outra ferramenta que permita a determinação dos erros nos coeficientes de regressão se faz necessária. No Capítulo 2, foram apresentadas juntamente com os demais conceitos fundamentais, expressões que permitem a determinação da matriz $\boldsymbol{S^2(b)}$ associada às estimativas de parâmetros em regressão linear múltipla. Tais expressões se limitam ao caso onde os parâmetros são estimados pelo método dos mínimos quadrados e, portanto, não são úteis quando a construção dos modelos é conduzida por métodos de redução de dimensionalidade. Nestes casos, especialmente nas situações onde o mapeamento da relação existente entre as variáveis latentes dos blocos $\boldsymbol{X}$ e $\boldsymbol{Y}$ é não linear, a obtenção de expressões analíticas para a determinação dos intervalos de confiança das computações pode vir a ser uma questão um tanto quanto complicada. Mesmo assim, são encontrados na literatura alguns estudos que visam a obtenção das mesmas. Como exemplo, podem ser citados os trabalhos de Pathak e Penlidis (1993) e Baffi et al (2002). Em ambos os casos, as expressões apresentadas são aproximações obtidas pela linearização dos estimadores *PLS* dos parâmetros por série de Taylor. Outra possível abordagem para esta questão é a utilização de técnicas de reamostragem, como os métodos *jackknife* e *bootstrap*, por exemplo. As técnicas de reamostragem têm sido utilizadas para a avaliação de incertezas em modelos obtidos por métodos de redução de dimensionalidade. Wold et al (1984), por exemplo, utilizaram o método *jackknife* para avaliar o erro presente na estimativa dos parâmetros do modelo *PLS* linear. Posteriormente, Wold et al (1989), ao propor uma extensão não linear para o algoritmo *NIPALS*, sugeriram que o mesmo método poderia ser utilizado para a análise estatística do modelo. No capítulo seis do livro de Höskuldsson (1996), a reamostragem é apresentada como ferramenta para validação de modelos e obtenção de intervalos de confiança. Ainda recentemente, são realizados estudos deste tipo. Duschesne e MacGregor (2001) aplicaram as técnicas *jackknife* e *bootstrap* no estudo de identificação de processos dinâmicos com modelos do tipo *FIR* (*finite impulse response*) e *ARX* (*auto-regressive with exogenous inputs*), utilizando *PLS* para a estimação dos parâmetros. Nicolaas e Faber (2002) conduziram um completo estudo comparando diferentes metodologias de reamostragem para a determinação de incertezas na estimação de coeficientes de regressão multivariável. Os autores avaliaram quatro metodologias que foram divididas em duas categorias, a categoria dos métodos que trabalham com os objetos (*jackknife* e *bootstrapping objects*) e a categoria dos métodos que trabalham com os resíduos (*bootstrapping residuals* e *noise addition*).

Neste capítulo, será proposta uma nova metodologia de reamostragem que, se baseando em informações a respeito do erro de medida das variáveis, é capaz de fornecer estimativas para a incerteza dos coeficientes em regressão multivariável. Retomando a divisão estabelecida anteriormente, a técnica proposta constituirá uma terceira categoria, a categoria dos métodos que trabalham explicitamente com o erro experimental. Nas próximas três seções, os métodos de cada uma das categorias serão apresentados. Posteriormente, a metodologia proposta será comparada com as demais através de simulações computacionais.

4.1. Reamostragem baseada nos objetos

Os métodos baseados nos objetos realizam a reamostragem trabalhando com as observações experimentais, ou seja, com as linhas da matriz $\boldsymbol{X}$ e com os elementos correspondentes do vetor $\boldsymbol{y}$. Nesta seção, serão descritos os dois métodos desta categoria considerados neste trabalho, o método *jackknife* e o método *bootstrapping objects*.

4.1.1. Método Jackknife

O método *jackknife* gera conjuntos de dados reduzidos pela remoção de amostras (observações experimentais) do conjunto de dados original. Chamando de $\boldsymbol{x}$*'s* os vetores correspondentes às linhas da matriz $\boldsymbol{X}$, o conjunto de dados reduzido resultante da remoção da observação i pode ser expresso por:

$$\begin{aligned} \boldsymbol{y}^{-i} &= \begin{bmatrix} \boldsymbol{y_1} & \cdots & \boldsymbol{y_{i-1}} & \boldsymbol{y_{i+1}} & \cdots & \boldsymbol{y_n} \end{bmatrix}^T \\ \boldsymbol{X}^{-i} &= \begin{bmatrix} \boldsymbol{x_1^T} & \cdots & \boldsymbol{x_{i-1}^T} & \boldsymbol{x_{i+1}^T} & \cdots & \boldsymbol{x_n^T} \end{bmatrix}^T \end{aligned} \tag{4.70}$$

Cada um dos conjuntos de dados reduzidos permite a obtenção da estimativa $\boldsymbol{b^{-i}}$ para os coeficientes do modelo. Combinando estas estimativas com a estimativa $\boldsymbol{b}$, obtida a partir do conjunto de dados completo, são computados os chamados pseudo-valores:

$$\boldsymbol{b}^{\boldsymbol{i}}_{\boldsymbol{pseudo}} = n\boldsymbol{b} - (n-1)\boldsymbol{b}^{-\boldsymbol{i}} \tag{4.71}$$

Denotando a média dos pseudo-valores por $\bar{\boldsymbol{b}}$, a aproximação para a matriz de covariância das estimativas é dada pela Eq. 4.72:

$$\boldsymbol{S^2(b)} = \frac{1}{n(n-1)} \sum_{i=1}^{n} \left(\boldsymbol{b}^{\boldsymbol{i}}_{\boldsymbol{pseudo}} - \bar{\boldsymbol{b}}\right)\left(\boldsymbol{b}^{\boldsymbol{i}}_{\boldsymbol{pseudo}} - \bar{\boldsymbol{b}}\right)^T \tag{4.72}$$

É importante notar que este procedimento não faz nenhuma consideração a respeito do erro presente em $\boldsymbol{X}$ ou $\boldsymbol{y}$. Na verdade, é assumido que as linhas reamostradas são uma amostra aleatória de alguma distribuição multivariável, o que implica no fato de que não devem ser utilizadas técnicas de planejamento de experimento para a obtenção dos dados.

4.1.2. Bootstrapping Objects

No método *botstrapping objects*, o conjunto de dados reamostrado é constituído por uma matriz com as mesmas dimensões de $\boldsymbol{X}(n,k)$ e por um vetor com as mesmas dimensões de $\boldsymbol{y}(n,1)$. A matriz e o vetor que constituem o conjunto reamostrado são construídos em n passos, sendo que em cada passo uma das linhas de $[\boldsymbol{X}\ \boldsymbol{y}]$ é selecionada para compor o conjunto reamostrado, que será composto somente por linhas presentes no conjunto original, porém algumas linhas aparecem mais de uma vez enquanto outras acabam sendo descartadas. O procedimento de reamostragem deve ser repetido um número R de vezes, gerando os novos conjuntos $[\boldsymbol{X^1}\ \boldsymbol{y^1}]$, $[\boldsymbol{X^2}\ \boldsymbol{y^2}]$, ..., $[\boldsymbol{X^R}\ \boldsymbol{y^R}]$, conforme descreve formalmente a Eq. 4.73.

$$\begin{aligned} \boldsymbol{y}_i^r &= y_{\xi_i^r} \\ \boldsymbol{X}_i^r &= \boldsymbol{X}_{\xi_i^r} \\ i &= 1,2,...,n \\ r &= 1,2,...,R \end{aligned} \tag{4.73}$$

Na Eq. 4.73, o subscrito *i* representa cada uma das *n* linhas de cada um dos *R* conjuntos reamostrados, diferenciados pelo subscrito *r*, enquanto ξ, um número aleatoriamente escolhido entre 1 e *n*, determina qual linha de [$\boldsymbol{X}$ $\boldsymbol{y}$] ocupa a *i*ésima linha do conjunto reamostrado *r*.

O número de conjuntos reamostrados *R* deve ser grande o bastante para assegurar a precisão do desvio padrão computado. Para cada um dos novos *R* conjuntos de dados reamostrados, é calculado o vetor dos coeficientes $\boldsymbol{b}^r$. Denotando a média dos vetores $\boldsymbol{b}^r$*'s* obtidos por $\overline{\boldsymbol{b}}$, a aproximação para a matriz de covariância das estimativas é dada pela Eq. 4.74:

$$\boldsymbol{S}^2(\boldsymbol{b}) = \frac{1}{R-1}\sum_{r=1}^{R}\left(\boldsymbol{b}^r - \overline{\boldsymbol{b}}\right)\left(\boldsymbol{b}^r - \overline{\boldsymbol{b}}\right)^T \tag{4.74}$$

O método bootstrapping objects é similar ao método *jackknife* no sentido de que não faz considerações a respeito do ruído presente nos valores de $\boldsymbol{X}$ ou $\boldsymbol{y}$, portanto as linhas reamostradas também devem ser uma amostra aleatória de alguma distribuição multivariável e não resultados de experimentos planejados.

4.2. Reamostragem Baseada nos Resíduos

Os métodos baseados nos resíduos realizam a reamostragem trabalhando com os desvios das predições do modelo em relação às observações experimentais. Nesta seção, serão descritos os dois métodos desta categoria considerados neste trabalho, o método *bootstrapping residuals* e o método da adição de resíduos.

4.2.1. Bootstrapping Residuals

O ponto de partida do método *bootstrapping residuals* é o próprio modelo de regressão do qual a variabilidade dos coeficientes deve ser estimada. Primeiramente, os resíduos são calculados de acordo com a Eq. 4.75:

$$e_i = \frac{y_i - \hat{y}_i}{\left(1 - \frac{gl}{n}\right)^{1/2}} \tag{4.75}$$

onde a variável *gl* representa o número de graus de liberdade consumidos pelos parâmetros do modelo. Posteriormente, *R* novos vetores de resíduos são gerados pelo reordenamento

aleatório dos elementos e_i do vetor original. O reordenamento é feito da mesma forma que é no método *bootstrapping objects*:

$$\begin{aligned} e_i^r &= e_{\xi_i^r} \\ i &= 1,2,...,n \\ r &= 1,2,...,R \end{aligned} \tag{4.76}$$

Com os novos resíduos gerados, R novos vetores para a variável de resposta são reamostrados de acordo com a Eq. 4.77:

$$y_i^r = \hat{y}_i + e_{\xi_i^r} \tag{4.77}$$

Para cada um dos vetores $\boldsymbol{y}^r$ gerados, é computado um novo vetor dos coeficientes $\boldsymbol{b}^r$. Do mesmo modo que no método *bootstrapping objects*, a matriz de covariância também é obtida pela Eq. 4.74.

$$\boldsymbol{S}^2(\boldsymbol{b}) = \frac{1}{R-1}\sum_{r=1}^{R}\left(\boldsymbol{b}^r - \overline{\boldsymbol{b}}\right)\left(\boldsymbol{b}^r - \overline{\boldsymbol{b}}\right)^T \tag{4.74}$$

Segundo Nicolaas e Faber (2002), o método *bootstrapping residuals* fornece melhores resultados que o *bootstrapping objects* para o modelo linear clássico (quando a matriz $\boldsymbol{X}$ apresenta posto completo). Entretanto, como os resultados do primeiro dependem crucialmente do fato do ruído estar normalmente distribuído, o último acaba sendo mais utilizado. Outro aspecto importante é que, ao contrário dos métodos *jackknife* e *bootstrapping objects*, o método *bootstrapping residuals* trabalha com a parte aleatória do modelo, não havendo problemas em o mesmo ser utilizado com dados provenientes de planejamento de experimentos.

4.2.2. Método da Adição de Resíduos

O ponto inicial é a computação do resíduo quadrático residual médio (*MSE – mean square error*):

$$MSE = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n - gl} \tag{4.78}$$

onde, novamente, *gl* representa o número de graus de liberdade consumidos pelos parâmetros do modelo. Então, R novos conjuntos de dados são reamostrados de acordo com:

$$\begin{aligned} y_i^r &= \hat{y}_i + \sqrt{MSE}.U \\ i &= 1,2,...,R \end{aligned} \tag{4.79}$$

onde U representa um número aleatório gerado a partir de uma distribuição normal de média nula e desvio padrão unitário. De posse dos R conjuntos de dados, podem ser obtidas as

estimativas $\boldsymbol{b_1}$, $\boldsymbol{b_2}$, ..., $\boldsymbol{b_R}$ para os vetores dos coeficientes e a aproximação para a matriz de covariância pode ser computada pela Eq. 4.74.

$$\boldsymbol{S}^2(\boldsymbol{b}) = \frac{1}{R-1}\sum_{r=1}^{R}\left(\boldsymbol{b}^r - \bar{\boldsymbol{b}}\right)\left(\boldsymbol{b}^r - \bar{\boldsymbol{b}}\right)^T \tag{4.74}$$

Do mesmo modo que no método *boostrapping residuals*, a hipótese de que as observações experimentais originais são amostras aleatórias de uma distribuição multivariável não precisa ser assumida, o que permite a utilização de técnicas de planejamento de experimentos no levantamento dos dados.

4.3. Reamostragem Baseada no Erro Experimental

Nesta seção, é descrita a metodologia proposta para a obtenção de aproximação de incertezas em coeficientes de modelos de regressão multivariável, o método da adição de erro. O método da adição de erro foi concebido visando superar algumas limitações do método da adição de resíduos. A principal limitação do método da adição de resíduos está no fato do ruído a ser adicionado à variável de saída ser gerado com base nos desvios das predições do modelo. De fato, o valor *MSE* é uma boa aproximação para o erro de medida da variável de saída em casos onde as hipóteses necessárias para a utilização do método dos mínimos quadrados são verdadeiras. Mas, obviamente, a utilização dos resíduos do modelo como aproximação para o erro de medida da variável de resposta pode não ter sentido em muitas situações, como, por exemplo, nos casos onde as entradas também são medidas com erro experimental considerável, a forma funcional utilizada é inadequada ou algum termo explicativo importante não está sendo considerado. Portanto, é proposto que, ao invés de se perturbar os dados com um ruído baseado nos resíduos do modelo, os conjuntos reamostrados sejam computados pela adição de ruído gerado diretamente com base no erro experimental das variáveis (tanto de saída como de entrada), de modo a simular uma perturbação tão idêntica quanto possível às perturbações originais (erros experimentais). Essa é a idéia básica do método da adição de erro.

O método da adição de erro é esquematizado para um sistema genérico, constituído de *n* amostras, *k* variáveis explicativas e uma variável de resposta, na Figura 4.1. A matriz $\boldsymbol{U_0(n,k)}$ e o vetor $\boldsymbol{v_0(n,1)}$, representam os valores verdadeiros das variáveis para as *n* observações disponíveis, que são, na verdade, desconhecidos. Na prática, trabalhamos com as medidas experimentais dos valores verdadeiros, denotadas por $\boldsymbol{X(n,k)}$ e $\boldsymbol{y(n,1)}$. As medidas experimentais correspondem à combinação aditiva dos valores verdadeiros $\boldsymbol{U}$ e $\boldsymbol{v}$ das variáveis em questão com um erro experimental aleatório: $\boldsymbol{X=U_0+D_0}$ e $\boldsymbol{y=v_0+e_0}$. A idéia básica da técnica proposta é a obtenção de *R* novos conjuntos de dados $\boldsymbol{X_1y_1}$, $\boldsymbol{X_2y_2}$, ..., $\boldsymbol{X_Ry_R}$ através da adição de ruído gerado artificialmente às medidas originais ($\boldsymbol{X_i=X+D_i}$ e $\boldsymbol{y_i=y+e_i}$). De posse dos $\boldsymbol{R}$ novos conjuntos de dados, podemos computar as estimativas $\boldsymbol{b_1}$, $\boldsymbol{b_2}$, ..., $\boldsymbol{b_R}$ para os coeficientes do modelo. A aproximação para a matriz de covariância dos coeficientes pode então ser obtida a partir destes *R* vetores através da Eq. 4.74.

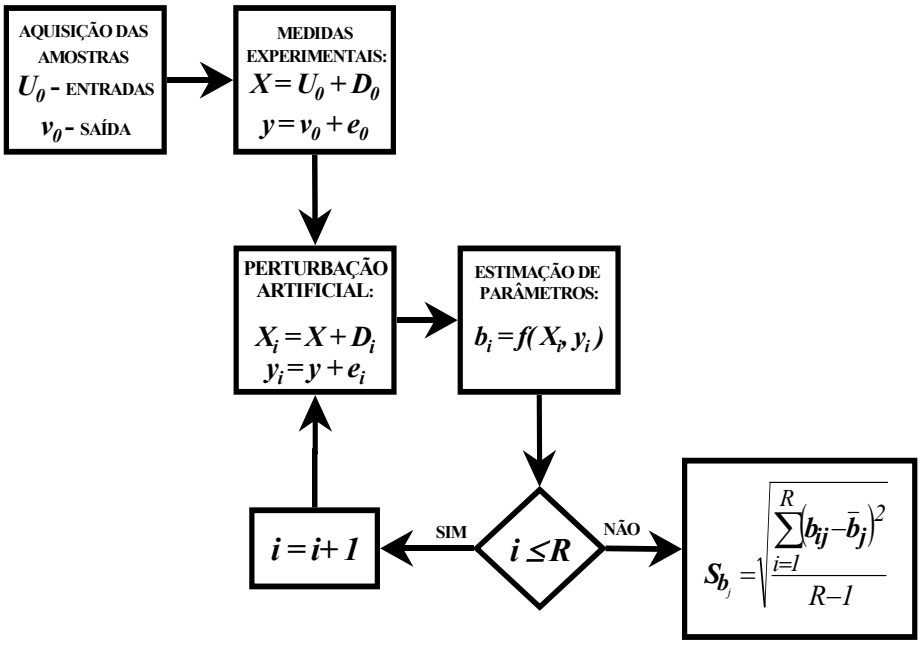

Figura 4.1: Fluxograma ilustrativo do método da adição de erro.

Do mesmo modo que os métodos que se baseiam nos resíduos, o método proposto não pressupõe que as observações experimentais disponíveis (linhas das matrizes $\boldsymbol{X}$ e $\boldsymbol{y}$) sejam amostras aleatórias de uma distribuição multivariável. Portanto, não há impedimentos no que se refere à utilização de técnicas de planejamento de experimentos para o levantamento dos dados. Entretanto, ao contrário das demais técnicas, o método da adição de erro requer, impreterivelmente, a disponibilidade de informações a respeito das incertezas presentes nos dados experimentais.

4.4. Comparação das Metodologias

Nesta seção, através de simulações computacionais, as aproximações para a matriz de covariância dos coeficientes de regressão obtidas pelas técnicas de reamostragem que trabalham com objetos e com resíduos serão comparadas com as aproximações fornecidas pelo método proposto, que trabalha com base nas incertezas experimentais das variáveis de entrada e saída. Vamos estudar dois problemas, um linear e o outro não linear. Em ambos os casos, a relação entre a qualidade da aproximação obtida e o esforço computacional despendido por cada uma das técnicas é avaliada. Todos os cálculos foram realizados em um microcomputador com processador *Pentium* III 933 MH*z*, 128 Mb de memória RAM, utilizando o *software* MATLAB 5.3.

4.4.1. Exemplo Linear

Primeiramente, vamos comparar as estimativas para a incerteza dos coeficientes de regressão fornecidas pelo método proposto com as fornecidas pelos métodos *jackknife*, *bootstrap* e *noise addition* estudando um problema onde a relação existente entre $\boldsymbol{X}$ e $\boldsymbol{y}$ é linear. A comparação será feita por simulações bastante similares às desenvolvidas no artigo Nicolaas e Faber (2002), referenciado anteriormente.

O ponto de partida das simulações é o conjunto de dados apresentado na Tabela 4.1. Estes dados, oriundos de um estudo da relação existente entre a reflectância de luz no infravermelho e o teor de proteínas contido em 24 amostras de trigo moído, não são fornecidos no artigo de Nicolaas e Faber, mas podem ser encontrados em Schacham e Brauner (2003).

Tabela 4.1: Reflectância das amostras para os 6 comprimentos de onda estudados, teor de proteína medido experimentalmente e predito pelo modelo *PLS*

Amostra	x_1	x_2	x_3	x_4	x_5	x_6	Proteína [%] (Real)	Proteína [%] (PLS)
01	468	123	246	374	386	-11.0	9.230	9.322
0'2	458	112	236	368	383	-15.0	8.010	8.099
03	457	118	240	359	353	-16.0	10.95	10.89
04	450	115	236	352	340	-15.0	11.67	11.25
05	464	119	243	366	371	-16.0	10.41	10.10
06	499	147	273	404	433	5.00	9.510	9.221
07	463	119	242	370	377	-12.0	8.670	9.086
08	462	115	238	370	353	-13.0	7.750	7.772
09	488	134	258	393	377	-5.00	8.050	7.739
10	483	141	264	384	398	-2.00	11.39	11.48
11	463	120	243	367	378	-13.0	9.950	9.996
12	456	111	233	365	365	-15.0	8.250	7.969
13	512	161	288	415	443	12.0	10.57	10.38
14	518	167	293	421	450	19.0	10.23	10.23
15	552	197	324	448	467	32.0	11.87	11.79
16	497	146	271	407	451	11.0	8.090	8.179
17	592	229	360	484	524	51.0	12.55	12.52
18	501	150	274	406	407	11.0	8.380	8.633
19	483	137	260	385	374	-3.00	9.640	9.939
20	491	147	269	389	391	1.00	11.35	11.51
21	463	121	242	366	353	-13.0	9.700	9.886
22	507	159	285	410	445	13.0	10.75	10.77
23	474	132	255	376	383	-7.00	10.75	10.92
24	496	152	276	396	404	6.00	11.47	11.51

Resumidamente, as variáveis $\boldsymbol{x_1}$, $\boldsymbol{x_2}$, ..., $\boldsymbol{x_6}$ são medidas experimentais reais da reflectância de luz para as 24 amostras em seis comprimentos de onda na freqüência do infra vermelho próximo, enquanto a variável de resposta ($\boldsymbol{y}$) é a medida experimental do teor de proteína contido em cada uma delas. Estes dados são então utilizados para gerar um novo sistema de 24 amostras fictícias, que são a base das simulações. Nestas simulações, é assumido que, para as 24 amostras fictícias, as medidas experimentais reais de $\boldsymbol{x_1}$, $\boldsymbol{x_2}$, ..., $\boldsymbol{x_6}$ são os valores "verdadeiros" das colunas da matriz $\boldsymbol{X}$. É assumido, ainda, que os valores "verdadeiros" de $\boldsymbol{y}$ são as predições fornecidas pelo modelo *PLS* utilizando três componentes, também apresentados na Tabela 4.1. De posse dos valores "verdadeiros" de $\boldsymbol{X}$ e $\boldsymbol{y}$, a obtenção de medidas "experimentais" é simulada através da adição de ruído aos mesmos.

Os valores "verdadeiros" dos parâmetros do modelo podem ser computados diretamente a partir dos dados da Tabela 4.1. Como o modelo em questão é linear, podemos

trabalhar com os coeficientes em termos das variáveis originais. Os parâmetros referentes a cada um dos comprimentos de onda estudados (b_1, b_2, ..., b_6), assim como o termo independente do modelo (b_0) são reportados na Tabela 4.2.

Tabela 4.2: Valores "verdadeiros" para os coeficientes do modelo *PLS*.

b_1	b_2	b_3	b_4	b_5	b_6	b_0
-0.0370	0.1524	0.1247	-0.1846	0.0129	-0.0653	40.57

Conforme explicado anteriormente, a realização de "medidas experimentais" para as variáveis das 24 amostras mostradas na Tabela 4.2 é simulada pela adição de ruído aos dados. Neste estudo, são realizadas seis simulações, A, B, C, D, E e F, que se diferenciam pelo modo como o ruído é adicionado às variáveis de entrada e saída. Nas simulações A, B e C, é adicionado um ruído normalmente distribuído com média nula desvio padrão igual a, respectivamente, 0.2, 1.0 e 5.0% do valor máximo à variável de saída, não sendo adicionado ruído algum às variáveis de entrada. Nas simulações D, E e F, o mesmo procedimento de adição de ruído à variável de resposta foi adotado, sendo que, nos três casos, foi também adicionado um ruído normalmente distribuído, com média nula e desvio padrão igual a 1% do valor máximo às variáveis explicativas. As estimativas para as incertezas fornecidas por cada uma das técnicas de reamostragem são comparadas entre si com base em estimativas "ideais" para os desvios padrões dos parâmetros, obtidas a partir de mil regressões realizadas com base em mil conjuntos de medidas "experimentais" independentes para as variáveis de entrada e saída. As estimativas "ideais" para o desvio padrão dos parâmetros σ_{b1}, σ_{b2}, ..., σ_{b6} são apresentadas na Tabela 4.3. O desvio padrão do termo independente do modelo é representado por σ_{b0}.

Tabela 4.3: Estimativas "ideais" para o desvio padrão dos parâmetros para as seis simulações.

caso	Ruído [%]		σ_{b1}	σ_{b2}	σ_{b3}	σ_{b4}	σ_{b5}	σ_{b6}	σ_{b0}
	X	***y***							
A	0.0	0.2	4.97E-04	6.38E-04	5.55E-04	9.27E-04	3.85E-04	1.71E-03	4.06E-01
B	0.0	1.0	2.43E-03	3.19E-03	2.75E-03	4.50E-03	1.96E-03	8.64E-03	1.98E+00
C	0.0	5.0	1.18E-02	1.62E-02	1.38E-02	2.28E-02	9.86E-03	4.24E-02	9.70E+00
D	1.0	0.2	2.72E-02	2.61E-02	2.24E-02	2.23E-02	9.53E-03	2.96E-02	7.57E+00
E	1.0	1.0	2.74E-02	2.57E-02	2.28E-02	2.10E-02	9.18E-03	3.17E-02	8.16E+00
F	1.0	5.0	3.15E-02	3.10E-02	2.81E-02	3.02E-02	1.32E-02	4.54E-02	1.08E+01

As estimativas para a incerteza dos parâmetros (S_{b1}, S_{b2}, ..., S_{b6} e S_{b0}) fornecidas pelos métodos *jackknife*, *bootstrap* e *noise addition* são comparadas com as fornecidas pelo método proposto (adição de erro) na Tabela 4.4 e na Tabela 4.5. Para métodos do tipo *bootstrap* e do tipo adição de ruído, os resultados apresentados nestas tabelas foram computados determinando-se o desvio padrão dos coeficientes resultantes de mil reamostragens dos dados da Tabela 4.1. Para o método jackknife, o número de conjuntos a serem reamostrados deve ser igual ao número de observações experimentais, que neste caso é 24. Na realidade, os valores apresentados estão normalizados em relação às respectivas estimativas "ideais". Esta transformação foi escolhida de modo a facilitar a interpretação dos resultados, pois a mesma

faz com que o valor meta passe a ser a unidade. Por fim, deve ser mencionado que, embora uma única realização "experimental" seja suficiente para a obtenção das estimativas para as incertezas nos coeficientes, neste trabalho, estamos trabalhando com a média de cem estimativas, obtidas a partir de cem realizações "experimentais" independentes. Deste modo, o peso de realizações "experimentais" isoladas se torna pequeno, fazendo com que a comparação entre os métodos seja mais confiável.

Tabela 4.4: Médias aritméticas das cem estimativas para o erro dos coeficientes sem erro de "medida" em ***X***. Resultados normalizados em relação às estimativas ideais.

Caso	Ruído [%] X	Ruído [%] y	Método	S_{b1}/σ_{b1}	S_{b2}/σ_{b2}	S_{b3}/σ_{b3}	S_{b4}/σ_{b4}	S_{b5}/σ_{b5}	S_{b6}/σ_{b6}	S_{b0}/σ_{b0}
A	0.0	0.2	Jackknife	49.82	13.72	23.95	21.01	14.42	62.61	54.74
			Bootstrap objects	35.64	16.72	19.40	18.61	12.78	42.40	37.63
			Bootstrap residuals	1.692	1.783	1.748	1.732	1.799	1.776	1.701
			Adição de Resíduos	1.693	1.791	1.761	1.738	1.814	1.785	1.700
			Adição de Erro	0.962	1.024	1.005	0.993	1.041	1.018	0.968
B	0.0	1.0	Jackknife	10.25	2.948	4.945	4.460	3.012	12.45	11.28
			Bootstrap objects	7.365	3.473	4.017	3.997	2.696	8.435	7.790
			Bootstrap residuals	0.972	1.010	0.997	1.012	0.999	0.988	0.983
			Adição de Resíduos	0.971	1.010	1.003	1.007	1.007	0.996	0.980
			Adição de Erro	0.992	1.027	1.016	1.026	1.016	1.011	1.001
C	0.0	5.0	Jackknife	2.453	1.189	1.435	1.407	1.202	2.797	2.616
			Bootstrap objects	1.865	1.158	1.247	1.234	1.102	1.994	1.895
			Bootstrap residuals	0.940	0.911	0.912	0.919	0.922	0.946	0.938
			Adição de Resíduos	0.943	0.917	0.919	0.919	0.922	0.955	0.942
			Adição de Erro	1.050	1.019	1.021	1.025	1.033	1.065	1.050

Em concordância com as conclusões de outros estudos publicados nesta área, como, por exemplo, Nicolaas e Faber (2002) e Hardy et al (1996), é notável o fato de que os métodos que trabalham com objetos (*jackknife* e *bootstrapping objects*) tendem a superestimar a variabilidade dos coeficientes de regressão. Também podemos notar que esta situação piora drasticamente à medida que o teor de ruído presente nos dados diminui. A explicação para esta observação está no fato de como estes métodos executam o procedimento de reamostragem. No caso do método *jackknife*, por exemplo, os novos conjuntos de dados são obtidos subdividindo-se o conjunto de dados original em diversos novos conjuntos com um número inferior de observações (objetos). Obviamente, as estimativas obtidas a partir dos subconjuntos são menos confiáveis do que as obtidas a partir do conjunto de dados completo. Para que o método *jackknife* forneça boas aproximações para o erro dos coeficientes, a diferença entre a confiabilidade das estimativas obtidas a partir do conjunto completo e a confiabilidade das estimativas obtidas a partir dos conjuntos reduzidas deve ser pequena. A princípio, esta parece ser uma hipótese razoável, uma vez que a diferença entre os dois conjuntos é de apenas uma observação. Entretanto, se as incertezas experimentais forem muito pequenas esta diferença pode passar a ser significativa. No que se refere aos métodos que trabalham com resíduos, podemos notar que a variabilidade dos parâmetros também é superestimada em casos onde o erro experimental apresenta baixa magnitude. Esta observação é conseqüência do compromisso entre *bias e* variância que está associado ao número de

componentes utilizados no modelo *PLS*. Quando o erro experimental é muito pequeno, o *bias* incorporado às predições pode passar a ser significativo em relação ao erro residual do modelo. Isso significa que a utilização dos resíduos como uma aproximação para o erro experimental pode passar a ser pessimista a medida que a precisão dos dados aumenta, o que explicaria os resultados que foram obtidos.

Tabela 4.5: Médias aritméticas das cem estimativas para o erro dos coeficientes com erro de "medida" em $\boldsymbol{X}$. Resultados normalizados em relação às estimativas ideais.

Caso	Ruído [%] X	Ruído [%] y	Método	S_{b1}/σ_{b1}	S_{b2}/σ_{b2}	S_{b3}/σ_{b3}	S_{b4}/σ_{b4}	S_{b5}/σ_{b5}	S_{b6}/σ_{b6}	S_{b0}/σ_{b0}
D	1.0	0.2	Jackknife	1.134	1.155	1.157	1.354	1.563	2.288	1.965
			Bootstrap objects	0.984	1.066	1.082	1.207	1.353	1.964	1.680
			Bootstrap residuals	0.687	0.678	0.788	0.950	1.063	1.106	1.063
			Adição de Resíduos	0.745	0.771	0.848	0.972	1.103	1.270	1.138
			Adição de Erro	0.671	0.763	0.797	0.775	0.822	0.853	0.838
E	1.0	1.0	Jackknife	1.116	1.209	1.228	1.594	1.698	2.312	1.817
			Bootstrap objects	1.007	1.132	1.163	1.427	1.511	1.952	1.604
			Bootstrap residuals	0.758	0.757	0.844	1.115	1.205	1.138	1.075
			Adição de Resíduos	0.824	0.858	0.913	1.170	1.260	1.337	1.160
			Adição de Erro	0.702	0.815	0.832	0.918	0.902	0.876	0.832
F	1.0	5.0	Jackknife	1.149	1.190	1.177	1.245	1.433	1.788	1.571
			Bootstrap objects	1.047	1.138	1.124	1.143	1.278	1.562	1.414
			Bootstrap residuals	0.822	0.792	0.856	0.948	1.029	1.020	1.023
			Adição de Resíduos	0.907	0.906	0.939	0.988	1.065	1.195	1.120
			Adição de Erro	0.895	0.999	1.006	0.924	0.973	1.028	1.011

Para todas as seis simulações realizadas, o método proposto forneceu resultados razoavelmente próximos à estimativa "ideal" das incertezas presentes nos coeficientes. Também foi observado que, especialmente nas situações onde é considerada a presença de ruído nas entradas, a qualidade das estimativas do método proposto em relação às demais metodologias tendem a variar menos de um coeficiente para o outro. Um outro aspecto a ser analisado na comparação das metodologias é a reprodutibilidade dos resultados, que, neste caso, pode ser quantificada pelo desvio padrão das cem estimativas para as incertezas nos coeficientes de regressão fornecidas por cada um dos métodos. Estas valores são apresentados na Tabela 4.6 e na Tabela 4.7 para todas as simulações realizadas.

Como era esperado, a variabilidade das estimativas para as incertezas dos coeficientes aumenta a medida que os erros "experimentais" são maiores. Em uma primeira análise, pode-se ter a falsa impressão de que a variabilidade das estimativas fornecidas pelos métodos *jackknife* e *bootstrap objects* diminuem ao adicionarmos ruído nas variáveis explicativas. Porém, é evidente que a comparação direta entre os valores dos desvios para estes casos não é adequada dada a considerável diferença de magnitude das estimativas médias. A análise da Tabela 4.6 e da Tabela 4.7 também revela que a metodologia proposta é mais robusta que as demais, pois, em todos os casos, o valor do desvio padrão das cem estimativas para o erro de cada um dos coeficientes apresentado pela mesma é visivelmente menor que o verificado na utilização dos outros métodos.

Tabela 4.6: Desvio padrão das cem estimativas para o erro dos coeficientes sem erro de "medida" em ***X***. Resultados normalizados em relação às estimativas ideais.

Caso	Ruído [%] X	Ruído [%] y	Método	S_{b1}/σ_{b1}	S_{b2}/σ_{b2}	S_{b3}/σ_{b3}	S_{b4}/σ_{b4}	S_{b5}/σ_{b5}	S_{b6}/σ_{b6}	S_{b0}/σ_{b0}
A	0.0	0.2	Jackknife	0.58	0.29	0.40	0.64	0.37	0.43	0.51
			Bootstrap objects	0.80	0.56	0.57	0.68	0.49	0.85	0.80
			Bootstrap residuals	0.110	0.106	0.105	0.109	0.107	0.113	0.112
			Adição de Resíduos	0.100	0.113	0.113	0.107	0.106	0.109	0.101
			Adição de Erro	0.023	0.025	0.024	0.023	0.023	0.023	0.023
B	0.0	1.0	Jackknife	0.56	0.304	0.381	0.662	0.359	0.41	0.49
			Bootstrap objects	0.349	0.202	0.235	0.283	0.211	0.308	0.328
			Bootstrap residuals	0.148	0.155	0.154	0.154	0.149	0.149	0.149
			Adição de Resíduos	0.148	0.149	0.150	0.153	0.156	0.151	0.150
			Adição de Erro	0.022	0.025	0.025	0.024	0.025	0.024	0.022
C	0.0	5.0	Jackknife	0.503	0.281	0.331	0.479	0.282	0.377	0.449
			Bootstrap objects	0.271	0.174	0.208	0.202	0.172	0.256	0.247
			Bootstrap residuals	0.159	0.148	0.152	0.151	0.155	0.163	0.158
			Adição de Resíduos	0.158	0.147	0.150	0.148	0.150	0.164	0.158
			Adição de Erro	0.048	0.045	0.051	0.044	0.049	0.056	0.047

Tabela 4.7: Desvio padrão das cem estimativas para o erro dos coeficientes com erro de "medida" em ***X***. Resultados normalizados em relação às estimativas ideais.

Caso	Ruído [%] X	Ruído [%] y	Método	S_{b1}/σ_{b1}	S_{b2}/σ_{b2}	S_{b3}/σ_{b3}	S_{b4}/σ_{b4}	S_{b5}/σ_{b5}	S_{b6}/σ_{b6}	S_{b0}/σ_{b0}
D	1.0	0.2	Jackknife	0.378	0.461	0.325	0.484	0.356	0.656	0.642
			Bootstrap objects	0.239	0.313	0.267	0.309	0.257	0.505	0.411
			Bootstrap residuals	0.173	0.181	0.194	0.208	0.183	0.294	0.248
			Adição de Resíduos	0.201	0.238	0.225	0.226	0.205	0.404	0.287
			Adição de Erro	0.109	0.129	0.133	0.158	0.114	0.187	0.122
E	1.0	1.0	Jackknife	0.331	0.452	0.453	0.535	0.408	0.697	0.507
			Bootstrap objects	0.232	0.339	0.346	0.392	0.295	0.436	0.384
			Bootstrap residuals	0.190	0.213	0.254	0.252	0.199	0.299	0.254
			Adição de Resíduos	0.230	0.286	0.285	0.304	0.228	0.433	0.323
			Adição de Erro	0.120	0.154	0.139	0.199	0.115	0.192	0.154
F	1.0	5.0	Jackknife	0.404	0.507	0.504	0.419	0.386	0.640	0.538
			Bootstrap objects	0.242	0.347	0.371	0.294	0.283	0.395	0.349
			Bootstrap residuals	0.203	0.234	0.257	0.233	0.189	0.296	0.250
			Adição de Resíduos	0.249	0.311	0.304	0.260	0.205	0.379	0.307
			Adição de Erro	0.129	0.180	0.174	0.157	0.121	0.215	0.158

De um modo geral, podemos dizer que o método da adição de erro apresentou melhor desempenho que os demais. O melhor desempenho do método proposto em relação às técnicas que trabalham com os resíduos do modelo deve-se ao fato do mesmo trabalhar explicitamente com os erros experimentais. Ao reamostrar o conjunto de dados com base nos resíduos das predições, os demais métodos se tornam suscetíveis às falhas do próprio modelo. O método proposto simplesmente perturba as variáveis ***X*** e ***y*** de maneira tão idêntica quanto

possível aos erros experimentais e, então, avalia como estas perturbações se propagam aos parâmetros estimados. Desta forma, limitações inerentes ao modelo, como a presença de bias ou de erros sistemáticos, não deverão prejudicar a obtenção de estimativas para as incertezas nos coeficientes de regressão. Entretanto, é importante lembrar que, ao contrário de todos os outros métodos avaliados, os resultados fornecidos pela técnica proposta estão relacionados à qualidade da informação a respeito dos erros experimentais disponível. Nas simulações conduzidas, obviamente, as informações a respeito do erro "experimental" são precisas e, portanto, os resultados obtidos são confiáveis.

Finalmente, os métodos devem ser comparados quanto ao esforço computacional despendido para a obtenção das estimativas. Tal comparação pode ser feita com base nos dados da Tabela 4.8, que apresenta o tempo em segundos gasto por cada um dos métodos para obter as estimativas para as incertezas dos coeficientes dos modelos a partir das cem réplicas "experimentais" avaliadas.

Tabela 4.8: Tempo em segundos gasto pelos métodos nas seis simulações conduzidas.

Caso	A	B	C	D	E	F
Erro em $\boldsymbol{X}$ [%]	0.0	0.0	0.0	1.0	1.0	1.0
Erro em $\boldsymbol{y}$ [%]	0.2	1.0	5.0	0.2	1.0	5.0
Jackknife	12.71	12.95	13.01	12.87	13.00	12.92
Bootstrap objects	831.7	845.9	859.3	863.4	845.4	840.7
Bootstrap residuals	836.0	842.1	847.3	846.6	839.6	842.4
Adição de Resíduos	1039	1046	1046	1048	1046	1048
Adição de Erro	1097	1098	1107	2599	2588	2595

Em todos os casos, o tempo de computação despendido pelo método *jackknife* foi nitidamente inferior ao despendido pelos outros métodos. Isso ocorre porque o número de conjuntos que podem ser reamostrados pelo método *jackknife* é limitado ao número de observações experimentais presentes no conjunto de dados original que, neste caso, é de 24 amostras. Como os demais métodos utilizaram mil conjuntos reamostrados para a obtenção das estimativas para as incertezas dos coeficientes, o método *jackknife*, embora seja mais rápido, tende a fornecer resultados menos precisos, o que de fato foi observado. É interessante notarmos que, em outras situações, podemos verificar o problema inverso, ou seja, se o número de amostras a serem consideradas for muito grande, o tempo computacional requerido pelo método *jackknife* poderá vir a ser muito maior que o requerido pelos outros métodos avaliados. Portanto, podemos notar que a utilização do método *jackknife*, por não possibilitar que o número de conjuntos reamostrados seja manipulado, não apresenta flexibilidade no que diz respeito ao compromisso existente entre a precisão dos resultados e o tempo computacional requerido para alcançá-los. Portanto, na discussões que seguem, vamos ignorar o método *jackknife* e considerar que apenas o método *bootstrapping objects* representa a classe de métodos que trabalham com os objetos.

No que se refere à comparação do esforço computacional requerido pelos demais métodos, percebe-se as duas variantes do método *bootstrap* são mais rápidas que os métodos de adição de ruído. A explicação para esta observação está no fato de que, nos dois últimos

métodos, o rúido a ser adicionado aos dados deve ser gerado novamente para cada um dos conjuntos reamostrados. Pelo mesmo motivo, o método da adição de erro torna-se visivelmente mais lento que o método da adição de resíduos quando é considerada a presença de erro nas variáveis explicativas. Como conclusão geral, pode ser dizer que, embora esteja vinculada a disponibilidade de informações a respeito dos erros experimentais presentes nos dados, a metodologia proposta fornece estimativas mais realistas para as incertezas dos coeficientes de regressão multivariável. O tempo computacional extra requerido pelo método da adição de erro em relação às técnicas do tipo *bootstrap* é de cerca de 20 a 30% em situações onde as varáveis explicativas são medidas sem erro é de cerca de 200 a 300% em situações onde a incerteza presente na determinação das entradas é considerável. Estes valores podem variar em função das dimensões do conjunto de dados original.

4.4.2. Exemplo Não Linear

Nesta seção, será realizado um estudo comparativo visando avaliar o comportamento das técnicas de determinação de incertezas em coeficientes de regressão dos métodos de modelagem não lineares apresentados no capítulo de revisão. Do mesmo modo que no estudo do caso linear, vamos simular a situação onde dispomos de um determinado número de amostras e desejamos avaliar o desvio padrão das estimativas dos parâmetros do modelo obtidos a partir das mesmas. Os valores "verdadeiros" das variáveis de entrada e saída são previamente computados, segundo algum critério estabelecido. De posse desses valores, a realização de "experimentos" é simulada adicionando-se ruído aos dados.

Vamos considerar uma situação onde temos uma única variável de saída, que deve ser relacionada com 4 variáveis de entrada não correlacionadas, cujos valores distribuem-se uniforme e aleatoriamente entre –0.25 e 0.25. A relação exata existente entre a variável de saída e as variáveis de entrada é dada pela Eq. 4.80, que apresenta uma função originalmente criada para comparar modelos de redes neurais com modelos obtidos por métodos estatísticos. Posteriormente, esta função foi utilizada por Baffi et al (1999) para ilustrar o desempenho do então proposto algoritmo *QPLS* modificado.

$$\boldsymbol{y} = \exp\left(2\boldsymbol{x_1}\,\mathrm{sen}\left(\pi \boldsymbol{x_4}\right)\right) + \mathrm{sen}\left(\boldsymbol{x_2}\boldsymbol{x_3}\right) \tag{4.80}$$

Escolhendo aleatoriamente cinqüenta valores para cada uma das variáveis de entrada, vamos, utilizando a Eq. 4.80, computar os cinqüenta valores correspondentes para a variável de saída. Desta forma, obtemos valores "verdadeiros" para os elementos da matriz $\boldsymbol{X}$ e do vetor $\boldsymbol{y}$. Os valores "verdadeiros" das variáveis de entrada e saída para as cinqüenta amostras para este estudo são mostrados na Tabela 4.9.

O nosso objetivo é a utilização destes valores na realização de uma análise comparativa entre os métodos de reamostragem encontrados na literatura e o método proposto para a estimação de incertezas em parâmetros de modelos obtidos por métodos de redução de dimensionalidade não lineares. A análise será feita utilizando o método *QPLS* para modelar a relação entre a variável de entrada e as variáveis de saída. Porém, antes de iniciarmos a análise, é importante que algumas particularidades do problema não linear sejam ressaltadas.

Primeiramente, é evidente que a versão não linear de algoritmo *NIPALS* requer um esforço computacional grande em comparação à versão linear. Portanto, algumas modificações serão feitas para reduzir o tempo de computação necessário para a condução das simulações. Iremos trabalhar com uma única determinação "experimental" das amostras da Tabela 4.9 que, ao invés de mil, será reamostrada apenas cinqüenta vezes por cada um dos métodos.

Tabela 4.9: Valores verdadeiros para as variáveis de entrada e saída das 50 amostras geradas.

n°	x_1	x_2	x_3	x_4	y	n°	x_1	x_2	x_3	x_4	y
01	0.133	0.001	-0.047	0.088	1.075	26	-0.014	0.054	0.222	0.173	0.997
02	0.242	0.173	-0.220	0.182	1.261	27	0.014	0.161	-0.082	0.034	0.990
03	-0.101	-0.090	-0.210	-0.042	1.046	28	0.051	-0.020	0.157	0.067	1.019
04	0.069	-0.057	-0.143	0.196	1.091	29	-0.008	0.021	-0.186	-0.101	1.001
05	0.136	0.165	0.066	0.147	1.140	30	-0.201	0.209	0.001	-0.096	1.127
06	0.016	0.241	0.096	0.039	1.027	31	-0.207	-0.041	-0.223	-0.126	1.182
07	0.177	0.014	-0.006	-0.001	0.999	32	-0.184	0.048	-0.107	-0.145	1.171
08	0.031	-0.232	0.121	0.067	0.985	33	-0.123	0.072	-0.063	-0.223	1.167
09	-0.175	0.129	0.106	-0.246	1.291	34	0.142	0.048	-0.186	0.152	1.131
10	-0.037	0.233	-0.093	-0.037	0.987	35	0.129	-0.181	0.185	0.104	1.053
11	0.126	-0.060	-0.110	0.075	1.068	36	-0.190	-0.043	0.068	0.122	0.864
12	0.109	0.028	-0.214	-0.167	0.890	37	0.012	0.155	0.240	0.226	1.053
13	-0.166	0.144	0.125	0.002	1.016	38	0.207	-0.049	-0.071	0.137	1.191
14	-0.101	0.080	-0.218	-0.245	1.133	39	-0.168	-0.113	-0.029	0.086	0.918
15	0.215	-0.115	0.089	0.182	1.252	40	0.226	0.241	0.197	0.015	1.069
16	0.232	-0.173	0.132	-0.052	0.905	41	0.098	0.159	-0.132	-0.105	0.918
17	-0.151	-0.024	-0.053	0.156	0.869	42	-0.155	-0.205	-0.092	-0.053	1.071
18	-0.038	0.087	-0.240	0.174	0.941	43	-0.073	-0.036	-0.162	0.190	0.927
19	0.121	0.159	-0.138	-0.227	0.832	44	-0.217	0.014	-0.196	-0.009	1.010
20	-0.209	0.228	-0.021	0.147	0.825	45	0.018	0.213	-0.142	-0.021	0.967
21	-0.225	-0.120	-0.217	0.044	0.965	46	-0.184	-0.201	-0.235	0.188	0.862
22	0.007	0.033	-0.124	-0.094	0.992	47	0.226	0.100	-0.046	-0.187	0.774
23	-0.080	-0.168	-0.087	0.159	0.941	48	-0.220	0.069	0.015	0.236	0.744
24	-0.052	0.029	0.247	0.239	0.939	49	-0.151	-0.178	-0.165	-0.019	1.048
25	0.232	0.184	0.203	-0.191	0.807	50	0.032	-0.178	0.104	0.086	0.999

Outra particularidade deste exemplo em relação ao anterior está no fato do método *QPLS* não fornecer uma equação relacionando diretamente as variáveis de entrada saída. Relembrando, o algoritmo *QPLS* extrai as variáveis latentes $\boldsymbol{t}$ e $\boldsymbol{u}$ das matrizes de dados originais utilizando, respectivamente, os vetores pesos $\boldsymbol{w}$ e $\boldsymbol{q}$. Posteriormente, a relação entre $\boldsymbol{t}$ e $\boldsymbol{u}$ é mapeada por um polinômio de segundo grau ($\hat{\boldsymbol{u}} = c_0 + c_1\boldsymbol{t} + c_2\boldsymbol{t}^2$). Como o algoritmo *QPLS* não trabalha com coeficientes em termos das variáveis originais, para computar as predições, são necessários, além dos coeficientes dos polinômios utilizados para mapear a relação existente entre os diferentes pares de variáveis latentes, os vetores pesos associados à extração das direções. Deste modo, uma análise completa requer que sejam avaliadas a estabilidade dos coeficientes dos polinômios e a estabilidade dos elementos dos vetores peso. Entretanto, para simplificar o estudo, vamos nos deter somente na avaliação das incertezas presentes nos coeficientes dos polinômios.

Outra diferença importante em relação ao exemplo linear é que, devido ao modo como os dados foram gerados, não sabemos de antemão o número correto de variáveis latentes a serem extraídas. Baffi et al (1999), ao utilizarem esta função para testar o novo algoritmo *QPLS*, demonstraram que, nitidamente, apenas as duas primeiras direções contribuem para a capacidade preditiva do modelo. Portanto, vamos considerar apenas os casos onde são extraídas uma e duas variáveis latentes das matrizes originais.

No demais, a análise é análoga à realizada no exemplo linear. Os valores "verdadeiros" para os coeficientes c_0, c_1 e c_2 podem ser computados diretamente a partir dos dados da Tabela 4.9. Estes valores são apresentados na Figura 4.2, onde também é ilustrado o mapeamento da relação existente entre os dois primeiros pares de variáveis latentes.

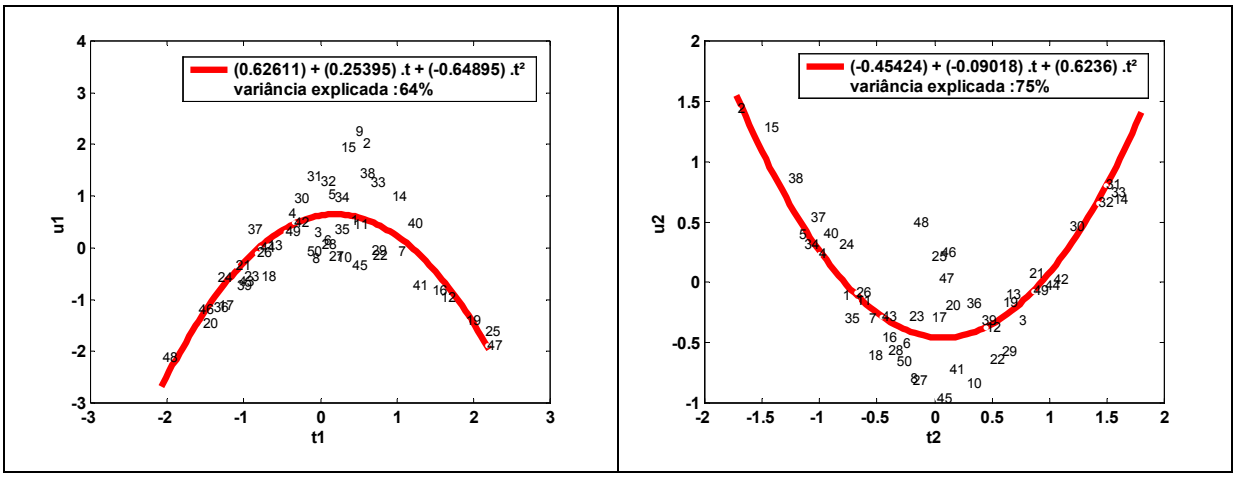

Figura 4.2: Mapeamento da relação entre as variáveis latentes do modelo *QPLS*.

Vamos estudar três casos A, B e C, que correspondem, respectivamente, às situações onde o desvio padrão do erro "experimental" é igual a 0.1, 1.0 e 5.0% do valor máximo da variável de resposta. Para os três casos estudados, podemos obter as estimativas "ideais" para o desvio padrão dos coeficientes c_0, c_1 e c_2 com base em mil realizações "experimentais" independentes, obtidas pela adição de ruído aos dados originais. As estimativas "ideais" para o desvio padrão dos coeficientes são apresentadas na Tabela 4.10

Do mesmo modo que no exemplo linear, vamos comparar o desempenho dos métodos apresentados na Seção 4.1 com o método proposto analisando os erros normalizados em relação às estimativas "ideais". Entretanto, neste caso, algumas questões a respeito das estimativas "ideais" apresentadas na Tabela 4.10 devem ser consideradas. Vamos, então, analisar a Figura 4.3 e a Figura 4.4 que apresentam os valores dos parâmetros c_0, c_1 e c_2 estimados para mapear a relação existente entre os dois primeiros pares de variáveis latentes nas mil regressões utilizadas na computação das estimativas 'ideais". Os pontos correspondem às estimativas obtidas a partir dos conjuntos perturbados ao passo que a linha escura representa o valor obtido com o conjunto original, ou seja, aos valores "verdadeiros" dos coeficientes, apresentados na Figura 4.2.

Tabela 4.10: Estimativas "ideais" para o desvio padrão dos coeficientes do modelo *QPLS*.

Caso	Ruído [%]		VL-1			VL-2		
	X	**y**	σ_{c0}	σ_{c1}	σ_{c2}	σ_{c0}	σ_{c1}	σ_{c2}
A	0.0	0.1	5.19E-04	1.30E-03	5.52E-04	1.90E-01	1.76E-01	5.77E+00
B	0.0	1.0	5.26E-03	1.27E-02	5.60E-03	2.07E-01	2.01E-01	6.78E+00
C	0.0	5.0	3.75E-01	7.85E-02	3.78E-01	2.56E-01	4.98E-01	1.37E+01

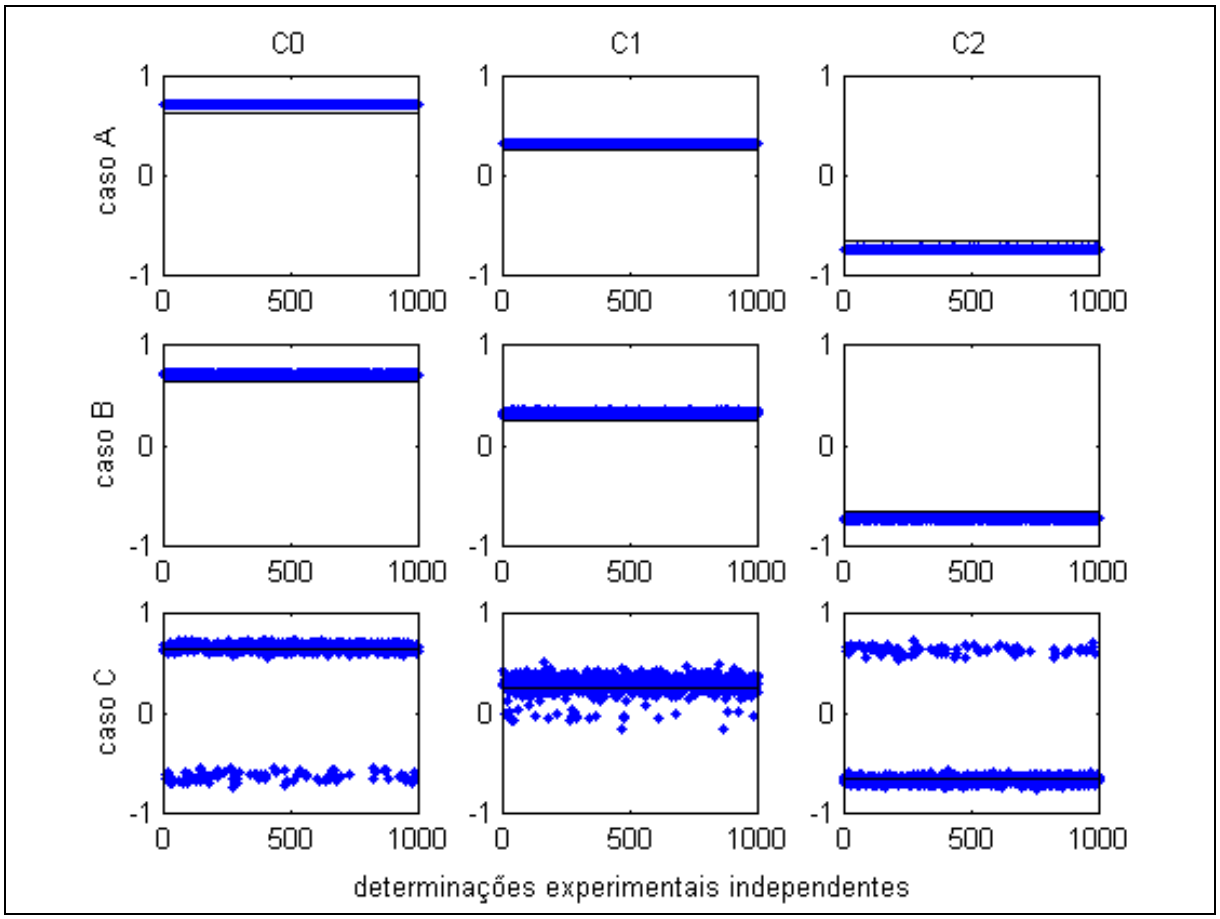

Figura 4.3: Estimativas dos coeficientes da primeira direção do modelo *QPLS*.

No que se refere à primeira variável latente, para os casos A e B, pode ser verificado que as estimativas obtidas se distribuem em uma determinada região próxima ao valor "verdadeiro" e que a tendência central das estimativas parece sempre estar levemente deslocada em relação ao valor verdadeiro. Entretanto, no caso C, é visível o fato de que as estimativas dos coeficientes passam a se distribuir em duas regiões distintas. Esta "anormalidade" observada no caso C é conseqüência da utilização do algoritmo *PLS* não linear na extração das variáveis latentes. A função objetivo utilizada na busca das direções ótimas pode apresentar uma relação um tanto o quanto complexa com as medidas experimentais. No caso C, verificamos que, em algumas das determinações "experimentais" realizadas, a perturbação nos dados originais modificou a função objetivo de uma maneira tal que uma direção completamente diferente da original passou a ser a ótima.

Na Figura 4.4, podemos notar que a situação é um pouco diferente. Como a maior parte da variabilidade de ***y*** é capturada no primeiro estágio da decomposição, as direções extraídas posteriormente se tornam muito mais suscetíveis aos efeitos do ruído. Conseqüentemente, quando estudamos a variabilidade dos coeficientes relacionados ao mapeamento da relação existente entre o segundo par de variáveis latentes, o problema da multiplicidade de direções pode ser verificado também nos casos A e B. Para o caso C, a análise da Figura 4.4 levanta suspeitas quanto às vantagens relacionadas à utilização do segundo par de variáveis latentes no modelo. Entretanto, para que se chegue a alguma conclusão definitiva a respeito desta questão, é necessário testar se esta direção contribui ou não para o aumento da capacidade preditiva do modelo, o que foge ao escopo desta discussão.

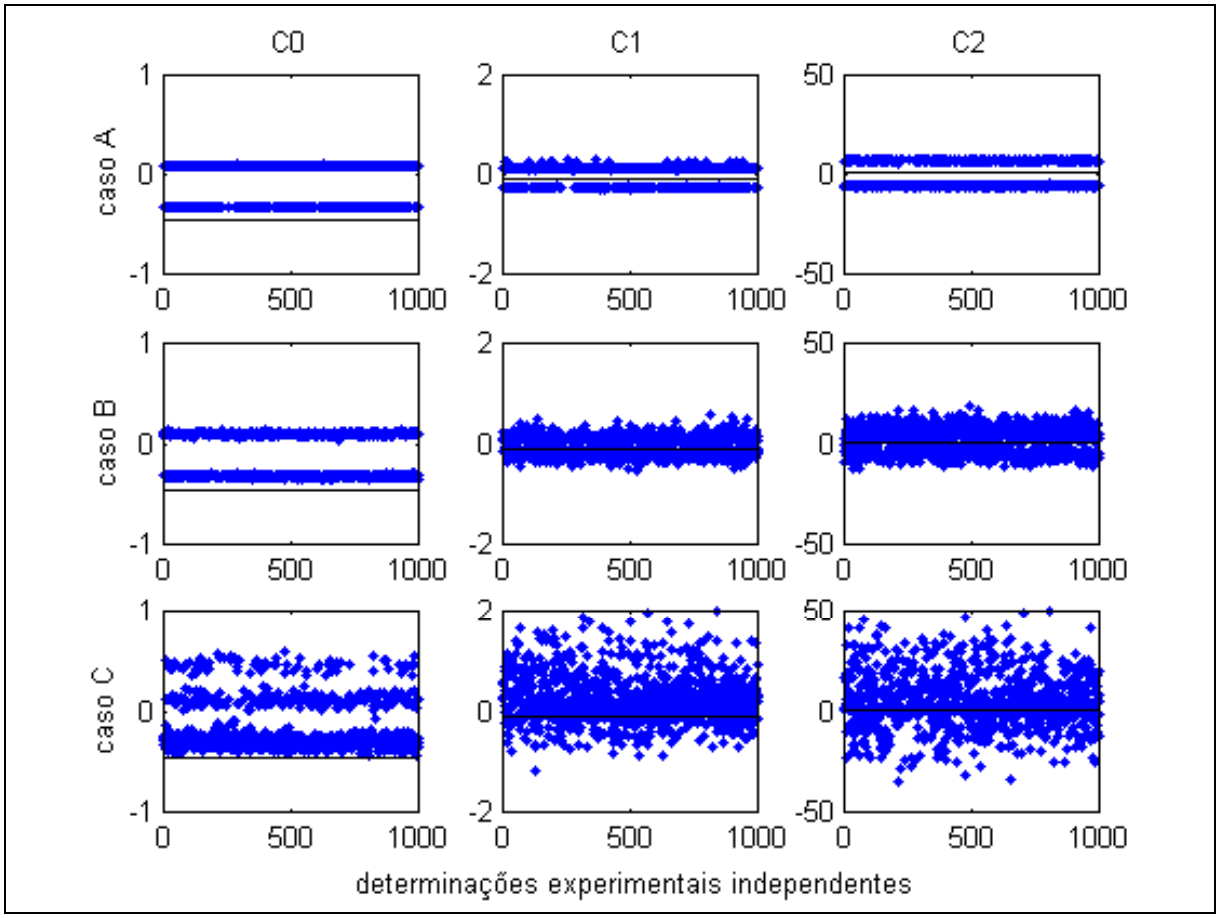

Figura 4.4: Estimativas dos coeficientes da segunda direção do modelo *QPLS*.

Obviamente, a comparação entre os métodos de reamostragem só faz sentido para os casos onde não há o surgimento de duas regiões de distribuição dos parâmetros. Por isso, nossa comparação se limitará às estimativas para os erros dos coeficientes no mapeamento da relação entre o primeiro par de variáveis latentes ***t*** e ***u***, nos casos A e B. As aproximações para os desvios padrões destes coeficientes normalizadas pelas respectivas estimativas "ideais" são apresentadas na Tabela 4.11.

Tabela 4.11: Aproximação para o desvio padrão dos coeficientes normalizada pelas estimativas ideais fornecidas pelos diferentes métodos de reamostragem.

Método	Caso A			Caso B		
	S_{c0}/σ_{c0}	S_{c1}/σ_{c1}	S_{c2}/σ_{c2}	S_{c0}/σ_{c0}	S_{c1}/σ_{c1}	S_{c2}/σ_{c2}
Jackknife	190.3	159.8	160.5	18.81	16.36	15.83
Bootstrap objects	1236	204.5	1190.8	123.6	20.95	118.8
Bootstrap residuals	253.4	42.00	241.2	10.03	2.717	9.620
Adição de Resíduos	416.4	48.72	396.9	16.01	2.868	15.27
Adição de Erro	1.014	0.964	1.028	0.988	0.990	1.034

Como podemos observar, do mesmo modo que no exemplo linear, o método proposto é o que fornece estimativas mais precisas das incertezas presentes nos coeficientes da regressão. Novamente, é verificado que quando o erro "experimental" é de pequena magnitude as técnicas baseadas nos objetos e nos resíduos tendem a superestimar a variância dos coeficientes do modelo. As explicações para esta observação são as mesmas do caso linear. Um outro fator que, neste exemplo em particular, pode contribuir de forma negativa para à performance dos métodos baseados nos resíduos é o fato de a relação mapeada entre $\boldsymbol{t_1}$ $\boldsymbol{u_1}$ apresentar erro sistemático. Como pode ser claramente observado na Figura 4.2, o valor absoluto dos resíduos da função mapeada aumentam à medida que o valor de $\boldsymbol{u}$ aumenta. Como há apenas uma variável de resposta, podemos dizer com segurança que os resíduos do modelo aumentam quando $\boldsymbol{y}$ aumenta, o que interfere diretamente na reamostragem realizada pelos métodos *bootstrapping objects* e adição de resíduos.

Por fim, vamos, com base nos dados apresentados na Tabela 4.12, realizar uma comparação do esforço computacional despendido por cada um dos métodos para a obtenção das estimativas para as incertezas nos coeficientes do modelo *QPLS*. Pela mesma razão discutida no exemplo anterior, o método *jackknife* será ignorado nas discussões que seguem.

Tabela 4.12: Tempo em segundos gasto pelos métodos nas três simulações conduzidas.

Caso	A	B	C
Erro em $\boldsymbol{X}$ [%]	0.0	0.0	0.0
Erro em $\boldsymbol{y}$ [%]	0.1	1.0	5.0
Jackknife	830.3	828.7	847.5
Bootstrap objects	1636	1677	1648
Bootstrap residuals	1662	1686	1712
Adição de Resíduos	1636	1653	1656
Adição de Erro	1663	1651	1656

Ao contrário do exemplo linear, onde verificamos uma significativa variação entre o esforço computacional requerido pelas diferentes técnicas de reamostragem estudadas, a diferença máxima os valores apresentados na Tabela 4.12 não chega a 5%. Na realidade, a tendência de diminuição na diferença entre os tempos no caso não linear já era esperada. Para explicar esta afirmativa, é importante ressaltarmos que estes valores são dependentes de dois tipos de computações. As computações referentes à extração e mapeamento das variáveis latentes, iguais para todos os métodos, e as computações referentes à realização da

reamostragem, que são o que diferenciam um método do outro. No caso linear, as computações referentes à extração das variáveis latentes são relativamente rápidas e, como os cálculos são executados em série, as computações referentes à reamostragem exercem um papel importante na determinação tempo total despendido por cada técnica. Já no caso não linear, a etapa de extração das direções é, em termos de esforço computacional, muito mais onerosa, fazendo com que as computações relacionadas à reamostragem exerçam pouca ou nenhuma influência no tempo de obtenção das aproximações para o desvio padrão dos coeficientes.

Capítulo 5 Sistemática de Análise e Estudos de Caso

Neste capítulo, a aplicação dos métodos revisados e propostos nos capítulos anteriores desta dissertação é organizada de uma maneira sistemática. A sistemática utilizada constitui a base para a implementação da ferramenta de regressão multivariável desejada. O capítulo está dividido em duas seções. A primeira seção é constituída da apresentação da sistemática propriamente dita. Na seção seguinte, a utilização da ferramenta desenvolvida para a obtenção de modelos empíricos é ilustrada através de alguns estudos de caso, baseados em dados obtidos na literatura. Serão estudados cinco problemas, quatro dos quais são referentes à obtenção de modelos para a predição de propriedades relacionadas à qualidade do produto final em diferentes tipos de processos industriais. Os dois primeiros exemplos foram escolhidos de modo a ressaltar a importância da utilização de métodos de redução de dimensionalidade para a estimação dos parâmetros. O terceiro exemplo foi escolhido visando demonstrar a importância da questão da seleção de variáveis na construção do modelo. Por fim, o quarto exemplo foi escolhido para testar a ferramenta desenvolvida quando a utilização de métodos de modelagem não lineares se faz necessária. Como será discutido posteriormente, o quarto exemplo não se mostrou adequado para ilustrar a aplicabilidade dos métodos não lineares e, portanto, tal ilustração foi realizada através de uma simulação matemática, que constitui o quinto estudo de caso.

Em todos os casos, utilizou-se o método *SRMP* para selecionar as variáveis que devem compor o modelo final e o método *bootstrapping residuals* para determinar os intervalos de confiança das estimativas, exceto no terceiro exemplo, onde foi utilizado o método da adição de erro. O método da adição de erro, proposto nos capítulos anteriores desta dissertação, não pôde ser utilizado nos demais exemplos devido a indisponibilidade de informações a respeito do erro experimental associado à medida das variáveis.

5.1. Sistemática de Análise

Nos capítulos anteriores, foram tratadas diferentes questões referentes à construção de modelos empíricos. No Capítulo 2, foi realizada uma revisão bibliográfica a respeito dos

diferentes métodos de modelagem. No Capítulo 3, foi proposto um procedimento que permite a seleção das variáveis explicativas que são de fato adequadas para compor o modelo. No Capítulo 4, foram revisadas algumas técnicas de estimação de incertezas em modelos de regressão multivariável e uma nova metodologia foi proposta. Nesta seção, a aplicação dos métodos revisados e propostos nas seções anteriores é organizada de uma maneira sistemática, constituindo a base para a implementação de uma completa ferramenta para a obtenção de modelos empíricos. A sistemática desenvolvida é representada esquematicamente na Figura 5.1.

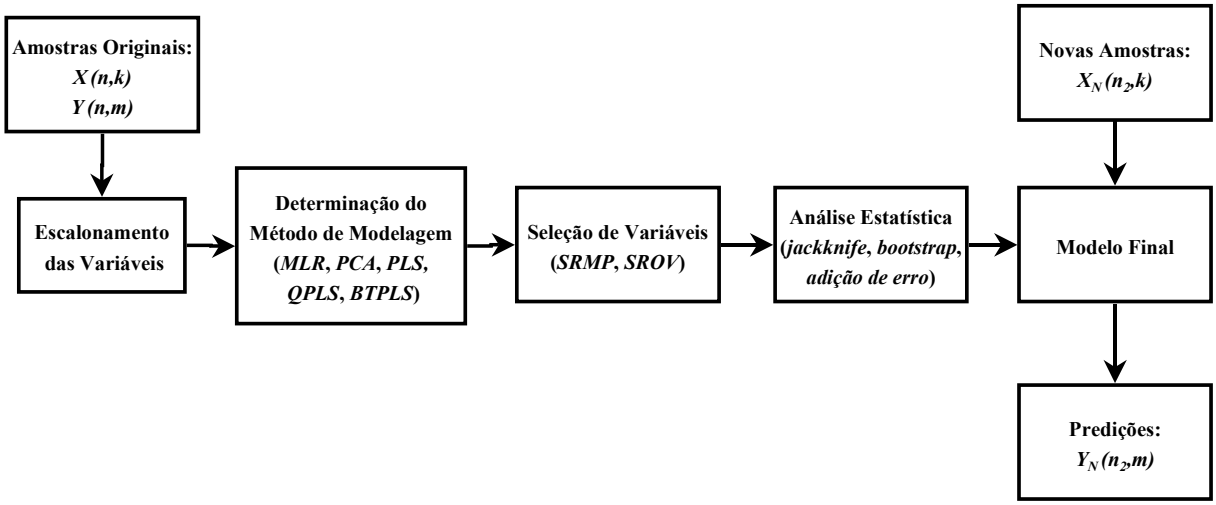

Figura 5.1: Ilustração da sistemática de análise para a obtenção de modelos empíricos.

Inicialmente, os dados referentes às variáveis de entrada e saída (***X*** e ***Y***) devem ser escalonados. Isso deve ser feito para evitar que a escala na qual as variáveis foram determinadas afete os resultados da modelagem. Normalmente, os dados devem ser escalonados de modo que as variáveis passem a apresentar média nula e variância unitária. Entretanto, em alguns casos, outros critérios de escalonamento podem ser adotados. Após a etapa de escalonamento, devemos escolher o método de modelagem que será utilizado. Conforme for conveniente, podemos selecionar qualquer um dos métodos revisados no Capítulo 2, como os métodos *MLR*, *PLS*, *QPLS* e *BTPLS*, por exemplo. No caso de optarmos por uma técnica de redução de dimensionalidade, devemos conduzir um estudo do número de componentes a serem incluídos no modelo.

Após a determinação da técnica de modelagem mais apropriada para o estudo em questão, é proposto que a questão da seleção de variáveis seja tratada. Ou seja, devemos verificar se todas as variáveis explicativas presentes na matriz ***X*** devem participar da composição do modelo final. Neste trabalho, a questão da seleção de variáveis será tratada pelo método *SRMP*, proposto no Capítulo 3. Finalmente, para que se tenha conhecimento da confiabilidade das predições fornecidas pelo modelo, é importante que uma análise estatística seja conduzida. Se houver disponibilidade de informações a respeito da variabilidade das variáveis, podemos conduzir a análise estatística pelo método da adição de erro, proposto no Capítulo 4. Se este não for o caso, devemos recorrer a alguma das técnicas utilizadas no estudo comparativo realizado no Capítulo 3, como a técnica *bootstrapping residuals*, por exemplo.

Finalizada a análise estatística, o modelo final obtido é capaz de fornecer predições com confiabilidade definida para os valores das variáveis de resposta de novas amostras do sistema modelado.

5.2. Caso 1: Dados da Planta de Processamento Mineral

Neste exemplo, vamos estudar os dados oriundos de uma planta de processamento mineral. As variáveis de entrada (***X***) são do tipo vazões, concentrações, pressão e pH, enquanto as variáveis de saída (***Y***) se referem a medidas de qualidade determinadas no limpador como, por exemplo, vazão de concentrado, teor de chumbo e teor de cobre. O conjunto de dados completo, que pode ser obtido em Höskuldsson (1996), é constituído por 291 amostras, para as quais dispomos de medidas de doze variáveis explicativas e de dez variáveis de resposta. Höskuldsson utilizou estes dados para comparar diferentes critérios para a determinação do número ótimo de direções a serem utilizadas em um modelo *PLS*. Uma questão importante levantada por este autor foi se, neste caso, devemos realizar a decomposição dos dados utilizando todas as variáveis de resposta simultaneamente ou se devemos realizar dez decomposições diferentes, uma para cada variável de resposta. No seu estudo, Höskuldsson concluiu que, em se tratando de dados desta natureza, raramente podemos esperar que um mesmo conjunto de componentes seja adequado para descrever todas as variáveis de resposta e, portanto, a abordagem mais adequada para o problema consiste em estudar cada uma das saídas separadamente. Então, para ser objetivo, o autor utilizou apenas y_4, a variável que apresenta maior variabilidade entre as dez respostas, para a condução de seu estudo, sugerindo que a mesma análise poderia ser realizada para as demais variáveis.

Tabela 5.1: Conjunto de dados do exemplo da planta de processamento mineral.

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	y
20,60	22,34	460,1	3,402	0,247	2,540	4,000	57,96	9,990	10,01	69,69	57,81	442,2
20,62	22,82	470,5	3,249	0,247	2,550	4,000	49,81	9,980	9,990	69,71	57,51	455,2
22,74	23,44	533,1	3,104	0,247	2,540	4,010	63,96	9,960	10,00	69,73	57,63	519,3
23,69	22,30	528,2	3,162	0,247	2,540	3,990	68,1	9,980	10,00	69,71	57,6	508,8
21,76	22,83	496,7	2,969	0,247	2,550	4,000	76,81	10,01	10,00	69,7	57,2	475,0
21,39	23,67	506,3	2,585	0,247	2,540	4,010	36,84	10,07	9,990	69,71	57,12	489,9
25,36	17,53	444,5	2,979	0,247	2,550	4,000	85,81	10,12	10,04	69,71	57,86	405,9
26,29	15,62	410,6	3,006	0,247	2,540	4,010	141,7	9,870	10,00	69,71	56,83	350,2
25,30	19,36	489,7	2,469	0,247	2,550	3,990	47,5	10,03	9,950	69,69	55,52	457,4
25,28	19,84	501,6	2,452	0,247	2,540	3,990	80,99	10,11	10,04	69,66	56,94	472,3
23,45	18,63	436,9	2,317	0,247	2,540	4,000	75,07	10,07	9,990	69,65	57,04	401,8
22,43	18,94	424,9	2,181	0,247	2,540	4,000	69,82	9,970	9,990	69,66	57,26	394,6
19,01	21,22	403,4	2,261	0,247	2,550	4,000	83,74	9,980	9,990	69,63	57,21	387,2
20,64	18,60	383,8	2,278	0,247	2,540	4,010	79,68	10,02	9,980	69,65	57,36	360,6
25,16	15,54	390,9	2,105	0,247	2,540	4,000	71,32	9,990	10,03	69,64	57,56	351,3
24,38	15,99	389,8	2,126	0,247	2,540	4,000	62,15	10,00	9,990	69,64	57,64	352,5
24,46	16,76	409,9	2,120	0,247	2,550	4,000	64,36	10,01	10,00	69,64	57,09	376,2
27,14	16,41	445,5	2,101	0,247	2,540	4,000	62,55	9,990	10,03	69,62	57,32	405,5
27,68	20,57	535,6	2,399	0,255	2,766	4,357	102,6	9,972	9,972	69,50	57,76	508,8
7,474	6,23	70,52	0,567	0,035	0,599	0,929	46,90	0,078	0,074	0,1279	1,432	74,23

Neste trabalho, os dados em questão serão utilizados para ilustrar a ferramenta desenvolvida para a construção de modelos empíricos. Pelas razões previamente apresentadas, nos limitaremos à modelagem da variável de resposta y_4 que, a partir de então, passará a ser referenciada simplesmente como y. A Tabela 5.1 apresenta parte do conjunto de dados que

iremos utilizar. Como o conjunto completo é relativamente grande, são apresentadas apenas as vinte primeiras linhas do conjunto original, que permitem que se tenha uma boa idéia da ordem da grandeza e da variabilidade de cada uma das variáveis que iremos utilizar. Nas duas últimas linhas da Tabela 5.1, são apresentadas a média e o desvio padrão das variáveis calculadas a partir das 291 amostras. Como o número de amostras é relativamente alto, o conjunto original foi dividido aleatoriamente em dois subconjuntos: o conjunto de treino, com 260 observações e o conjunto de teste, com 31 observações. Vamos então nos fazer valer das amostras do conjunto de treino para, através da ferramenta desenvolvida, obter um modelo que relacione $\boldsymbol{y}$ com $\boldsymbol{X}$ e, posteriormente, avaliar a capacidade do mesmo no que se refere ao fornecimento de predições precisas para o conjunto de teste.

Seguindo o fluxograma apresentado na seção anterior, a primeira questão que surge é o escalonamento das variáveis. A princípio, as colunas deveriam ser centradas em suas médias e divididas pelos respectivos desvios padrões. Mas, como foi observado por Höskuldsson (1996), algumas das variáveis, principalmente $\boldsymbol{x_5}$ e $\boldsymbol{x_{10}}$, apresentam um desvio padrão muito baixo, o que torna arriscado o escalonamento para a variância unitária. Embora as diferentes variáveis de entrada apresentem valores com grandezas relativamente diferentes, Höskuldsson preferiu apenas centrar os dados na média e não escaloná-los para a variância unitária. Como acreditamos que a diferença entre as grandezas dos valores das saídas não é grande o suficiente para ocasionar problemas de escala nas computações, vamos escalonar os dados da mesma maneira que este autor. Como iremos utilizar o método *PLS* linear para determinar a relação entre $\boldsymbol{X}$ e $\boldsymbol{y}$, após o escalonamento dos dados, deve ser conduzida uma análise visando a identificação do número de componentes a serem incluídos no modelo. A Tabela 5.2 apresenta a percentagem da variabilidade de $\boldsymbol{X}$, $\boldsymbol{y}$ e $\boldsymbol{b}$ utilizada por cada componente do modelo.

Tabela 5.2: Variabilidade relativa e acumulada de $\boldsymbol{X}$, $\boldsymbol{y}$ e $\boldsymbol{b}$ em cada etapa da decomposição.

comp	dX	X	dy	y	db	b
1	68,55	68,55	94,91	94,91	0,000	0,000
2	26,67	95,22	0,653	95,56	0,000	0,000
3	4,617	99,83	2,891	98,45	0,000	0,000
4	0,031	99,87	0,026	98,48	0,005	0,006
5	0,100	99,97	0,004	98,48	0,004	0,010
6	0,031	100,0	0,006	98,49	0,015	0,024
7	0,003	100,0	0,006	98,49	0,292	0,316
8	0,000	100,0	0,024	98,52	42,16	42,47
9	0,000	100,0	0,006	98,52	50,62	93,09
10	0,000	100,0	0,000	98,52	3,709	96,80

É notável o fato de que os primeiros componentes contribuem muito pouco para a variabilidade total de $\boldsymbol{b}$. A explicação para esta observação está no fato de que a solução se torna instável quando muitos componentes são adicionados ao modelo. Esta é uma indicação clara de que a utilização do método *MLR* não é adequada nesta situação, uma vez que a solução do método de mínimos quadrados é equivalente a solução do método *PLS* com k componentes. Como podemos observar, a maior parte da variabilidade de $\boldsymbol{y}$ é explicada

quando são utilizados três componentes. O mesmo pode ser afirmado a respeito da variabilidade de ***X***. Vamos, portanto, trabalhar com três componentes no modelo.

Conforme mencionado na primeira seção deste capítulo, a seleção de variáveis será conduzida pelo método *SRMP*, proposto anteriormente. A Tabela 5.3 apresenta o valor da *PRESS* (valores em relação aos dados centrados na média) obtido nas diferentes etapas do procedimento *SRMP*, que também é plotado na Figura 5.2. Neste exemplo, a *PRESS* foi calculada de forma idêntica à explicada no Capítulo 3. Na Tabela 5.3, são apresentados também o desvio padrão dos cem valores utilizados no cálculo da *PRESS* e a significância da hipótese de a *PRESS* obtida na etapa atual ser maior do que o valor mínimo obtido entre todas as etapas anteriores.

Tabela 5.3: Sumário dos resultados do procedimento *SRMP*.

var	PRESS	desvio	SIG
3	282.0	76.48	-
1	106.37	27.06	0.00
2	90.68	31.57	0.40
8	86.95	26.57	26.1
11	86.23	26.01	44.6
9	87.15	27.37	56.9
4	86.88	29.19	54.7
10	84.67	26.43	38.3
7	85.45	25.63	56.0
12	87.91	27.51	72.6
5	83.87	28.47	44.2
6	93.39	28.19	95.3

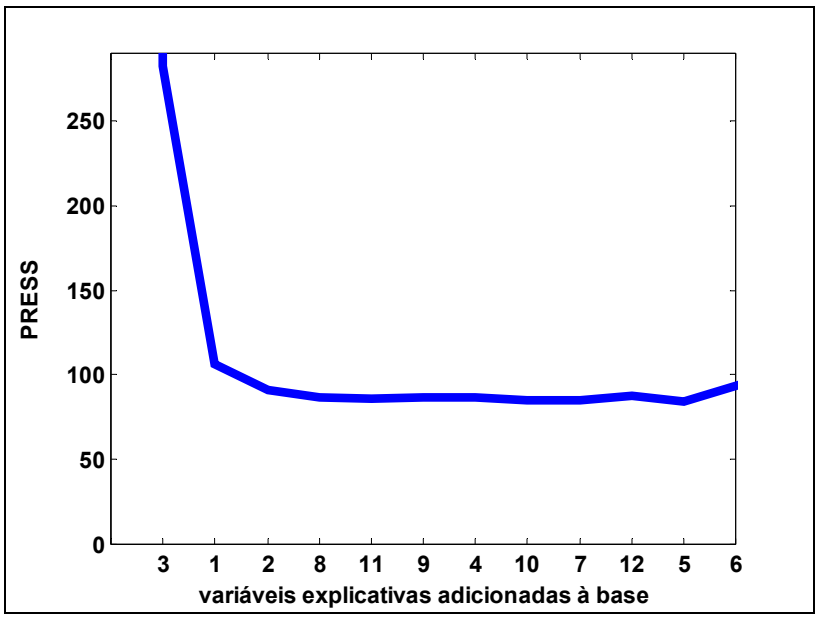

Figura 5.2: *PRESS* em função das variáveis adicionadas ao modelo.

Como podemos observar, os índices avaliados não nos permitem afirmar que alguma das variáveis tenha contribuído negativamente para a capacidade preditiva do modelo e,

portanto, todas as variáveis serão incluídas na base. Ou seja, em nenhuma das etapas do procedimento *SRMP* o valor da *PRESS* se mostrou maior que o valor mínimo verificado em todas as etapas anteriores com 99% de significância. Portanto, o modelo final será um modelo *PLS* com três componentes, utilizando todas as variáveis explicativas presentes no conjunto de dados original.

Na Figura 5.3, os valores da variável de resposta preditos pelo modelo final são plotados contra os valores experimentais para os conjuntos de treino e teste. Como podemos observar, o modelo se mostra adequado tanto no que se refere ao ajuste dos dados do conjunto de treino quanto no que diz respeito a predição dos valores da resposta do conjunto de teste, o que comprova a sua validade.

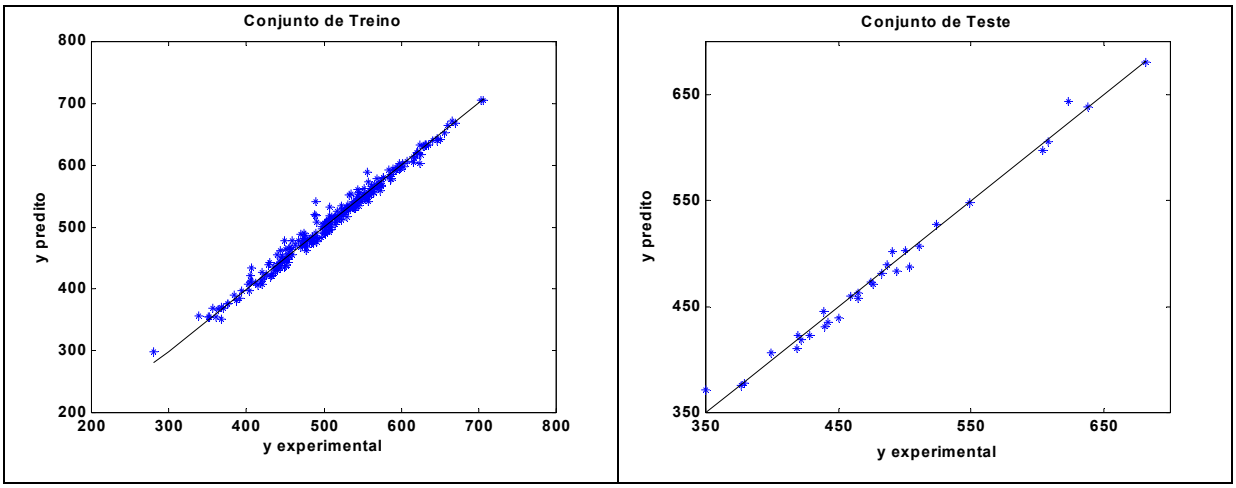

Figura 5.3: Predições do modelo final para os conjuntos de treino e teste

Tabela 5.4: Erros nos coeficientes do modelo *PLS* com 3, 4 e 12 componentes.

	3 direções		4 direções		12 direções	
b_1	-12.46	0.552	-12.99	1.912	-12.16	1.274
b_2	13.34	0.596	13.25	2.596	14.10	1.804
b_3	1.111	0.016	1.112	0.018	1.105	0.019
b_4	-0.428	0.575	55.03	15.03	182.3	186.7
b_5	-1.872	0.595	-22.48	17.73	5130	42416
b_6	1.942	0.637	-53.58	20.36	620.0	2674
b_7	1.962	0.636	-54.72	20.37	-324.1	1127.9
b_8	-0.052	0.019	-0.061	0.020	-0.064	0.021
b_9	1.192	0.597	80.00	13.47	16077	7302
b_{10}	1.377	0.679	1.815	17.06	-1806	9673
b_{11}	-2.299	0.565	-15.98	19.30	-2416	2611
b_{12}	3.209	0.622	-6.943	13.91	25.48	27.01
b_0	-2E-13	0.530	-1E-12	0.580	-3E-10	0.536

Por fim, vamos utilizar o método *bootstrapping residuals* para estimar as incertezas presentes nos coeficientes do modelo. A Tabela 5.4 apresenta o valor dos coeficientes do

modelo *PLS* com três, quatro e doze direções e as respectivas estimativas para o desvio padrão dos mesmos (em relação aos dados centrados na média). Como podemos observar, o modelo com três componentes é de um modo geral consideravelmente mais estável mais estável do que o modelo com quatro direções e muito mais estável que o modelo *MLR* (equivalente ao modelo de doze direções).

Então, o modelo composto por três componentes utiliza 99% de $\boldsymbol{X}$ e é capaz de explicar mais de 98% da variabilidade de $\boldsymbol{y}$. Para dados de processos desta natureza, este resultado pode ser considerado satisfatório. No que se refere a aplicação da ferramenta desenvolvida para a construção do modelo, neste caso em particular, os índices avaliados permitiram visualizar de maneira nítida que todas as variáveis devem ser incluídas na base e que três componentes devem ser utilizados.

5.3. Caso 2: Dados da Indústria Tabagista

Neste exemplo, vamos estudar os dados provenientes da análise de 25 amostras de folhas de tabaco. O objetivo do estudo é a determinação da relação existente entre a fração mássica de alguns constituintes químicos das folhas ($\boldsymbol{X}$) e a taxa de queima dos cigarros produzidos a partir das mesmas ($\boldsymbol{y}$), em polegadas por 1000 segundos. Os dados de $\boldsymbol{X}$ e $\boldsymbol{y}$, apresentados na Tabela 5.6, também foram retirados de Höskuldsson (1996). As variáveis $\boldsymbol{x_1}$, $\boldsymbol{x_2}$, ..., $\boldsymbol{x_6}$ são, respectivamente, o teor percentual em massa de nitrogênio, cloro, potássio, fósforo, cálcio e magnésio presente nas folhas.

Vamos então escalonar os dados de modo que todas as variáveis envolvidas passem a apresentar média nula e variância unitária. Posteriormente, vamos conduzir a determinação do número ótimo de componentes a serem utilizados no modelo *PLS* na situação onde todas as entradas são incluídas no modelo final. A Tabela 5.5 apresenta a percentagem da variabilidade de $\boldsymbol{X}$, $\boldsymbol{y}$ e $\boldsymbol{b}$ utilizada por cada componente do modelo.

Tabela 5.5: Variabilidade relativa e acumulada de $\boldsymbol{X}$, $\boldsymbol{y}$ e $\boldsymbol{b}$ em cada etapa da decomposição.

comp	dX	X	dy	y	db	b
1	29.99	29.99	57.89	57.89	44.80	44.80
2	34.74	64.73	7.714	65.60	17.49	62.29
3	11.90	76.62	2.605	68.21	15.49	77.78
4	8.876	85.50	1.384	69.59	17.69	95.47
5	8.870	94.37	0.087	69.68	1.901	97.37
6	5.629	100.0	0.075	69.75	2.633	100.0

Podemos observar que, ao adicionarmos o terceiro componente no modelo, cerca de 68% da variabilidade da variável de resposta é explicada. Como a inclusão de novos componentes ao modelo não foi capaz de aumentar consideravelmente este valor, vamos continuar a análise utilizando apenas três componentes. Vamos então, seguindo o fluxograma apresentado na primeira seção deste capítulo, passar para a etapa de seleção de variáveis. Como já foi mencionado, a escolha das variáveis explicativas que devem permanecer no modelo final será conduzida pelo método *SRMP*. A Tabela 5.7 apresenta o valor da *PRESS*

(valores em relação aos dados escalonados) obtido nas diferentes etapas do procedimento *SRMP*, que também é plotado na Figura 5.4. Na Tabela 5.7, são apresentados também o desvio padrão dos cem valores utilizados no cálculo da *PRESS* e a significância da hipótese de a *PRESS* obtida na etapa atual ser maior do que o valor mínimo obtido entre todas as etapas anteriores.

Tabela 5.6: Conjunto de dados do exemplo da indústria tabagista.

x_1	x_2	x_3	x_4	x_5	x_6	y
-0.495	0.692	-0.273	0.610	-0.309	-0.428	-0.827
1.677	0.494	-2.026	0.305	2.484	2.265	-0.347
-0.278	0.329	0.623	-1.830	-0.182	-1.140	-0.168
0.156	1.138	-0.701	0.915	0.250	0.048	-1.006
0.808	0.065	-0.234	-2.135	1.062	1.235	0.072
-0.459	0.131	1.285	-1.525	-2.035	-1.140	-0.048
2.582	0.312	1.558	0.305	0.834	0.760	0.551
-1.002	0.164	-0.078	0.000	-0.029	0.364	-0.707
-0.821	-0.364	-0.351	2.135	-0.055	-0.348	-0.527
1.496	-1.223	-2.337	0.610	2.001	2.027	-1.006
-0.749	-0.546	0.935	-0.305	-0.791	-1.219	-0.048
-0.459	-0.794	0.545	0.305	-0.715	0.127	0.311
1.242	-0.678	0.312	-0.305	0.326	0.602	1.449
-1.581	-0.397	0.039	0.915	-1.248	-1.457	0.491
1.351	-1.223	1.558	0.305	-0.283	-0.269	1.509
-0.930	-1.685	-1.052	-0.915	-0.283	-0.507	0.851
0.084	-2.873	0.857	-0.305	-1.096	-0.823	2.408
0.012	0.593	0.467	0.000	0.224	-0.111	0.192
-0.061	1.352	-0.779	-0.305	-0.791	0.760	-1.186
-0.640	0.692	-0.312	-0.305	-0.080	-0.982	-1.006
-0.966	0.560	-0.779	1.220	-0.080	0.443	-0.287
-0.314	0.510	-0.351	0.915	-0.258	0.602	-1.725
-0.278	1.088	1.402	0.305	-0.740	-1.299	0.551
0.193	0.543	-0.234	-1.525	1.443	-0.348	1.449
-0.568	1.121	-0.078	0.610	0.351	0.839	-0.946

Como podemos observar, novamente, os índices avaliados não nos permitem afirmar que alguma das variáveis tenha contribuído negativamente para a capacidade preditiva do modelo e, portanto, todas as variáveis serão incluídas na base. Uma análise superficial da curva plotada na Figura 5.4 pode passar a falsa impressão de que a *PRESS* passa a crescer após a adição da variável x_2 ao modelo. Entretanto, deve ser lembrado que a computação da *PRESS* é uma variável aleatória com média e desvio padrão dados na Tabela 5.7. Podemos verificar que, em todas as etapas do procediemnto *SRMP*, o desvio padrão das cem computações da *PRESS* é relativamente alto quando comparado com a respectiva média. Na verdade, podemos notar que os valores computados para a *PRESS* a partir da segunda etapa parecem pertencer à mesma distribuição, o que é comprovado pelo teste de significância da diferença entre as médias. Devido a esta alta variabilidade, obviamente, os valores médios computados podem apresentar uma visível diferença e, como o procedimento *SRMP* tende a selecionar as variáveis em ordem crescente de *PRESS*, podemos ter a impressão de que a *PRESS* está aumentando. Mas, como comprova o teste de significância, tal afirmativa não

pode ser sustentada. Desta forma, o modelo final, composto por três componentes, utilizará todas as variáveis explicativas presentes no conjunto de dados original.

Tabela 5.7: Sumário dos resultados do procedimento *SRMP*.

var	PRESS	desvio	SIG
3	0.024	0.013	-
2	0.014	0.010	0.002
4	0.015	0.008	60.48
6	0.016	0.008	82.45
5	0.016	0.009	83.10
1	0.016	0.009	84.71

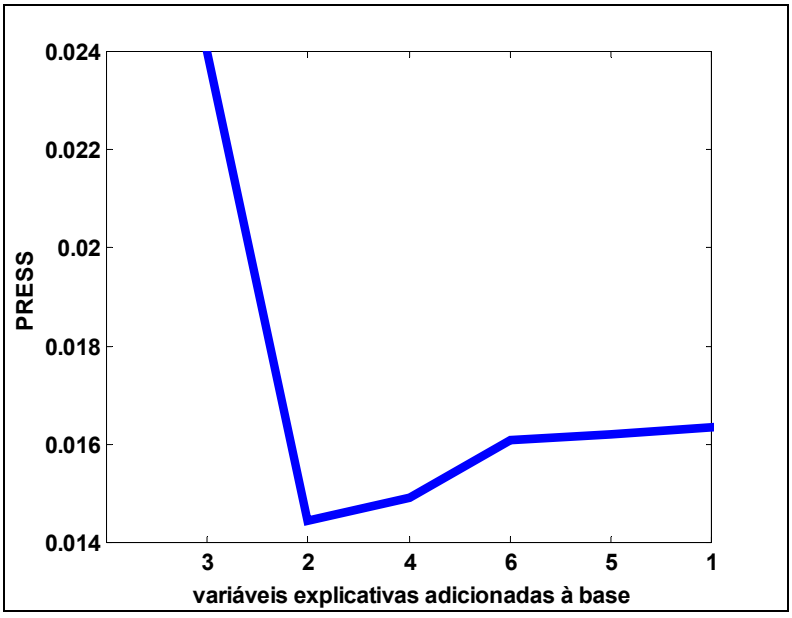

Figura 5.4: *PRESS* em função das variáveis adicionadas ao modelo.

Na Figura 5.5, os valores da variável de resposta preditos pelo modelo final são plotados contra os valor experimental. Como o número de amostras disponíveis é baixo, não foi separado um conjunto de teste para validar o modelo. Embora a análise de validação não tenha sido conduzida, podemos ter uma boa idéia da capacidade preditiva do modelo através dos dados da Tabela 5.7.

Vamos então, utilizando o método *bootstrapping residuals*, estimar as incertezas presentes nos coeficientes do modelo. A Tabela 5.8 apresenta o valor dos coeficientes do modelo *PLS* com três, quatro e seis direções e as respectivas estimativas para o desvio padrão dos mesmos (em relação aos dados escalonados).

Como podemos observar, os modelos com três e quatro componentes são bastante parecidos em termos de estabilidade de modo que é difícil a obtenção de uma conclusão definitiva. Contudo, é fato que ambos são consideravelmente mais estáveis que o modelo de mínimos quadrados (equivalente ao modelo de seis direções), o que indica que a utilização de métodos de redução de dimensionalidade também é vantajosa neste caso. Resumindo, o

modelo composto por três componentes utiliza 76% de ***X*** e é capaz de explicar mais de 68% da variabilidade de ***y***, enquanto o modelo com quatro componentes utiliza 85% de ***X*** e é capaz de explicar 69% de ***y***. Como a diferença entre a variabilidade de ***y*** explicada nos dois casos é muito pequena, vamos considerar que o modelo final deve utilizar três componentes.

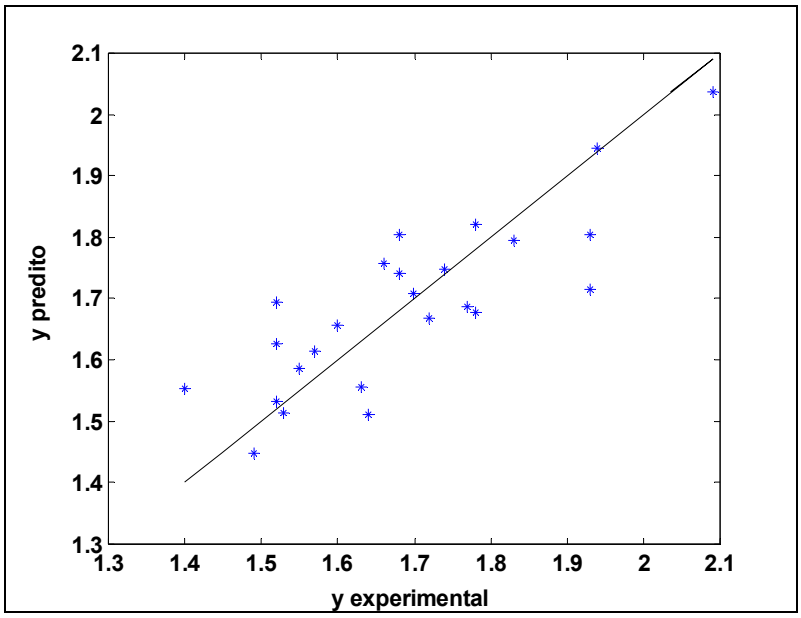

Figura 5.5: Predições do modelo para as amostras disponíveis.

Tabela 5.8: Erros nos coeficientes do modelo *PLS* com 3, 4 e 6 componentes.

	3 direções		4 direções		6 direções	
b_1	0.210	0.140	0.178	0.178	0.103	0.263
b_2	-0.605	0.114	-0.563	0.128	-0.581	0.137
b_3	0.335	0.139	0.389	0.154	0.449	0.245
b_4	-0.093	0.143	-0.121	0.125	-0.129	0.136
b_5	0.201	0.130	0.368	0.195	0.408	0.233
b_6	-0.284	0.136	-0.387	0.201	-0.324	0.238
b_0	0.000	0.109	0.000	0.115	0.000	0.122

5.4. Caso 3: Dados da Indústria de Alimentos

Neste exemplo, vamos estudar novamente os dados referentes ao teor de proteínas presente em 24 amostras de trigo moído. O conjunto de dados, retirado de Shacham e Brauner (2003), é o mesmo que utilizamos no estudo da determinação de incertezas em coeficientes de regressão (Capítulo 4). O objetivo da análise é a obtenção de um modelo que relacione o teor de proteínas presente nas amostras com o espectro infravermelho das mesmas. A utilização de modelos que utilizam o espectro infravermelho como entrada é de grande importância na indústria de alimentos, uma vez que os mesmos permitem que medidas referentes à qualidade dos produtos sejam feitas *on line*, durante o processo de produção. Porém, dados relacionados à espectroscopia são difíceis de serem trabalhados. Normalmente, há um alto grau de

colinearidade entre as variáveis explicativas e, em alguns casos, o número de variáveis de entrada pode ser várias vezes maior que o número de amostras, o que faz com que a utilização de técnicas de seleção de variáveis seja importante.

O conjunto de dados que será utilizado é apresentado na Tabela 5.9, onde são fornecidos os valores da reflectância da luz infravermelha para seis comprimentos de onda na faixa de 1380 a 2610 nm e as respectivas medidas experimentais do teor de proteína contido nas amostras de trigo, determinadas pelo método de Kjeldahl. Ao contrário dos exemplos estudados até então, neste caso, dispomos de informações que podem fornecer boas aproximações para o valor do erro experimental presente nos dados. Segundo Nicolaas e Faber (2002), o método de Kjeldahl apresenta um erro de aproximadamente 0,2% e, sendo assim, vamos assumir que esta é a incerteza presente nas medidas de $\boldsymbol{y}$. Estes mesmos autores afirmam, ainda, que um erro de 0,25% pode ser considerado como uma hipótese conservativa na determinação do espectro infravermelho de uma amostra, portanto, neste estudo, será assumido que este é o erro presente nas medidas de $\boldsymbol{X}$. Estas informações são interessantes, pois permitem que o método de reamostragem proposto para a estimação da incerteza presente nos coeficiente do modelo (método da adição de erro) seja utilizado.

Tabela 5.9: Espectro infravermelho e teor de proteínas para as 24 amostras de trigo.

x_1	x_2	x_3	x_4	x_5	x_6	y
468.0	123.0	246.0	374.0	386.0	-11.00	9.230
458.0	112.0	236.0	368.0	383.0	-15.00	8.010
457.0	118.0	240.0	359.0	353.0	-16.00	10.95
450.0	115.0	236.0	352.0	340.0	-15.00	11.67
464.0	119.0	243.0	366.0	371.0	-16.00	10.41
499.0	147.0	273.0	404.0	433.0	5.000	9.510
463.0	119.0	242.0	370.0	377.0	-12.00	8.670
462.0	115.0	238.0	370.0	353.0	-13.00	7.750
488.0	134.0	258.0	393.0	377.0	-5.000	8.050
483.0	141.0	264.0	384.0	398.0	-2.000	11.39
463.0	120.0	243.0	367.0	378.0	-13.00	9.950
456.0	111.0	233.0	365.0	365.0	-15.00	8.250
512.0	161.0	288.0	415.0	443.0	12.00	10.57
518.0	167.0	293.0	421.0	450.0	19.00	10.23
552.0	197.0	324.0	448.0	467.0	32.00	11.87
497.0	146.0	271.0	407.0	451.0	11.00	8.090
592.0	229.0	360.0	484.0	524.0	51.00	12.55
501.0	150.0	274.0	406.0	407.0	11.00	8.380
483.0	137.0	260.0	385.0	374.0	-3.000	9.640
491.0	147.0	269.0	389.0	391.0	1.000	11.35
463.0	121.0	242.0	366.0	353.0	-13.00	9.700
507.0	159.0	285.0	410.0	445.0	13.00	10.75
474.0	132.0	255.0	376.0	383.0	-7.000	10.75
496.0	152.0	276.0	396.0	404.0	6.000	11.47

Após escalonar os dados de modo que as variáveis passem a apresentar média nula e variância unitária, vamos conduzir a determinação do número ótimo de componentes a serem utilizados no modelo *PLS* na situação onde todas as entradas são incluídas no modelo final. A

Tabela 5.10 apresenta a percentagem da variabilidade de ***X***, ***y*** e ***b*** utilizada por cada componente do modelo.

Tabela 5.10: Variabilidade relativa e acumulada de ***X***, ***y*** e ***b*** em cada etapa da decomposição.

comp	dX	X	dy	y	db	b
1	97.77	97.77	22.46	22.46	0.074	0.074
2	1.552	99.33	40.317	62.78	11.55	11.62
3	0.462	99.79	34.997	97.77	52.23	63.85
4	0.205	99.99	0.132	97.91	0.533	64.39
5	0.003	100.0	0.296	98.20	31.51	95.89
6	0.003	100.0	0.013	98.21	4.109	100.0

Podemos observar que, ao adicionarmos o terceiro componente no modelo, praticamente toda a variabilidade presente em ***X*** é utilizada e mais de 97% da variabilidade de ***y*** é explicada. A adição de novos componentes ao modelo não é capaz de aumentar consideravelmente este valor e, portanto, vamos conduzir a etapa de seleção de variáveis através do método *SRMP*, utilizando o modelo *PLS* com três direções. Tabela 5.11 apresenta o valor da *PRESS* (valores em relação aos dados escalonados) obtido nas diferentes etapas do procedimento *SRMP*, que também é plotado na Figura 5.6. Na Tabela 5.11, são apresentados também o desvio padrão dos, neste caso, quinhentos valores utilizados no cálculo da *PRESS* e a significância da hipótese de a *PRESS* obtida na etapa atual ser maior do que o valor mínimo obtido entre todas as etapas anteriores.

Tabela 5.11: Sumário dos resultados do procedimento *SRMP*.

var	PRESS	desvio	SIG
2	0.757	0.406	-
4	0.076	0.051	0.000
3	0.034	0.017	0.000
1	0.045	0.028	100.0
6	0.053	0.029	100.0
5	0.053	0.053	100.0

Os índices avaliados na Tabela 5.11 nos mostram nitidamente que a adição das variáveis $\boldsymbol{x_2}$, $\boldsymbol{x_3}$ e $\boldsymbol{x_4}$ contribuem de maneira significativa para a capacidade preditiva do modelo *PLS* com três componentes. Por outro lado, também é claro o fato de que a adição das variáveis $\boldsymbol{x_1}$, $\boldsymbol{x_5}$ e $\boldsymbol{x_6}$ ao modelo faz com que o valor da *PRESS* aumente significativamente. Portanto, o modelo final será um modelo *PLS* com três componentes, utilizando as variáveis explicativas $\boldsymbol{x_2}$, $\boldsymbol{x_3}$ e $\boldsymbol{x_4}$.

Substituindo os valores estimados para os coeficientes na expressão do modelo linear, o modelo final (em termos das variáveis não escalonadas) é dado por: $\boldsymbol{y} = -0.110\boldsymbol{x_2} + 0.355\boldsymbol{x_3} - 0.229\boldsymbol{x_4} + 20.90$. É interessante notarmos que, como estamos utilizando apenas três variáveis explicativas, o modelo final é, na verdade, equivalente ao

modelo que seria obtido pela aplicação direta do método dos mínimos quadrados. Na Figura 5.7, os valores da variável de resposta preditos pelo modelo final são plotados contra os valores experimentais. Como podemos observar, o modelo com três componentes fornece um bom ajuste aos dados. Novamente, devido à não disponibilidade de amostras, não foi separado um conjunto de teste para a validação do modelo. Porém, de forma similar ao exemplo anterior, podemos ter uma boa idéia da capacidade preditiva do modelo através dos dados da Tabela 5.11

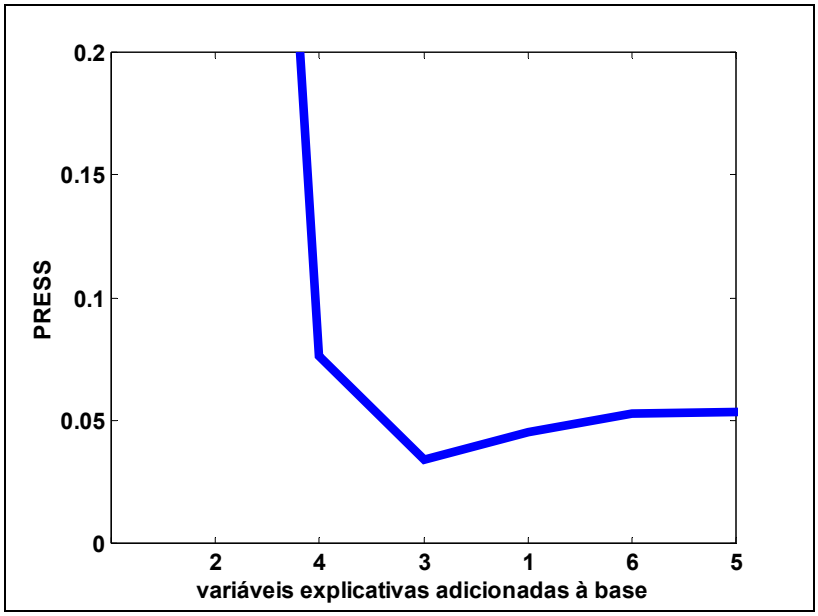

Figura 5.6: *PRESS* em função das variáveis adicionadas ao modelo.

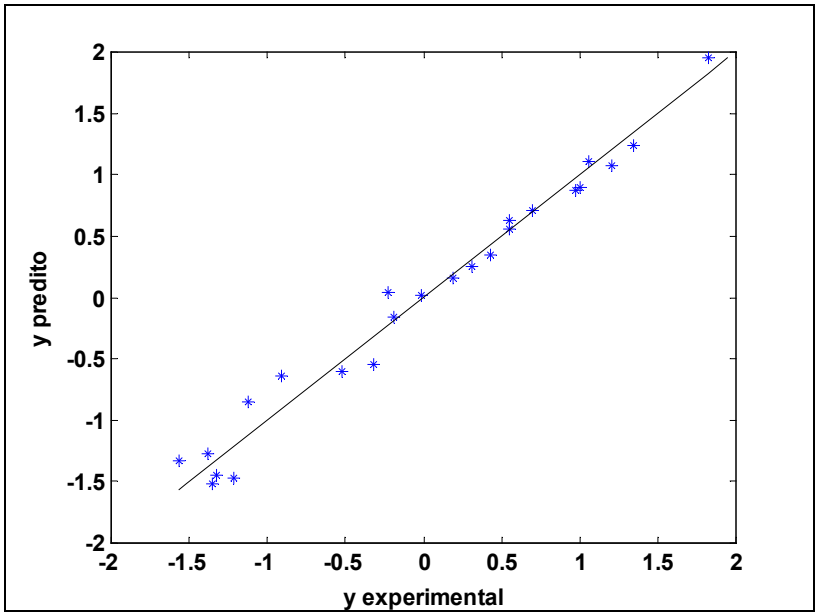

Figura 5.7: Predições do modelo para as amostras disponíveis.

Como foi mencionado no início desta seção, aproximações para o desvio padrão de cada um dos coeficientes do modelo podem ser obtidas pelo método da adição de erro. Considerando que o erro padrão associado à medida de $\boldsymbol{X}$ e $\boldsymbol{y}$ são de, respectivamente, 0.20 e 0.25%, foram obtidas as seguintes aproximações para o desvio padrão dos coeficientes: S_{b0} =

5.823, S_{b2} = 0.060, S_{b3} = 0.062 e S_{b4} = 0.01. Convém lembrar que a confiabilidade destes valores está diretamente relacionada a confiabilidade das estimativas para o erro de medida das variáveis utilizados em suas computações. Como, neste caso, acreditamos que a incerteza na determinação das variáveis explicativas está superestimada, provavelmente, as aproximações para o desvio padrão dos coeficientes também estão.

5.5. Caso 4: Dados da Indústria de Cosméticos

Neste exemplo, vamos estudar a relação existente entre a composição química e a qualidade de um creme de aplicação facial. O estudo será baseado em $n = 17$ formulações (amostras) do creme, que se distinguem entre si pelo teor em que os $k = 8$ constituintes químicos do produto (tais como glicerina, água, emulsificador e vaselina) estão presentes. A Tabela 5.12 apresenta a aqui chamada matriz de composição dos cremes, $\boldsymbol{X}$.

O produto final é avaliado pela análise de uma bateria de indicadores de qualidade, que são determinados em relação a um creme padrão. Os 17 cremes foram submetidos a um teste onde cada creme é aplicado em uma das metades do rosto de dez modelos enquanto, ao mesmo tempo, o creme padrão é aplicado à outra metade do rosto. Juntamente com as modelos, avaliadores treinados forneceram sua opinião a respeito de $m = 11$ indicadores de qualidade do produto (tais como facilidade de aplicação, oleosidade, maciez e brilho) em relação ao creme padrão. A Tabela 5.13 apresenta a aqui chamada matriz de qualidade dos cremes, $\boldsymbol{Y}$, que contém os valores médios computados a partir da opinião das dez modelos sobre cada um dos 17 cremes.

Estes dados foram levantados com o objetivo de se desenvolver um modelo que relacione a composição dos cremes ($\boldsymbol{X}$) com os indicadores de qualidade ($\boldsymbol{Y}$), o que permitiria a obtenção de formulações ótimas para os cremes através da escolha da composição apropriada. Neste trabalho, pretendíamos utilizar estes dados para ilustrar o caso onde a ferramenta desenvolvida é aplicada utilizando-se métodos de redução de dimensionalidade não lineares para a construção dos modelos. Entretanto, como será discutido em seguida, isso não foi possível e, novamente, o modelo linear teve de ser utilizado. A escolha deste conjunto de dados foi motivada pelo fato de que diferentes autores recorreram a este exemplo para ilustrar o desempenho de extensões não lineares para o algoritmo do método *PLS*. Por exemplo, Wold et al (1989) utilizaram estes dados para ilustrar o desempenho do algoritmo *QPLS* e, posteriormente, Li et al (2001) se fizeram valer deste mesmo exemplo para comparar os métodos *QPLS* e *BTPLS*. Wold et al (1989) verificaram que o modelo *QPLS* com apenas dois componentes é capaz de explicar praticamente a mesma parcela da variabilidade de $\boldsymbol{Y}$ que o modelo *PLS* com quatro componentes, concluindo então que, no caso do modelo linear, os dois últimos componentes simplesmente compensam a não linearidade da relação existente entre $\boldsymbol{X}$ e $\boldsymbol{Y}$. Li et al (2001) demonstraram que, para o mesmo número de componentes, a capacidade de explicar a variabilidade de $\boldsymbol{Y}$ do método *BTPLS* é visivelmente superior a dos métodos *QPLS* e *PLS*. As conclusões de ambos os autores podem ser verificadas na Tabela 5.14, que apresenta uma reprodução dos resultados obtidos no estudo comparativo de Li et al (2001). Os resultados para o método *QPLS* não são exatamente os mesmos que os encontrados por Wold et al (1989) porque Li et al (2001) utilizaram o procedimento de atualização dos vetores pesos baseados no erro (ver Capítulo 2 para detalhes).

Tabela 5.12: Composição química do creme facial para as 17 formulações avaliadas.

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
1.7301	-1.0629	0.7277	-0.6592	0.7324	-0.6409	-0.5196	1.4716
-0.8590	0.8351	-0.8685	-0.6592	-1.1656	-0.6409	-0.5196	-0.8790
0.2506	0.0759	-0.8685	-0.6592	0.7324	1.9533	1.6887	-0.8790
-0.8590	1.2147	-0.8685	-0.6592	0.7324	-0.6409	-0.5196	-0.8790
-0.8590	1.5943	1.5258	-0.6592	-0.4064	-0.6409	-0.5196	1.2198
0.9904	-1.0629	0.7277	-0.6592	1.8712	1.9533	-0.5196	0.8000
0.8794	-1.0629	0.3286	0.1272	-0.1786	1.3048	-0.5196	0.3802
1.2123	-1.0629	0.7277	-0.6592	0.7324	-0.6409	0.9526	-0.8790
-0.8590	1.2147	1.5258	-0.6592	-1.9248	-0.6409	-0.5196	0.3802
0.9904	-1.0629	0.7277	-0.6592	0.7324	1.3048	-0.5196	-0.8790
-0.8590	-1.0629	-0.8685	2.2900	-1.5452	-0.6409	-0.5196	-0.8790
0.2506	0.0759	1.5258	-0.6592	0.7324	-0.6409	-0.5196	-0.8790
-0.8590	0.9110	-0.8685	1.3069	0.7324	-0.6409	-0.5196	1.2198
-0.8590	0.9110	-0.8685	1.3069	-0.7860	-0.1221	2.4248	1.2198
1.4268	-1.0629	-0.8685	-0.6592	-0.0268	-0.6409	1.6887	-0.8790
-0.8590	0.1518	-0.8685	0.9137	-0.4064	-0.6409	-0.5196	-0.8790
-0.8590	0.4555	-0.8685	1.3069	-0.5582	0.6562	-0.5196	1.2198

Tabela 5.13: Indicadores de qualidade para as 17 formulações do creme facial avaliadas.

y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8	y_9	y_{10}	y_{11}
-0.3190	1.2460	1.9574	1.7677	1.7788	0.2649	-0.4490	0.1288	0.1475	1.8100	-0.8346
1.7056	-1.5052	-2.3426	-1.9721	0.2086	0.2649	-1.6328	-1.4750	0.1457	-1.0344	-0.8346
-0.3190	-0.6044	-1.4142	-0.3173	0.5125	-1.7190	0.1225	1.4773	-0.9038	-1.0344	1.3362
0.1148	-1.0426	-1.5608	-1.7735	-1.3870	0.2649	-0.5715	-0.5638	0.1475	0.0756	-0.6020
1.4887	0.4669	-0.1926	0.1129	0.5125	0.2649	-1.0613	-0.5638	1.1988	-0.8263	-1.3385
-1.0059	1.5382	0.2716	1.3374	0.0566	0.6403	-1.6328	0.8213	0.1475	0.0756	-0.0980
-2.5967	-2.1139	0.1250	0.7417	0.3605	-1.0756	0.4490	1.0400	0.6327	0.4919	1.1036
0.1148	0.4669	0.5648	-0.5159	-1.0577	0.6403	0.6531	-0.0899	1.1988	-0.5835	0.6385
0.8018	-0.2879	-0.0460	-0.7145	0.3605	0.2649	-0.7348	0.8213	0.1475	0.9428	-0.6020
-0.5721	-0.4340	0.7358	0.3115	-1.3870	-2.0943	0.3674	-0.7824	1.1988	-1.4506	0.6385
-0.3190	-0.4340	0.1250	-0.3173	-0.9058	-1.0756	0.9797	-1.0376	0.6327	0.2873	0.1345
0.5849	1.0756	0.7358	-0.3173	-1.5389	-0.3785	-0.8981	-1.0376	0.1475	-0.3753	-1.3385
-0.1021	-0.2879	0.4182	0.7417	0.3605	0.6403	0.8981	-0.3451	0.1475	0.4919	0.6385
0.5849	0.6373	0.2716	0.7417	0.6898	1.9808	-0.1225	0.5661	-2.3595	-1.2425	-1.5711
0.8018	0.9295	-0.3392	-0.5119	-0.2726	-0.0568	1.8369	-0.5638	-0.9038	0.9428	1.6075
-0.3913	-0.2879	0.5648	-0.2511	-0.0700	0.5330	0.6531	-0.5658	0.1475	-0.3772	0.9873
-0.5721	0.6373	0.1250	0.9403	1.7788	0.6403	1.1430	2.1699	-1.8743	1.8100	0.1345

Tabela 5.14: Variância extraída pelos componentes dos modelos *PLS*, *QPLS* e *BTPLS*.

comp	PLS				QPLS				BTPLS			
	dX	X	dY	Y	dX	X	dY	Y	dX	X	dY	Y
1	28.54	28.54	16.76	16.76	13.63	13.63	25.84	25.84	14.22	14.22	29.96	29.96
2	19.31	47.85	17.64	34.40	16.74	30.37	19.93	45.76	14.86	29.08	18.01	47.97
3	19.51	67.36	11.09	45.49	10.61	40.98	14.33	60.09	13.72	42.80	14.01	61.98
4	11.77	79.13	8.10	53.58	13.96	54.94	7.510	67.60	5.761	48.56	5.262	67.24
5	10.38	89.5	7.246	60.83	9.946	64.89	4.961	72.56	8.664	57.22	6.098	73.34
6	3.660	93.2	5.299	66.13	8.097	72.99	2.890	75.45	12.78	70.01	4.799	78.14
7	6.708	99.9	1.545	67.67	21.83	94.82	2.030	77.48	7.790	77.80	2.604	80.74
8	0.123	100.0	5.582	73.26	5.178	100.0	2.127	79.61	22.20	100.0	1.562	82.31

Li et al (2001) afirmaram que o melhor ajuste aos dados fornecido pelo modelo *BTPLS* deve-se ao fato do mesmo ser capaz de modelar adequadamente a relação existente entre as variáveis latentes $\boldsymbol{u}$*'s* e $\boldsymbol{t}$*'s*. Por exemplo, analisando o gráfico mostrado na Figura 5.8, onde a $\boldsymbol{u}_1$ é plotada contra $\boldsymbol{t}_1$, os autores afirmaram que os métodos *QPLS* e *PLS* não são capazes de

mapear adequadamente a relação existente entre estas variáveis latentes, justificando assim o fato de o modelo *BTPLS* capturar uma parcela maior da variabilidade de ***Y***.

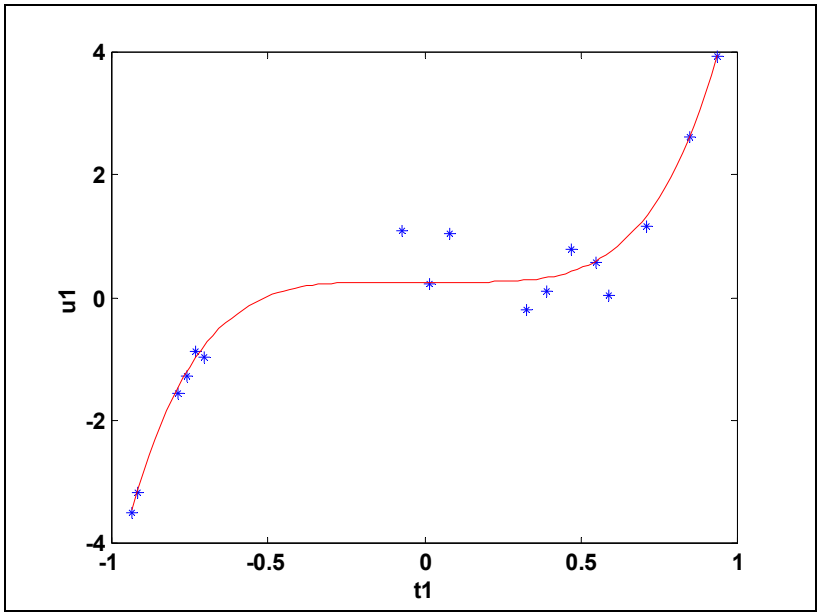

Figura 5.8: Relação entre o primeiro par de variáveis latentes pelo método *BTPLS*.

Entretanto, devemos chamar a atenção para o fato de que a variância de ***Y*** capturada não é necessariamente um indicativo da adequabilidade do modelo. Um modelo capaz de se ajustar bem aos dados utilizados em sua construção mas incapaz de fornecer predições precisas para outras amostras não é adequado. Quando esta situação é verificada, diz-se que houve a *overfit* na construção do modelo. Em casos como este, onde o número de amostras presentes no conjunto de dados é relativamente pequeno quando comparado ao número de variáveis, o risco de ocorrência de *overfit* é alto, pois o número de graus de liberdade para a estimação dos parâmetros torna-se baixo, principalmente para os métodos não lineares.

Baseando-se na análise do valor do índice $BIC = \ln(SQR) + (p\log(n))/n$ (*Bayesian Information Criterion*). Li et al (2001) afirmaram que o fato dos métodos não lineares serem capazes de explicar uma maior parcela da variabilidade de ***Y*** não é conseqüência de *overfit*. O índice *BIC* é uma função da soma de quadrados residuais (*SQR*), do número de parâmetros a serem estimados no mapeamento da relação existente entre $\boldsymbol{t_i}$ e $\boldsymbol{u_i}$ e do número de amostras disponíveis. Obviamente, quanto menor o valor do índice melhor o modelo. Este índice é penalizado pelo acréscimo do número de parâmetros a serem estimados e pelo decréscimo do número de observações experimentais disponíveis. Como, de um modo geral, o valor do índice *BIC* verificado por Li et al (2001) apresentou valores mais baixos para o modelo *BTPLS*, estes autores concluíram que a maior capacidade de ajuste deste método não se trata de *overfit*.

Porém, é importante notarmos que a computação do índice *BIC* se baseia única e exclusivamente em informações retiradas do conjunto de dados utilizado para a estimação dos parâmetros, o que levanta dúvidas quando a extrapolação das conclusões obtidas a partir do mesmo para outras amostras. Para avaliarmos esta questão, é conveniente que seja conduzido

um estudo de validação cruzada, onde separamos uma fração dos dados, deixando-a de fora da etapa de construção do modelo e, posteriormente, a utilizamos para testar a capacidade preditiva do modelo. O teste da capacidade preditiva do modelo é feito pela computação da *PRESS* (*PREdictive Sum of Squares*), que corresponde à soma dos quadrados dos desvios das predições do modelo em relação às medidas experimentais para as amostras separadas para teste. Basicamente, este procedimento é repetido um número grande de vezes, digamos cem, e a *PRESS* média é então computada. A Tabela 5.15 apresenta, para os três métodos em questão, o valor médio obtido a partir de cem computações para a *PRESS* em função do número de componentes incluídos no modelo. Como, neste caso, estamos trabalhando com onze variáveis de resposta, calculamos a *PRESS* para cada uma das variáveis de maneira idêntica à explicada e, posteriormente, computamos a média destes onze valores (é importante ressaltar que os dados de $\boldsymbol{X}$ e $\boldsymbol{Y}$ estão centrados na média e escalonados para a variância unitária, de modo que a escala na qual as variáveis de resposta foram determinadas não afeta a análise).

Tabela 5.15: Valor da *PRESS* em função do número de componentes nos modelos.

	1	2	3	4	5	6	7	8
PLS	0.1055	0.1006	0.0932	0.0906	0.0923	0.1067	0.1259	0.1417
QPLS	0.2023	0.2864	0.6823	0.7510	0.4475	0.7290	1.1955	0.8869
BTPLS	0.2975	0.2700	2.1417	28.012	1480	1018	1430	21011

Podemos notar que a capacidade preditiva do modelo linear é visivelmente superior a dos demais modelos. Isso revela que os modelos fornecidos pelos métodos *QPLS* e *BTPLS* são muito mais dependentes das amostras utilizadas na etapa de estimação, o que sugere que o melhor ajuste aos dados fornecido pelos mesmos trata-se, na verdade, de *overfit*. Quando aplicamos os métodos *QPLS* e *BTPLS* aos dados deste exemplo, o grau de *overfit* é tão alto que as predições dos modelos para amostras diferentes das pertencentes ao conjunto de treino são inaceitáveis em termos práticos. Esta afirmação pode ser comprovada pela análise da Tabela 5.16, que apresenta o desvio padrão das computações da *PRESS* para os modelos *QPLS* e *BTPLS*. As computações da *PRESS* apresentam uma variabilidade exagerada, a ponto de tornar sem sentido qualquer tentativa de interpretação de seus valores.

Tabela 5.16: Desvio padrão das estimativas para a *PRESS* apresentadas na Tabela 5.15.

	1	2	3	4	5	6	7	8
PLS	0.0260	0.0261	0.0240	0.0230	0.0276	0.0280	0.0463	0.0573
QPLS	0.3064	0.4301	3.0337	3.0943	0.6399	1.3773	2.7543	1.1990
BTPLS	0.8427	0.5112	17.907	227.87	10467	7067	10751	109889

Portanto, infelizmente, isso implica no fato de que não será possível utilizar este exemplo para testar a utilização ferramenta desenvolvida para a construção de modelos não lineares, uma vez que, neste caso, os modelos não lineares não fazem sentido. Portanto, novamente, vamos recorrer ao método *PLS* linear para modelar a relação existente entre $\boldsymbol{X}$ e $\boldsymbol{Y}$. Como os dados já estão devidamente escalonados, vamos, seguindo o fluxograma apresentado na primeira seção deste capítulo, determinar o número de direções que devem ser

utilizadas durante a etapa de seleção de variáveis. Como pode ser notado, ao contrário dos exemplos anteriores, a análise da Tabela 5.14 não fornece uma sugestão clara do número ótimo de componentes a serem utilizados. Vamos, portanto, fundamentar a nossa escolha na análise da Tabela 5.15, que fornece uma medida da capacidade preditiva do modelo *PLS* em função do número de componentes utilizados. Em concordância com as conclusões de Wold et al (1989), os dados da Tabela 5.15 indicam que o modelo com quatro componentes apresenta maior capacidade preditiva e, por isso, é o mais adequado. A Figura 5.9 apresenta os gráficos que ilustram a relação existente entre os quatro primeiros pares de variáveis latentes, incluídos no modelo. Como pode ser observado, todos parecem apresentar significância considerável, e são capazes de explicar cerca de 54% da variabilidade total de $\boldsymbol{Y}$.

Uma vez definido o número de componentes que serão utilizados no modelo, vamos efetuar o procedimento de avaliação das variáveis que devem compor o modelo final. Novamente, utilizaremos o método *SRMP* para esta tarefa. A Tabela 5.17 apresenta o valor da *PRESS* obtido nas diferentes etapas do procedimento *SRMP*. Na Tabela 5.17, também são apresentados o desvio padrão dos cem valores utilizados no cálculo da *PRESS* e a significância da hipótese de a *PRESS* obtida na etapa atual ser maior do que o valor mínimo obtido entre todas as etapas anteriores.

Tabela 5.17: Sumário dos resultados do procedimento *SRMP*.

var	PRESS	desvio	SIG
2	0.0927	0.0229	-
8	0.0819	0.0178	0.000
6	0.0809	0.0228	0.000
3	0.0780	0.0178	0.000
5	0.0825	0.0183	0.000
4	0.0852	0.0201	0.000
1	0.0872	0.0192	0.000
7	0.0985	0.0252	0.000

Como podemos observar, os índices avaliados não nos permitem afirmar que alguma das variáveis tenha contribuído negativamente para a capacidade preditiva do modelo e, portanto, todas as variáveis serão incluídas na base. Ou seja, o modelo final será um modelo *PLS* com quatro componentes, utilizando todas as variáveis explicativas presentes no conjunto de dados original.

Finalmente, seguindo o fluxograma da Figura 5.1, deveríamos determinar as incertezas presentes nos coeficientes estimados para o modelo. Estes valores não serão apresentados neste texto por uma questão de objetividade, pois, como temos oito variáveis explicativas e onze variáveis de resposta, existem oitenta e oito coeficientes presentes no modelo linear em termos das variáveis originais. De qualquer forma, como nos exemplos anteriores, estes valores podem ser facilmente calculados. Se informações a respeito da variabilidade dos dados estivessem disponíveis poderíamos utilizar o método da adição de erro, proposto nos capítulos anteriores. Como este não é o caso, a estimativa das incertezas para os coeficientes da regressão pode ser obtida pelo método *bootstrapping residuals*, por exemplo.

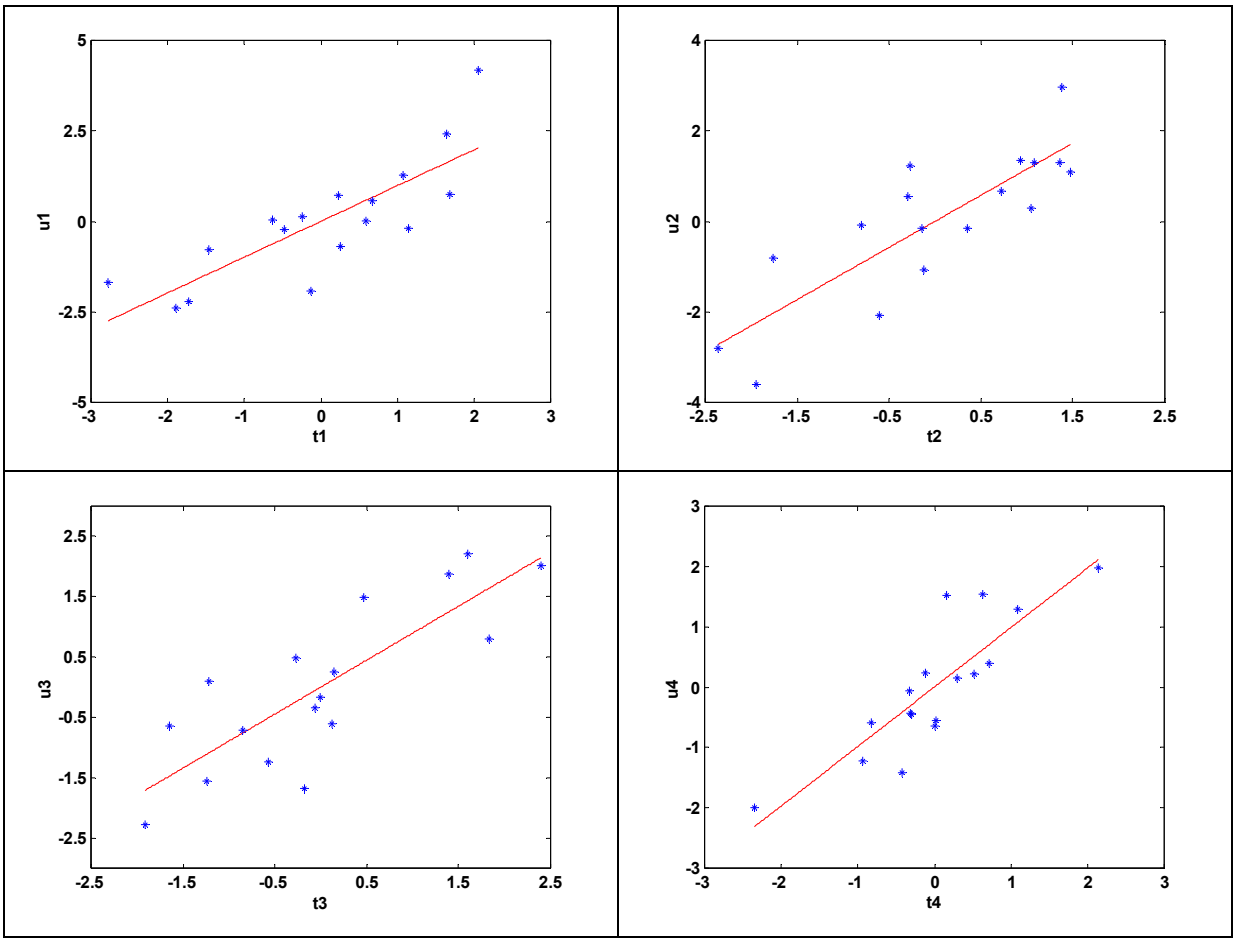

Figura 5.9: Relação entre os quatro primeiros pares de variáveis latentes do modelo *PLS*.

5.6. Caso 5: Simulação Matemática

Conforme discutido anteriormente, os dados da indústria de cosméticos, estudados na seção anterior, não permitiram a ilustração de um caso onde a ferramenta desenvolvida fosse aplicada utilizando métodos de regressão não lineares. Portanto, vamos recorrer a uma simulação matemática para analisar esta situação. Basicamente, teremos quatro variáveis explicativas, denominadas de $\boldsymbol{x_1}$, $\boldsymbol{x_2}$, $\boldsymbol{x_3}$ e $\boldsymbol{x_4}$, as quais foram geradas independentemente com valores aleatoriamente distribuídos entre –0.25 e 0.25 e uma variável de resposta, denominada de $\boldsymbol{y}$, que foi gerada a partir das entradas de acordo com a função $\boldsymbol{y} = \text{senh}(\boldsymbol{x_1} + \boldsymbol{x_2})$. Sendo assim, devemos esperar que apenas uma única combinação linear das variáveis seja necessária para descrever o comportamento da variável de resposta e, ainda, que o procedimento *SRMP* selecione apenas as variáveis $\boldsymbol{x_1}$ e $\boldsymbol{x_2}$ para compor o modelo. Além da validade do modelo final, estas são as duas questões que serão avaliadas nesta simulação. Iremos utilizar dois conjuntos de dados, um conjunto de treino e um conjunto de teste. O conjunto de treino, composto por cinqüenta "amostras", será utilizado para a construção do modelo ao passo que o conjunto de teste, composto por 25 amostras, será utilizado para a validação do mesmo. As 75 amostras foram geradas conforme descrito anteriormente e um ruído de magnitude igual a 15% foi adicionado a todas as variáveis para simular a presença de erro experimental. Os valores das variáveis $\boldsymbol{x_1}$, $\boldsymbol{x_2}$, $\boldsymbol{x_3}$, $\boldsymbol{x_4}$ e $\boldsymbol{y}$ são apresentados na Tabela 5.18 para as cinqüenta amostras do conjunto de treino.

Tabela 5.18: Conjunto treino gerado para a simulação matemática.

x_1	x_2	x_3	x_4	y	x_1	x_2	x_3	x_4	y
0,089	0,049	0,102	0,130	0,066	-0,124	-0,037	-0,071	-0,082	-0,172
0,237	-0,180	-0,174	0,094	0,098	0,238	-0,170	0,216	-0,194	0,134
0,057	0,098	0,046	-0,131	0,167	0,058	-0,175	0,040	0,029	-0,140
-0,083	-0,101	0,069	-0,187	-0,265	-0,020	-0,196	0,034	0,134	-0,251
0,058	-0,181	-0,026	0,005	-0,152	-0,093	-0,240	0,221	0,255	-0,323
0,120	-0,061	-0,072	-0,055	-0,008	0,233	0,141	0,191	0,087	0,367
-0,163	0,243	-0,230	-0,082	0,070	-0,052	-0,082	-0,016	-0,184	-0,085
0,083	-0,128	0,062	0,213	0,105	-0,161	-0,118	0,181	-0,262	-0,255
-0,113	0,034	-0,127	-0,118	-0,062	0,201	0,081	0,128	-0,107	0,195
-0,198	0,242	-0,132	0,040	0,129	-0,026	0,113	-0,062	-0,072	0,053
-0,072	0,004	0,006	0,185	-0,156	-0,146	0,225	0,018	-0,181	0,013
-0,111	0,096	-0,269	0,196	-0,012	-0,069	-0,232	-0,217	-0,037	-0,275
0,216	0,099	0,229	0,221	0,353	-0,202	0,164	0,061	-0,033	-0,096
0,299	-0,076	-0,151	-0,181	0,097	0,197	0,006	0,059	0,077	0,236
0,004	0,166	-0,170	-0,019	0,296	-0,214	-0,086	0,129	-0,014	-0,279
-0,022	0,072	-0,095	0,094	0,225	0,097	0,031	-0,126	0,044	0,067
0,056	-0,209	-0,166	-0,083	-0,186	-0,018	-0,142	0,011	-0,012	-0,156
-0,029	0,134	0,098	-0,034	0,018	-0,073	0,161	-0,139	0,188	0,099
-0,122	-0,259	0,215	0,166	-0,259	0,194	-0,085	-0,216	0,219	0,145
-0,057	0,220	0,212	0,103	0,196	0,115	-0,106	-0,154	0,086	0,072
-0,263	-0,053	-0,258	-0,165	-0,181	0,180	0,027	0,016	-0,046	0,230
-0,245	-0,059	0,158	0,014	-0,361	-0,033	-0,213	0,000	0,063	-0,262
-0,200	-0,205	-0,147	-0,287	-0,389	-0,204	-0,083	0,053	0,018	-0,356
0,075	0,083	-0,113	0,204	0,204	-0,144	0,020	-0,097	0,093	-0,105
0,018	0,249	-0,119	-0,241	0,211	-0,191	0,069	0,090	0,021	-0,146

Vamos então escalonar os dados de modo que todas as variáveis envolvidas passem a apresentar média nula e variância unitária. Posteriormente, vamos conduzir a determinação do número ótimo de componentes a serem utilizados no modelo *BTPLS* na situação onde todas as entradas são incluídas no modelo final. A Tabela 5.19 apresenta a percentagem da variabilidade de $\boldsymbol{X}$ e $\boldsymbol{y}$ utilizada por cada componente do modelo.

Tabela 5.19: Variabilidade relativa e acumulada de $\boldsymbol{X}$ e $\boldsymbol{y}$ em cada etapa da decomposição.

comp	dX	X	dy	y
1	25.49	25.49	93.75	93.75
2	26.19	51.68	0.245	93.99
3	25.28	76.96	0.046	94.04
4	22.52	100.00	0.014	94.06

Como podemos verificar, o primeiro componente do modelo *BTPLS* é capaz de explicar a maior parte da variabilidade da variável de resposta e os demais componentes explicam insignificantes parcelas da variabilidade residual de $\boldsymbol{y}$. Esta observação está de acordo com o esperado, uma vez que os dados foram gerados de modo que uma única combinação linear das entradas fosse necessária para explicar o comportamento de $\boldsymbol{y}$. Portanto, vamos continuar a análise utilizando apenas o primeiro componente.

Novamente, a seleção de variáveis será conduzida pelo método *SRMP*, detalhado no Capítulo 3. A Tabela 5.20 apresenta os valores médios de quinhentas computações da *PRESS*

(valores em relação aos dados escalonados) obtidos nas diferentes etapas do procedimento *SRMP*. Estes valores também são plotados na Figura 5.10. Na Tabela 5.20, são apresentados, ainda, o desvio padrão dos cem valores utilizados no cálculo da *PRESS* e a significância da hipótese de a *PRESS* obtida na etapa atual ser maior do que o valor mínimo obtido entre todas as etapas anteriores.

Tabela 5.20: Sumário dos resultados do procedimento *SRMP*.

var	PRESS	desvio	SIG
1	0,0252	0,00670	0,000
2	0,0023	0,00073	0,000
4	0,0025	0,00086	99,18
3	0,0025	0,00083	99,87

De acordo com o previsto, os índices avaliados sugerem que as variáveis x_3 e x_4 devem ser descartadas. Portanto, o modelo final será um modelo *BTPLS* com um componente, utilizando apenas as variáveis x_1 e x_2.

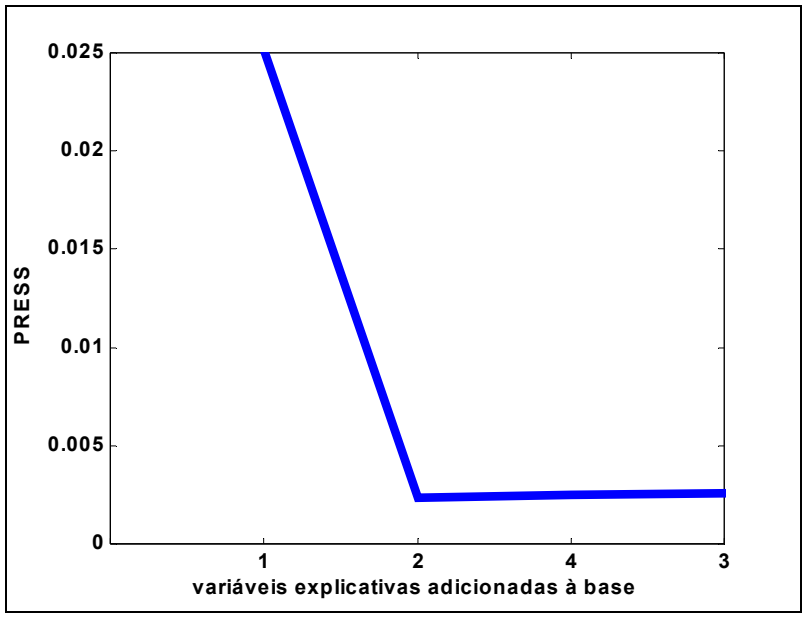

Figura 5.10: *PRESS* em função das variáveis adicionadas ao modelo.

Uma vez determinado o modelo final, devemos recorrer às amostras do conjunto de teste para testar a capacidade preditiva do mesmo. Na Figura 5.11, os valores da variável de resposta preditos pelo modelo final são plotados contra os valores "experimentais" para os conjuntos de treino e teste. Como podemos observar, o modelo mostra-se adequado tanto no que se refere ao ajuste dos dados do conjunto de treino quanto no que diz respeito a predição dos valores da resposta do conjunto de teste, o que comprova a validade do mesmo.

Após a validação do modelo final, vamos, do mesmo modo que nos exemplos anteriores, utilizar o método *bootstrapping residuals* para estimar as incertezas nos coeficientes do mesmo. Devemos lembrar que, como estamos trabalhando com o método

BTPLS, não podemos obter um modelo em termos das variáveis originais. Portanto, para uma análise completa das incertezas, precisamos avaliar, além da variabilidade dos coeficientes b_0, b_1, δ e α, a variabilidade dos elementos w_1 e w_2 do vetor peso $\boldsymbol{w}$ utilizado para a extração da primeira variável latente da matriz $[\boldsymbol{x_1 x_2}]$. Os valores estimados para cada um destes parâmetros a partir das cinqüenta amostras do conjunto de treino são mostrados na Tabela 5.21 juntamente com as respectivas variabilidades, obtidas pelo método *bootstrapping residuals*. Como podemos verificar, com exceção de b_0, todos os parâmetros apresentam uma baixa variabilidade, o que indica a adequabilidade do modelo. Por curiosidade, também na Tabela 5.21, também são mostrados os resultados obtidos a partir de uma repetição desta simulação, onde o conjunto de treino é composto, ao invés de cinqüenta, por quinhentas amostras. Obviamente, a variabilidade das estimativas diminui à medida que o número de observações "experimentais" aumenta. No caso onde a simulação foi repetida com quinhentas amostras, todos os parâmetros avaliados apresentaram desvio padrão com um valor pelo menos uma ordem de grandeza abaixo do valor estimado.

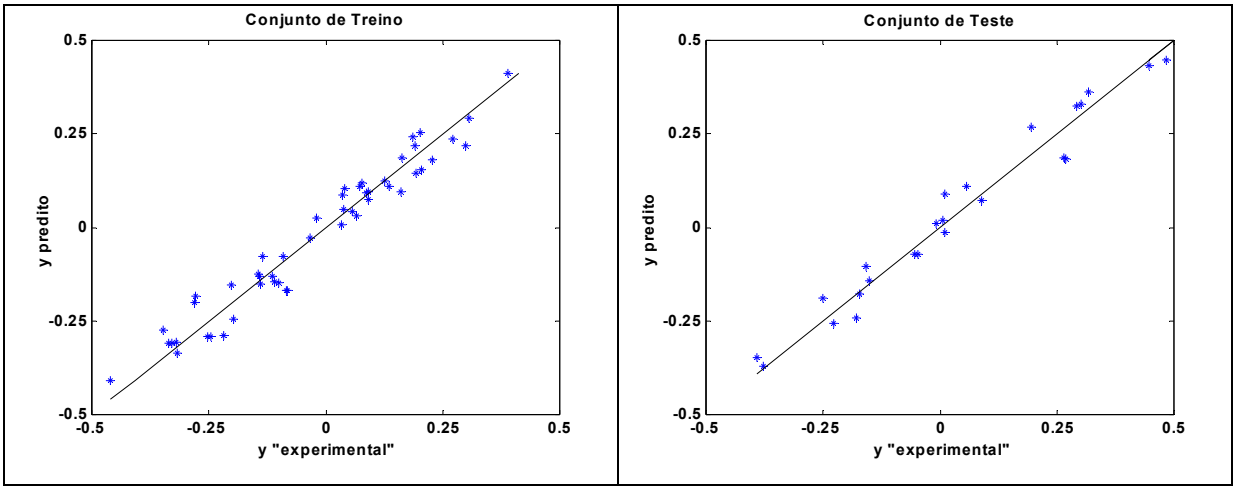

Figura 5.11: Predições do modelo final para os conjuntos de treino e teste

Tabela 5.21: Erros nos coeficientes do modelo *BTPLS* com 50 e 500 amostras.

	n=50		n=500	
b_0	-7.32E-04	5.75E-04	5.36E-05	1.32E-06
b_1	1.19E+00	1.60E-01	1.49E+00	1.34E-03
α	1.00E+00	0.00E+00	1.00E+00	0.00E+00
δ	9.03E-01	7.31E-02	1.02E+00	5.32E-04
w_1	7.23E-01	2.12E-02	7.07E-01	1.43E-04
w_2	6.91E-01	2.25E-02	7.07E-01	1.43E-04

Concluindo, a simulação conduzida foi capaz de ilustrar bem a aplicação da ferramenta desenvolvida no caso onde recorremos métodos de redução de dimensionalidade não lineares para modelar a relação existente entre a variável de resposta e as variáveis explicativas de um sistema.

Capítulo 6 Conclusões e Sugestões

Neste trabalho, foi conduzido um estudo visando a obtenção de uma ferramenta capaz de fornecer modelos empíricos que determinem a relação existente entre as diversas variáveis de processos industriais e a qualidade final dos produtos produzidos nos mesmos. Foram revisadas diferentes técnicas de modelagem empírica e propostas novas metodologias para a estimação das incertezas nas predições e também para a seleção das variáveis explicativas que devem compor o modelo final. Utilizando as metodologias revisadas e propostas, foi sugerida uma sistemática para o tratamento da questão da modelagem empírica de dados industriais, que constitui a base para a implementação da ferramenta desejada. Visando a maior flexibilidade possível, a sistemática foi elaborada de maneira bastante genérica, de modo que questões como o método de modelagem a ser utilizado, a técnica de seleção de variáveis e a metodologia para a estimação das incertezas do modelo foram deixadas como opções a serem especificadas em cada caso. Cada uma destas questões foi tratada, respectivamente, nos capítulos 2, 3 e 4 desta dissertação. Nos parágrafos que seguem, são feitas algumas colocações finais a respeito deste trabalho, onde as principais conclusões são destacadas e possibilidades para a continuidade do mesmo são identificadas.

No Capítulo 2, foi feita uma revisão englobando diversos métodos de regressão multivariável. De um modo geral, foram tratados, além da clássica técnica de regressão linear múltipla por mínimos quadrados, métodos de regressão utilizando transformação de variáveis para levar em conta a presença de não linearidades nas relações e métodos de redução de dimensionalidade para lidar com casos onde as entradas apresentam correlacionamento múltiplo. Também foram estudados métodos mistos, adequados para situações onde ambos os casos são verificados. Uma questão deixada de fora no desenvolvimento deste trabalho foi a utilização de redes neurais para a construção dos modelos. A utilização de redes neurais é amplamente difundida em processos industriais de grande porte e, conforme referências de Li et al (2001), podem ser encontrados na literatura trabalhos onde a utilização das mesmas é acoplada a algoritmos de métodos de redução de dimensionalidade. Com certeza, um estudo avaliando os méritos e ônus associados à utilização destas metodologias na ferramenta então desenvolvida seria um interessante complemento para este estudo.

Como já foi discutido, a seleção das variáveis explicativas que devem compor o modelo final é vital no processo de construção de um modelo empírico. No Capítulo 3, foi proposto o método *SRMP* (*Stepwise Regression based on Model Predictions*), uma nova metodologia para a abordagem deste problema, baseada na capacidade preditiva do modelo. Este critério foi adotado porque, na prática, a principal utilidade de um modelo é o fornecimento de predições para o valor da variável de resposta. O método proposto foi formulado de modo que, definida a técnica de modelagem a ser utilizada, devem ser especificados critérios para a medição da capacidade preditiva do modelo e para a determinação do momento em que a seleção de variáveis deve ser encerrada. A especificação destes critérios foi deixada como opcional para permitir mais flexibilidade ao procedimento *SRMP*. Neste trabalho, os critérios foram formulados visando eficiência em situações onde temos diversas variáveis de entrada altamente correlacionadas. Entretanto, é fato que a especificação dos mesmos é um fator chave na determinação dos resultados e que, portanto, a especificação de diferentes critérios para a execução do procedimento *SRMP* é um importante foco para o direcionamento de trabalhos futuros. Seria de grande utilidade o estudo de critérios específicos para situações mais particulares como aquelas onde a relação existente entre as entradas e as saídas é altamente não linear ou ainda onde a disponibilidade de observações experimentais é baixa.

Outra questão de fundamental importância é a determinação das incertezas associadas às predições, que foi abordada no Capítulo 4. Quando um modelo é utilizado para predizer o valor da variável de resposta para uma nova amostra, existem duas fontes de erro presentes nas computações: os erros referentes à medição das entradas para a amostra em questão (erros futuros) e os erros associados à construção do modelo (erros passados). Neste trabalho, a nossa atenção deve se voltar apenas para os erros de modelagem. Como já foi mencionado anteriormente, em muitos casos, a obtenção de expressões analíticas que forneçam aproximações para as incertezas presentes em um modelo de regressão podem não ser de fácil obtenção. Nestes casos, uma alternativa interessante é a obtenção de aproximações para a variância dos coeficientes do modelo pela análise estatística de diversas estimativas para os mesmos, obtidas a partir da reamostragem do conjunto de dados original. No Capítulo 4, foram revisadas algumas técnicas de reamostragem e também foi proposta uma nova metodologia, o método da adição de erro. Basicamente, os métodos revisados se baseiam ou nas observações experimentais ou nos resíduos do modelo para realizar a reamostragem, enquanto a metodologia proposta se baseia no erro experimental associado à medição das variáveis. Foram realizadas simulações para comparar as metodologias e, de um modo geral, a metodologia proposta se mostrou capaz de fornecer aproximações mais precisas para as incertezas do modelo. Por outro lado, é importante lembrar que a metodologia proposta requer, impreterivelmente, que informações a respeito dos erros experimentais associados à medida das variáveis esteja disponível e que a obtenção, assim como a confiabilidade, dos resultados é altamente dependente da qualidade das mesmas. Uma verificação importante, que foi comentada nos estudos comparativos realizados no Capítulo 4 é o fato de que, dependendo da forma da função objetivo utilizada para a estimação dos parâmetros do modelo, existe o risco da solução obtida a partir de cada um dos conjuntos reamostrados convergir para mínimos locais diferentes, o que torna sem sentido a computação da variância dos mesmos como uma aproximação para as respectivas incertezas. Neste trabalho, as discussões relativas à estimação de incertezas em problemas não lineares se limitaram ao caso onde este problema

não foi verificado, ficando o desenvolvimento de alternativas para lidar com tal situação como uma sugestão para futuros trabalhos.

A ferramenta proposta é capaz de modelar a relação existente entre diversas variáveis e também de fornecer estimativas para a precisão das predições do modelo, o que é fundamental para a determinação da confiabilidade do mesmo. Obviamente, a qualidade do modelo obtido, assim como a das estimativas para sua precisão, está intimamente relacionada com as opções que devem ser especificadas para a utilização da ferramenta. No Capítulo 5, a utilização da ferramenta desenvolvida foi ilustrada através de alguns estudos de caso. Os resultados obtidos nos cinco exemplos estudados foram, de uma forma geral, satisfatórios, sendo capazes de ilustrar bem a aplicabilidade da ferramenta desenvolvida. Para finalizar, cabe ressaltar que, embora o desenvolvimento deste trabalho tenha dado mais ênfase para a utilização de métodos de redução de dimensionalidade, a especificação do método *SRMP* como técnica de seleção de variáveis torna a ferramenta desenvolvida bastante genérica, permitindo que qualquer técnica de regressão multivariável seja utilizada na construção dos modelos.

Referências Bibliográficas

BAFFI, G., MARTIN, E., MORRIS, J., **PREDICTION INTERVALS FOR NON-LINEAR PROJECTION TO LATENT STRUCTURES REGRESSION MODELS**, CHEMOMETRICS AND INTELLIGENT LABORATORY SYSTEMS, V.61, 151-165, 2002.

BAFFI, G., MARTIN, E.B., MORRIS, A.J., **NON-LINEAR PROJECTION TO LATENT STRUCTURES REVISTED: THE QUADRATIC PLS ALGORITHM**, COMPUTERS AND CHEMICAL ENGINEERING, V.23, 395-411, 1999.

BOX, G.E.P., HUNTER, W. G., HUNTER, J. S., **STATISTICS FOR EXPERIMENTERS**, JOHN WILEY & SONS, 1978.

BOX, G.E.P., TIDWELL, P.W., **TRANSFORMATION OF THE INDEPENDENT VARIABLES**, TECHNOMETRICS, V. 4, Nº 4, 531-550, 1962.

DEMING, W.E., **STATISTICAL ADJUSTMENT OF DATA**, DOVER PUBLICATIONS, 1964.

FOGLER, H. S., **ELEMENTS OF CHEMICAL REACTION ENGINEERING**, PRENTICE HALL PTR, 2ª ED., 1992.

GELADI P., KOWALSKI, B.R., **PARTIAL LEAST SQUARES REGRESSION: A TUTORIAL**, ANALYTICA CHIMICA ACTA, V.185, 1-17, 1986.

GNANADESIKAN, R., **METHOS FOR STATISTICAL DATA ANALYSIS OF MULTIVARIATE OBSERVATIONS**, WILEY, 1977.

HARDY, A.J., MACLAURIN, P., HASWELL, S.J., JONG, S., VANDEGINSTE, B.G.M., **DOUBLE-CASE DIAGNOSTIC FOR OUTLIER IDENTIFICATION**, CHEMOMETRICS AND INTELLIGENT LABORATORY SYSTEMS, 34, 117-129, 1996.

HELLAND, I.S., **ON THE STRUCTURE OF PARTIAL LEAST SQUARES REGRESSION**, COMUNICATION IN STATISTICS: SIMULATION AND COMPUTATIONS, V.17(2), 581-607, 1988.

HIMMELBLAU, D.M., **PROCESS ANALYSIS BY STATISTICAL METHODS**, JOHN WILEY AND SONS, 1970.

HÖSKULDSSON, A., **PLS REGRESSION METHODS**, JOURNAL OF CHEMOMETRICS, V.2, 211-228, 1988.

HÖSKULDSSON, A., **PREDICTION METHODS IN SCIENCE AND TECHNOLOGY**, THOR PUBLISHING, V.1, 1996.

HÖSKULDSSON, A., **VARIABLE AND SUBSET SELECTION IN PLS REGRESSION**, CHEMOMETRICS AND INTELLIGENT LABORATORY SYSTEMS, 55, 23-38, 2001.

JAMES, B.R., **PROBABILIDADE: UM CURSO EM NÍVEL INTERMEDIÁRIO**, PROJETO EUCLIDES ,2ª ED., 1996.

KVALHEIM, O.M., KARSTANG, T.V., **INTERPRETATION OF LATENT-VARIABLE REGRESSION MODELS**, CHEMOMETRICS AND INTELLIGENT LABORATORY SYSTEMS, 7, 39-51, 1989.

LATADO, A., EMBIRUÇU, M., NETO, A.G.M., PINTO, J.C., **MODELING OF END-USE PROPERTIES OF POLY(ETHYLENE/PROPYLENE) RESINS**, POLYMER TESTING, V.20, P. 419-439, 2001.

LI, B., MARTIN, E.B., MORRIS, A.J., **BOX-TIDWELL BASED PARTIAL LEAST SQUARES REGRESSION**, COMPUTERS & CHEMICAL ENGINEERING, V.25, 1219-1233, 2001.

MARDIA, K.V., KENT J.T., BIBBY, J.M., **MULTIVARIATE ANALYSIS**, ACADEMIC PRESS, 7ª ED., 2000.

MILLER, R. G., **THE JACKKNIFE – A REVIEW,** BIOMETRIKA, 61, 1-15, 1974.

NICOLAAS, FABER, M., **UNCERTAINTY ESTIMATION FOR MULTIVARIATE REGRESSION COEFFICIENTS**, CHEMOMETRICS AND INTELLIGENT LABORATORY SYSTEMS, 64, 169-179, 2002

PHATAK, A., REILLY, P.M., PENLIDIS A., **AN APPROACH TO INTERVAL ESTIMATION IN PARTIAL LEAST SQUARES REGRESSION**, ANALYTICA CHIMICA ACTA, 277, 495-501, 1993.

SHACHAM, M. BRAUNER N., **CONSIDERING ERROR PROPAGATION IN STEPWISE POLYNOMIAL REGRESSION**, IND. CHEM. ENG. RES., V.38, 4477-4485, 1999-A.

SHACHAM, M. BRAUNER N., **CONSIDERING PRECISION OF EXPERIMENTAL DATA IN CONSTRUCTION OF OPTIMAL REGRESSION MODELS**, CHEMICAL ENGINEERING AND PROCESSING, V.38, 477-486, 1999-B.

SHACHAM, M. BRAUNER N., **THE SROV PROGRAM FOR DATA ANALYSIS AND REGRESSION MODEL IDENTIFICATION**, CHEMICAL ENGINEERING AND PROCESSING, V.27, 701-714, 2003.

STEPHENSON, R.M., MALANOWSKI, S., **HANDBOOK OF THERMODYANAMICS OF ORGANICS COMPOUNDS**, ELSEVIER, 1987.

STEWART, G.W., **COLLINEARITY AND LEAST SQUARES REGRESSION**, STATISTICAL SCIENCE, V.2, Nº 1, 68-100, 1987.

STEWART, G. W., **MATRIX ALORITHMS**, V.1, SIAM, 1998.

WERKEMA, M. C. C., AGUIAR, S., **ANÁLISE DE REGRESSÃO: COMO ENTENDER O RELACIONAMENTO ENTRE AS VARIÁVEIS DE UM PROCESSO,** FUNDAÇÃO CHRISTIANO OTTONI**,** V.7, 1996

WOLD, H., **NONLINEAR ESTIMATION BY ITERATIVE LEAST SQUARES PROCEDURES**, IN F. DAVID (ED.), RESEARCH PAPERS IN STATISTICS, FESTSCHRIFT FOR LERZY NEWMAN, 411-444, WILEY, NEW YORK, 1966.

WOLD, S. RUHE A., WOLD H., DUNN, W.J., **THE COLLINEARITY PROBLEM IN LINEAR REGRESSION: THE PARTIAL LEAST SQUARES APPROACH TO GENERALIZED INVERSES,** SIAM J. SCI. STAT. COMPUT., V.5, N° 3, 735-743, 1984.

WOLD, S. RUHE A., KETTANEH N., SKAGERBERG B., **NONLINEAR PLS MODELING,** CHEMOMETRICS AND INTELLIGENT LABORATORY SYSTEMS, 7, 53-65, 1989.