PRODUCTION
PRODUÇÃO

Production, v. 24, n. 1, p. 84-93, jan./mar. 2014
doi: http://dx.doi.org/10.1590/S0103-65132013005000030

Um novo método para seleção de variáveis preditivas com base em índices de importância

Juliano Zimmer[a]*, Michel José Anzanello[b]

[a]*zimmer@producao.ufrgs.br, UFRS, Brasil
[b]anzanello@producao.ufrgs.br, UFRS, Brasil

Resumo

O grande volume de variáveis coletadas em processos industriais impõe dificuldades ao controle e monitoramento de tais processos. A regressão PLS (*partial least squares*) vem sendo amplamente utilizada em procedimentos de seleção de variáveis por sua capacidade de operar com grande número de variáveis correlacionadas e afetadas por ruído. Este artigo propõe um método para identificar o melhor subconjunto de variáveis de processo para a predição das variáveis de resposta. Indicadores de importância das variáveis são desenvolvidos a partir de parâmetros da regressão PLS e guiam a eliminação das variáveis irrelevantes. Tais índices são então testados em termos de seu desempenho. Ao ser aplicado em cinco bancos de dados industriais, o método utilizando o índice recomendado reteve apenas 31% das variáveis originais e aumentou a acurácia de predição do conjunto de teste em 6%. O método proposto também superou a acurácia do método Stepwise, tradicionalmente utilizado em procedimentos de seleção com propósitos de predição.

Palavras-chave

Seleção de variáveis. Regressão PLS. Indicador de importância das variáveis.

1. Introdução

Diversos processos industriais envolvem elevado número de variáveis correlacionadas e contaminadas por ruído para seu controle e monitoramento. Exemplos de tais processos incluem refino e processamento de petróleo, siderurgia e produção de alimentos, bem como processos químicos em geral (GAUCHI; CHAGNON, 2001; CHONG; JUN, 2005; ZHAI; CHEN; HU, 2006; FERRER et al., 2008; PIERNA et al., 2009; ERIKSSON; WOLD, 2010; XIAOBO et al., 2010; ANZANELLO; ALBIN; CHAOVALITWONGSE, 2012). Kourti e MacGregor (1995) e Montgomery (2004) ressaltam que na presença de variáveis de processo correlacionadas é imperativo valer-se de métodos multivariados para o monitoramento do processo, uma vez que os métodos univariados podem levar a interpretações equivocadas. Martin, Morris e Kiparissides (1999) destacam que a utilização de técnicas de modelagem multivariada tornou-se relevante por possibilitar alertas acerca de mudanças no processo, potenciais falhas, mau funcionamento de componentes e distúrbios no processo.

O elevado volume de informações coletadas de processos industriais, no entanto, pode inviabilizar o monitoramento preciso dos mesmos, visto que grande parte dessas informações é inflada com ruído, colinearidade e dados faltantes (KOURTI; MacGREGOR, 1995; ANDERSEN; BRO, 2010; KONDYLIS; WHITTAKER, 2010). Nesse contexto, engenheiros e técnicos de produção são desafiados a identificar um conjunto reduzido de variáveis relevantes que descrevam características do processo e viabilizem o monitoramento e controle do mesmo (GAUCHI; CHAGNON, 2001; CHONG; JUN, 2005; FERRER et al., 2008; ANZANELLO; ALBIN; CHAOVALITWONGSE, 2009; ANDERSEN; BRO, 2010).

Métodos para seleção de variáveis têm sido continuamente propostos na literatura (LAZRAQ; CLÉROUX; GAUCHI, 2003; GAUCHI; CHAGNON, 2001; ANZANELLO; ALBIN; CHAOVALITWONGSE, 2009; CHIANG; PELL, 2004; PIERNA et al., 2009). Entre os métodos com propósito de predição,

*UFRS, Porto Alegre, RS, Brasil
Revisado 03/06/2011; Aceito 27/09/2012

destacam-se aqueles baseados em regressões PLS (*partial least squares*). A regressão PLS consiste em uma análise multivariada que transforma as variáveis de resposta e de processo em um número reduzido de combinações lineares. Seu amplo uso na indústria decorre de sua habilidade em lidar com um elevado número de variáveis de produto, múltiplas variáveis de resposta, dados com elevado nível de ruído, colinearidade e observações incompletas (KOURTI; MacGREGOR, 1995; WOLD; SJÖSTRÖM; ERIKSSON, 2001; FERRER et al., 2008; ANDERSEN; BRO, 2010).

Apesar do grande número de métodos para seleção de variáveis com propósitos de predição em PLS, não existe um método unânime para tal propósito (LAZRAQ; CLÉROUX; GAUCHI, 2003). Essa condição é justificada pelas características peculiares dos processos industriais, conforme pode ser visto nos estudos comparativos de Gauchi e Chagnon (2001), Lazraq, Cléroux e Gauchi (2003), Chong e Jun (2005) e Zhai, Chen e Hu (2006). Observa-se ainda que diversas possibilidades de utilização dos parâmetros gerados pela regressão PLS permanecem inexploradas e, por consequência, há espaço para abordagens mais eficientes para seleção de variáveis. Complementarmente, setores e aplicações específicas ainda carecem de métodos mais robustos para predição das variáveis de resposta, especialmente quando as variáveis de processo apresentam correlação elevada.

Este artigo apresenta um método para seleção de variáveis de processo com propósito de predição. Para tanto, os coeficientes gerados pela regressão PLS dão origem a índices de importância das variáveis de processo, os quais identificam as variáveis mais relevantes para explicação da variabilidade na variável de resposta. Inicia-se então um processo de eliminação de variáveis do tipo *backward*, sendo a ordem de eliminação definida pelo índice de importância. O desempenho do modelo de regressão resultante após cada eliminação de variável é avaliado por intermédio do indicador RMSE (*root mean square error*). Por fim, o método proposto é comparado com o tradicional método Stepwise.

O artigo inova ao adaptar o método de seleção proposto por Anzanello, Albin e Chaovalitwongse (2009), desenvolvido com propósito de classificação, para a seleção de variáveis com fins de predição através da regressão PLS. O artigo também desenvolve um novo índice de importância com base nos parâmetros oriundos de tal regressão, além de testar outro, ainda não utilizado em contexto de predição. A comparação com o método Stepwise também aparece como contribuição relevante, visto que o confronto de distintos métodos para seleção de variáveis auxilia na identificação dos métodos mais adequados em aplicações específicas.

O artigo está organizado em quatro seções, além desta introdução. A revisão bibliográfica é apresentada na seção 2, abordando os fundamentos da regressão PLS e métodos para seleção de variáveis. A seção 3 descreve os procedimentos metodológicos do trabalho, enquanto que a seção 4 apresenta os resultados obtidos. Por fim, tem-se a conclusão do trabalho, na seção 5.

2. Fundamentação teórica

As seções a seguir apresentam os fundamentos da regressão PLS e os parâmetros utilizados no método proposto, bem como abordagens para seleção de variáveis em contexto de predição.

2.1. *Regressão PLS*

A regressão PLS relaciona a matriz X (composta por variáveis de processo *x*) à matriz Y (composta por variáveis de produto *y*), permitindo analisar dados com forte correlação, elevados níveis de ruído e desequilíbrio entre o número de variáveis e observações. Tal regressão gera um conjunto de parâmetros que fornecem informações sobre a estrutura e comportamento de X e Y, o que corrobora para sua ampla aplicação em procedimentos de seleção de variáveis (WOLD; SJÖSTRÖM; ERIKSSON, 2001).

Ferrer et al. (2008) ressaltam que poucas ferramentas de análise estatística possuem a versatilidade da regressão PLS, a qual tem oferecido suporte em aplicações de diferentes naturezas, como discriminação e classificação de observações, modelagem e análise de processo e identificação de desvios. Suas aplicações não estão restritas a áreas industriais, mas também são verificadas em setores de negócios (avaliação de desempenho e comportamento humano), finanças e marketing (preferência por marcas, satisfação e fidelidade dos clientes) e áreas de ciências sociais (FERRER et al., 2008; ESPOSITO-VINZI et al., 2007).

Os fundamentos matemáticos da regressão PLS são agora apresentados. Considere uma matriz X, de dimensão ($K \times N$), e uma matriz Y, de dimensão ($M \times N$), na qual K denota o número de variáveis de processo, M o número de variáveis de resposta e N o número de observações. O vetor $\mathbf{x}_i (x_{i1}, x_{i2}, ..., x_{ik})$ representa a observação i para cada variável de processo k, enquanto que o vetor $\mathbf{y}_i (y_{i1}, y_{i2}, ..., y_{im})$ representa a observação i para cada variável de resposta m. A regressão PLS gera A variáveis latentes (combinações lineares) $\mathbf{t}_a$ ($a = 1, 2, ..., A$) a partir das variáveis originais, as quais são usadas com propósitos de predição e controle de processo (WOLD;

SJÖSTRÖM; ERIKSSON, 2001). Além de serem em número reduzido, geralmente de duas a cinco, as variáveis t_a são ortogonais entre si, ou seja, reduzem os problemas oriundos da elevada colinearidade das variáveis originais (WOLD; SJÖSTRÖM; ERIKSSON, 2001; FERRER et al., 2008; ANZANELLO; ALBIN; CHAOVALITWONGSE, 2009).

Para a escolha do número de componentes a serem mantidos no modelo, avalia-se a significância em termos de predição de cada componente *a*; a inclusão de componentes no modelo é interrompida quando eles deixam de ser significativos (WOLD; SJÖSTRÖM; ERIKSSON, 2001). Wold, Sjöström e Eriksson (2001) e Höskuldosson (2001) sugerem o uso da técnica de validação cruzada, a qual se destaca por sua praticidade e robustez, para definir o número de componentes a serem retidos. Adicionalmente, pode-se optar pelo Algoritmo Inferencial de Lazraq e Cleroux (2001), pelo método de minimização da média quadrada do erro preditivo (DENHAM, 2000), pelo critério de Kaiser (onde somente são retidos os componentes cujos autovalores são maiores que 1), ou pela soma da variância explicada pelos componentes retidos. Ressalta-se ainda que os limitados componentes retidos descrevem grande parte da variância das variáveis de processo e produto, bem como a covariância entre as mesmas (ANZANELLO; ALBIN; CHAOVALITWONGSE, 2009).

As variáveis latentes t_a são combinações lineares independentes das variáveis *x* com coeficientes w_a (w_{1a}, w_{2a}, ..., w_{ka}), conforme a Equação 1. O vetor w_a representa o peso da variável de processo *k* no componente *a*, sendo importante ressaltar que também leva em conta a influência das variáveis de produto (WOLD; SJÖSTRÖM; ERIKSSON, 2001; FERRER et al., 2008; ANZANELLO; ALBIN; CHAOVALITWONGSE, 2009).

$$t_{ia} = w_{1a}x_{i1} + w_{2a}x_{i2} + ... + w_{ja}x_{ij} = \mathbf{w}'_a\mathbf{x}_i \quad (1)$$

Da mesma forma, geram-se as variáveis latentes u_a (a = 1, 2, ..., A), que são combinações lineares das variáveis *y*. O vetor c_a (c_{1a}, c_{2a}, ..., cm_a) representa o peso de cada variável de produto m no componente *a* (WOLD; SJÖSTRÖM; ERIKSSON, 2001; FERRER et al., 2008; ANZANELLO; ALBIN; CHAOVALITWONGSE, 2009).

$$u_{ia} = c_{1a}y_{i1} + c_{2a}y_{i2} + ... + c_{ma}y_{im} = \mathbf{c}'_a\mathbf{y}_i \quad (2)$$

De acordo com Ferrer et al. (2008) e Anzanello, Albin e Chaovalitwongse (2009), os vetores w_a e c_a são selecionados de forma a maximizar a covariância entre os componentes t_a e u_a. Além disso, tais componentes aglutinam informações sobre as observações e suas semelhanças em relação ao modelo (WOLD; SJÖSTRÖM; ERIKSSON, 2001). Wold, Sjöström e Eriksson (2001) afirmam ainda que w_a e c_a fornecem informações sobre como as variáveis se combinam para formar a relação quantitativa entre X e Y, sinalizando as variáveis *x* de maior relevância (maiores valores de w_a).

Os parâmetros da regressão PLS podem ser estimados através do algoritmo NIPALS, sugerido por Geladi e Kowalski (1986). Os pesos **w** das correlações lineares das variáveis independentes e **c** das variáveis dependentes são definidos através da maximização da covariância entre as combinações lineares. Para tanto, parte-se de vetores de pesos zerados, os quais passam a ser alterados em um procedimento baseado em busca; a cada alteração nos pesos, monitora-se o valor da covariância entre os componentes, retendo os pesos que aumentarem a covariância. Os pesos que conduzem ao valor máximo de covariância nesse procedimento de busca são utilizados nas etapas subsequentes da regressão.

Multiplicando o vetor de cargas das variáveis de processo, p_a (p_{1a}, p_{2a}, ..., p_{ka}) pelo vetor t_a pode-se reconstituir a matriz X com valores reduzidos dos resíduos e_{ik} (WOLD; SJÖSTRÖM; ERIKSSON, 2001), conforme a Equação 3. (Entre parênteses é apresentada a representação matricial do procedimento.)

$$x_{ik} = \sum_a t_{ia}p_{ak} + e_{ik} \quad (\mathbf{X} = \mathbf{TP}' + \mathbf{E}) \quad (3)$$

Por sua vez, a predição das variáveis de resposta *y* pode ser obtida pela multiplicação de u_a pelos coeficientes c_a (WOLD; SJÖSTRÖM; ERIKSSON, 2001):

$$y_{im} = \sum_a u_{ia}c_{am} + g_{im} \quad (\mathbf{Y} = \mathbf{UC}' + \mathbf{G}) \quad (4)$$

Por fim, os coeficientes da regressão PLS podem ser reescritos como apresentado na equação (5), onde $w^*_{ka} = w_{ka}\left(p_{ka}w_{ka}\right)^{-1}$ e f_{im} são os resíduos da predição, os quais podem ser utilizados para diagnóstico da qualidade do modelo (WOLD; SJÖSTRÖM; ERIKSSON, 2001; ANZANELLO; ALBIN; CHAOVALITWONGSE, 2009).

$$b_{mk} = \sum_a c_{ma}w^*_{ka} + f_{im} \quad (\mathbf{B} = \mathbf{W} \times \mathbf{C}') \quad (5)$$

Substituindo-se as equações anteriores pode-se chegar ao formato tradicional do modelo de regressão (WOLD; SJÖSTRÖM; ERIKSSON, 2001).

$$y_{im} = \sum_a b_{mk}x_{ik} + f_{im} \quad (\mathbf{Y} = \mathbf{XB} + \mathbf{F}) \quad (6)$$

A qualidade da predição gerada pelo modelo pode ser avaliada através da soma do resíduo médio, $\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{N}\left(y_i - \hat{y}_i\right)^2}{N}}$, onde y_i é o valor observado e $\hat{y}_i$ é o valor estimado a partir da regressão PLS (GAUCHI; CHAGNON, 2001; MONTGOMERY; RUNGER, 2009).

2.2. *Abordagens para seleção de variáveis*

A presença de um grande número de variáveis de produto e processo tem incentivado engenheiros e pesquisadores a buscarem modelos compostos por um número reduzido de variáveis com vistas à redução dos custos de coleta de dados, aumento da precisão das informações geradas e maior possibilidade de aplicações práticas (GAUCHI; CHAGNON, 2001; MONTGOMERY; RUNGER, 2009; PIERNA et al., 2009; ANDERSEN; BRO, 2010).

O uso de abordagens explanatórias (por exemplo, gráficos de normalidade) com escolha manual das variáveis pode se tornar impraticável quando as variáveis são numerosas e correlacionadas (GAUCHI; CHAGNON, 2001; ANDERSEN; BRO, 2010). Além disso, um modelo de regressão com bom ajuste aos dados não conduz necessariamente a boas predições, evidenciado por situações de *overfitting* ou quando o processo apresenta alterações no intervalo decorrido entre a construção do modelo e sua efetiva aplicação (HÖSKULDSSON, 2001; LAZRAQ; CLÉROUX; GAUCHI, 2003; CHONG; JUN, 2005; PIERNA et al., 2009; ANDERSEN; BRO, 2010).

Dentre as abordagens para seleção de variáveis aplicadas no contexto de regressões lineares múltiplas, o método Stepwise é possivelmente o mais amplamente difundido (MONTGOMERY; RUNGER, 2009). O método também vem sendo usado para a seleção de variáveis em regressões PLS com propósito de predição (GAUCHI; CHAGNON, 2001; CHONG; JUN, 2005; ZHAI; CHEN; HU, 2006). Sua operacionalização ocorre através da sistemática adição ou remoção de variáveis na regressão, realizada com base em um teste estatístico de significância de cada variável (MONTGOMERY; RUNGER, 2009). Apesar de amplamente difundido, o desempenho do método Stepwise é afetado por variáveis correlacionadas e ruidosas (GAUCHI; CHAGNON, 2001).

Métodos mais robustos vêm sendo propostos para a seleção de variáveis em aplicações preditivas de PLS (HÖSKULDSSON, 2001; GAUCHI; CHAGNON, 2001; LAZRAQ; CLÉROUX; GAUCHI, 2003; CHONG; JUN, 2005; ZHAI; CHEN; HU, 2006; PIERNA et al., 2009). Gauchi e Chagnon (2001) comparam 20 métodos de seleção baseados em diferentes critérios de avaliação, incluindo ajuste do modelo e capacidade de predição. Dentre os métodos, destacam-se o BCOR (*backward correlations*), BQ (*backward* Q^2_{cum}) e algoritmo genético (AG). O método BCOR usa os parâmetros da regressão PLS para rodar uma sequência de eliminação de variáveis a partir da significância dos coeficientes de cada variável x em cada componente a. O método BQ, por sua vez, elimina sistematicamente a variável associada ao menor coeficiente da regressão PLS e registra Q^2_{cum} para avaliar a qualidade da predição a cada eliminação. Por fim, o conjunto de variáveis que maximiza o Q^2_{cum} é escolhido. Já o AG, baseado num critério de busca, retém um número reduzido de variáveis e conduz a bons resultados na predição, porém apresenta alta variabilidade e requer demasiado processamento computacional.

Com propósitos semelhantes, Chong e Jun (2005) comparam o desempenho de três métodos para seleção de variáveis: método PLS-VIP (*variable importance in the projection*), regressão Lasso (*least absolute shrinkage and selection operator*) e regressão Stepwise. Experimentos simulados em cenários com alta colinearidade apontaram o método PLS-VIP como o mais adequado para previsões. Pierna et al. (2009) desenvolveram um método para seleção de variáveis espectrais baseado na regressão PLS e remoção *backward* a partir do desempenho de predição do modelo, mensurado através do RMSE. O método proposto manteve ou aumentou a capacidade de predição ao ser aplicado em dois bancos de dados com múltiplas variáveis de resposta. O método assemelha-se ao aqui proposto, porém não faz uso de indicadores de importância das variáveis para escolher a variável a ser eliminada a cada iteração (ao invés disso, utiliza uma parte do banco de dados para iterativamente identificar a variável que menos colabora com o RMSE). Além disso, o método em Pierna et al. (2009) não é comparado com outros métodos propostos pela literatura, o que dificulta a conclusão sobre seu desempenho.

A proposição de índices de importância das variáveis também tem encontrado elevada aplicação em procedimentos de seleção; tais índices atuam como guias no processo de eliminação ou inclusão sistemática de variáveis no modelo. Wold, Sjöström e Eriksson (2001) desenvolveram um índice de importância das variáveis, VIP, a partir do coeficiente modificado de peso w^*_{ka} e da fração de variância explicada pelo componente a em Y, R^2_{Ya}. Esse índice foi testado em Lazraq, Cléroux e Gauchi (2003) e Anzanello, Albin e Chaovalitwongse (2009). Outros índices com propósitos semelhantes podem ser obtidos em Eriksson e Wold (2010).

Por sua vez, Anzanello, Albin e Chaovalitwongse (2009) propuseram um método para seleção de variáveis de processo para fins de classificação das variáveis de resposta a partir do uso combinado de índices de importância das variáveis e técnicas de mineração de dados. Através de um processo de eliminação do tipo *backward*, as variáveis com o menor índice de importância são sequencialmente removidas do conjunto de variáveis retidas. O desempenho de classificação é avaliado a cada iteração, sendo

escolhido o subconjunto que maximiza a acurácia de classificação. No método proposto neste artigo, as variáveis são sistematicamente eliminadas com base em novos índices de importância, porém com objetivo de predição (e não de classificação).

Cinco índices de importância foram propostos em Anzanello, Albin e Chaovalitwongse (2009), dentre os quais três são aqui abordados: o índice v_w, baseado no indicador VIP proposto por Wold, Sjöström e Eriksson (2001) [ver Equação 7], é amplamente usado para seleção de variáveis visando predição. O índice v_k, na Equação 8, é uma variação do índice VIP e ainda não foi aplicado com propósitos de predição, sendo gerado com base nos pesos w_{ka} e na fração da variação de Y, R^2_{Ya}, explicada pelo componente a (a = 1, ..., A). O índice v_b, na Equação 9, define a importância da variável de processo k com base no coeficiente b_{mk} da regressão PLS, o qual mede a magnitude da relação entre X e Y. Esses índices são combinados na seção 3 para a geração de um novo índice de importância das variáveis.

$$v_w = \frac{\sum_{a=1}^{A}\left(w^*_{ka}\right)^2 R^2_{Ya}}{\max_{k\in K}\left(\sum_{a=1}^{A}\left(w^*_{ka}\right)^2 R^2_{Ya}\right)} \quad k = 1,\ldots,K \tag{7}$$

$$v_k = \frac{\sum_{a=1}^{A}|w_{ka}|\, R^2_{Ya}}{\max_{k\in K}\left(\sum_{a=1}^{A}|w_{ka}|\, R^2_{Ya}\right)} \quad k = 1,\ldots,K \tag{8}$$

$$v_b = \frac{\sum_{m=1}^{M}|b_{mk}|}{\max_{k\in K}\left(\sum_{m=1}^{M}|b_{mk}|\right)} \quad k = 1,\ldots,K \tag{9}$$

3. Método proposto

O método proposto é operacionalizado em cinco etapas: (*i*) divisão do banco de dados em conjuntos de treino e teste; (*ii*) aplicação da regressão PLS no conjunto de treino e geração de índices de importância das variáveis; (*iii*) predição da variável de resposta y para o conjunto de treino e eliminação das variáveis irrelevantes e ruidosas; (*iv*) construção de um gráfico para identificação do melhor subconjunto de variáveis e validação das variáveis selecionadas no conjunto de teste; e (*v*) comparação do desempenho dos diferentes índices frente ao método Stepwise. Enfatiza-se que o método proposto assume as variáveis de produto y como contínuas e, por isso, adaptações podem ser necessárias para uso com variáveis de produto discretas. Os passos propostos são detalhados na sequência.

3.1. *Etapa 1: Dividir o banco de dados em conjuntos de treino e teste*

Considere as matrizes X e Y, introduzidas na seção 2, com N observações, K variáveis de processo e uma variável de produto. Inicia-se dividindo aleatoriamente as observações do banco de dados em um conjunto de treino tr com N_{tr} observações e em um conjunto de teste ts com N_{ts} observações tais que $N_{tr} + N_{ts} = N$. As variáveis relevantes são identificadas a partir do conjunto de treino. Já o conjunto de teste é utilizado para avaliação da capacidade predita do modelo gerado. Recomenda-se usar uma proporção de 3:2 entre as observações de N_{tr} e N_{ts}, respectivamente (ANZANELLO; ALBIN; CHAOVALITWONGSE, 2009).

3.2. *Etapa 2: Aplicar a regressão PLS no conjunto de treino e gerar índices de importância das variáveis*

Para evitar efeitos de escala nos resultados, sugere-se normalizar os dados antes de aplicar a regressão. Os parâmetros de interesse incluem os coeficientes de regressão b_{mk}, pesos w_{ka} e o percentual de variação em Y explicada pelo componente a, R^2_{Ya}. Tais parâmetros são utilizados para gerar os índices de importância das variáveis de processo.

Três são os indicadores de importância usados no presente método. O índice v_w foi escolhido por ser baseado no índice VIP, amplamente usado para seleção de variáveis com propósitos de predição (WOLD; SJÖSTRÖM; ERIKSSON, 2001)). O índice v_k, elaborado por Anzanello, Albin e Chaovalitwongse (2009), é aqui utilizado de forma inovadora com a finalidade de guiar a escolha das variáveis de processo mais significativas para a predição de y. Por fim, com base nas Equações 7 e 9 propõe-se um novo índice, v_{bw}, apresentado na Equação 10, o qual integra três parâmetros da regressão PLS para definir a importância da variável k: o coeficiente de regressão b_{mk}, os pesos w_{ka} e a fração da variação de Y, R^2_{Ya}, explicada pelo componente a (a = 1, ..., A).

$$v_{bw} = \frac{\sum_{a=1}^{A}|w_{ka}|\, R^2_{Ya}}{\max_{k\in K}\left(\sum_{a=1}^{A}|w_{ka}|\, R^2_{Ya}\right)} \frac{\sum_{m=1}^{M}|b_{mk}|}{\max_{k\in K}\left(\sum_{m=1}^{M}|b_{mk}|\right)} \quad k = 1,\ldots,K \tag{10}$$

3.3. *Etapa 3: Predizer a variável de resposta y para o conjunto de treino e eliminar as variáveis irrelevantes e ruidosas*

Uma primeira predição é feita valendo-se de K variáveis de processo e o desempenho da predição é medido através do RMSE. Na sequência, remove-se do conjunto de treino a variável com o menor índice

de importância da variável, roda-se a regressão PLS com as K–1 variáveis de processo e registra-se novo RMSE. Repete-se o processo, removendo-se a variável com menor índice e aplicando-se a regressão PLS para predição de y até que reste apenas uma variável de processo.

3.4. *Etapa 4: Construir um gráfico para identificar o melhor subconjunto de variáveis e testar essas variáveis no conjunto de teste*

O RMSE calculado a cada eliminação de variável é relacionado ao percentual de variáveis retidas através de um gráfico, como apresentado na Figura 1. O subconjunto responsável pelo menor RMSE é selecionado para predição do conjunto de observações de teste; o desempenho de predição para novas observações é estimado via RMSE.

3.5. *Etapa 5: Comparar o desempenho dos índices de importância e identificar o melhor método para a seleção de variáveis*

Aplica-se o método Stepwise no conjunto de treino; as variáveis selecionadas são inseridas como independentes em um modelo PLS. Compara-se então o desempenho do método proposto utilizando-se os índices v_w, v_k e v_{bw}, e o método Stepwise em termos de RMSE e percentual de variáveis retidas.

4. Resultados e discussão

Para aplicação do método proposto e avaliação do seu desempenho foram utilizados os cinco bancos de dados em Gauchi e Chagnon (2001), os quais também constam nos trabalhos de Lazraq, Cléroux e Gauchi (2003) e Anzanello, Albin e Chaovalitwongse (2009). As análises foram realizadas em MATLAB® versão 7.10.

O número de variáveis de processo de cada banco de dados assim como a divisão das observações em conjuntos de treino e teste são apresentados na Tabela 1. O banco de dados ADPN, com 71 observações, é procedente de um processo intermediário da produção de nylon. As 262 observações do LATEX foram extraídas de um processo de manufatura de látex. Os dados de OXY, com 30 observações, correspondem ao processo de produção do óxido de titânio, o qual é usado na mistura de tintas. O processo SPYRA refere-se à etapa de fermentação para a produção de antibiótico. Por fim, o banco GRANU é proveniente de um processo de emulsões antiespuma utilizado na indústria de papel. Os bancos estão disponíveis mediante solicitação aos autores.

As variáveis desses bancos são caracterizadas por elevados níveis de colinearidade. O banco ADPN, por exemplo, apresenta correlações acima de 0,7 para 30% das variáveis (as matrizes de correlação não são apresentadas por restrições de espaço). A implicação de tais níveis de correlação na predição da PLS, no entanto, é minimizada pela sistemática de operacionalização da regressão, a qual apoia-se na geração de combinações lineares independentes

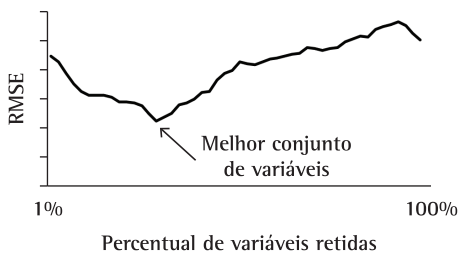


Figura 1. Perfil hipotético de RMSE à medida que as variáveis de processo são eliminadas do conjunto de treino.

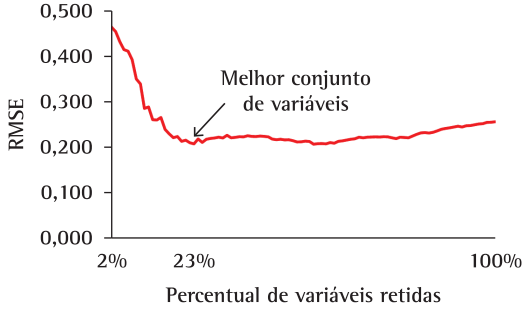


Figura 2. Desempenho da predição no conjunto de treino do processo OXY usando o índice v_{bw}.

Tabela 1. Bancos de dados analisados

Banco de dados	Número de variáveis de processo	Número de observações	
		Conjunto de treino	Conjunto de teste
ADPN	100	57	14
LATEX	117	210	52
OXY	95	18	12
SPIRA	96	115	29
GRANU	78	23	6

Tabela 2. Desempenho dos métodos de seleção de variáveis para os conjuntos de treino e teste.

Banco de dados	RMSE – conjunto de treino				RMSE – conjunto de teste				Variáveis retidas (%)			
	v_k	v_w	v_{bw}	Stepwise	v_k	v_w	v_{bw}	Stepwise	v_k	v_w	v_{bw}	Stepwise
ADPN	1,110	1,216	0,961	1,102	1,051	1,228	0,968	1,225	56	36	37	31
LATEX	0,602	0,586	0,546	0,588	0,594	0,573	0,549	0,610	7	16	24	15
OXY	0,227	0,225	0,208	0,213	0,126	0,121	0,089	0,172	39	24	23	9
SPIRA	0,160	0,157	0,156	0,161	0,148	0,147	0,147	0,182	57	47	41	10
GRANU	0,654	0,638	0,601	0,669	0,742	0,455	0,448	0,978	28	37	28	6
Média	0,551	0,564	0,494	0,547	0,532	0,505	0,440	0,633	37,4	32	30,6	14,2

Tabela 3. Intervalos de confiança para a variável de resposta.

Banco	Intervalo de confiança
ADPN	98,34 ± 2,91
LATEX	5,6 ± 0,25
OXY	5,38 ± 0,83
SPIRA	2,18 ± 0,093
GRANU	5,18 ± 0,51

das variáveis de processo e de resposta. O peso de tais combinações, por serem ortogonais para os diferentes componentes retidos, confere aos coeficientes de regressão PLS condições de predição sem influência da correlação, a qual foi desdobrada na concepção das combinações lineares.

A regressão PLS foi aplicada ao conjunto de treino de cada banco de dados. Foram retidos três componentes da regressão para cada banco de dados através de validação-cruzada (ver WOLD; SJÖSTRÖM; ERIKSSON, 2001), resultando nos seguintes R^2_{Ya}'s: ADPN, 94%; LATEX, 77%; OXY, 94%; SPIRA, 71%; e GRANU, 86%.

A Figura 2 apresenta o perfil de RMSE à medida que as variáveis são eliminadas do conjunto de treino para o banco de dados OXY, ao aplicar-se o índice v_{bw}. A escolha do melhor conjunto de variáveis considerou uma solução de compromisso entre o menor percentual de variáveis retidas e o menor RMSE. O método proposto reteve apenas 23% das variáveis, gerando um RMSE de 0,208 (valor próximo ao menor valor possível de RMSE). A utilização de todas as variáveis conduz a um RMSE de 0,257, representando um aumento de acurácia preditiva de 19% no conjunto de treino para o banco OXY. A mesma lógica foi aplicada aos demais bancos de dados e índices de importância das variáveis.

A comparação do desempenho dos distintos índices de importância nas porções de treino e teste é apresentada na Tabela 2. O índice v_{bw} apresentou desempenho superior de predição no conjunto de treino de todos os processos analisados, com destaque para o processo ADPN, onde se verificou acurácia 26% superior ao segundo melhor índice. O RMSE médio para os cinco bancos é 0,494, o que incrementa a acurácia de predição média em 6% (o RMSE médio, utilizando todas as variáveis, é 0,525) valendo-se de 31% das variáveis originais (em média sobre todos os bancos). Tal índice também conduziu aos melhores resultados para as predições da porção de teste, com RMSE médio igual a 0,440. Os índices v_k e v_w alternam seu desempenho de acordo com o banco analisado, conduzindo a maiores valores de RMSE e retendo mais variáveis do que o índice v_{bw}.

Por fim, o método composto pelo índice selecionado (v_{bw}) é comparado ao tradicional método Stepwise. O método proposto conduz a predições mais precisas tanto para a porção de treino (RMSE = 0,494 contra RMSE = 0,547 do método Stepwise), quanto para a porção de teste (RMSE = 0,440 contra RMSE = 0,633 do método Stepwise). Essa última vantagem é expressiva, pois, apesar de o método proposto reter mais variáveis do que o Stepwise, conduz a predições 44% mais acuradas do que o modelo gerado pela seleção tradicional.

A qualidade de ajuste dos modelos compostos pelas variáveis selecionadas aos dados pode ser averiguada através do gráfico de resíduo, apresentados no apêndice. Não são verificadas tendências de maior expressão nos resíduos, os quais aparecem distribuídos de forma randômica e sem presença de homoscedasticidade. Por sua vez, os intervalos de confiança para a variável de resposta com um nível de significância de 95% nas porções de teste dos bancos de dados analisados são apresentados na Tabela 3.

5. Conclusões

Processos industriais caracterizados por elevado número de variáveis correlacionadas e ruidosas demandam métodos de seleção para assegurar boa capacidade de predição dos modelos gerados. O presente artigo apresentou um método de seleção de variáveis de processo mais relevante com vistas à predição das variáveis de resposta.

O método proposto se apoia nas seguintes etapas: (1) divisão dos bancos de dados compostos por variáveis de processo e resposta em conjuntos de treino e teste; (2) aplicação da regressão PLS no conjunto de treino e geração dos índices de importância das variáveis v_w

v_k e v_{bw}; (3) predição dos valores de Y para o conjunto de treino e eliminação das variáveis com base nos índices de importância, registrando-se o desempenho preditivo via RMSE; (4) construção de um gráfico associando RMSE e percentual de variáveis retidas, seleção do subconjunto recomendado e predição da variável de resposta para o conjunto de teste usando tal subconjunto de variáveis; e (5) comparação do desempenho do método composto pelos três índices frente ao método Stepwise.

O novo índice de importância de variáveis, v_{bw}, foi comparado aos índices v_w e v_k, também gerados com base nos coeficientes da regressão PLS. O novo índice apresentou a melhor acurácia de predição de Y, quando comparado aos índices v_k e v_w e ao método tradicional Stepwise. Além disso, o índice v_{bw} reteve um percentual de variáveis inferior ao obtido com os índices v_k e v_w. Ao valer-se do índice v_{bw}, o método proposto utilizou 31% das variáveis para predição, gerando uma acurácia de predição 6% superior ao valor obtido quando todas as variáveis são utilizadas. Portanto, o método com o novo indicador de importância das variáveis é recomendado para aplicações que necessitam de elevada acurácia de predição a partir de um conjunto reduzido de variáveis.

Pesquisas futuras incluem a comparação do método de seleção de variáveis proposto e do novo indicador de importância das variáveis com relação a outros métodos para seleção de variáveis. Da mesma forma, sugere-se a utilização de outros indicadores de acurácia da predição de Y, além do RMSE, para corroborar resultados e conclusões do presente artigo.

Referências

ANDERSEN, C. M.; BRO, R. Variable selection in regression – a tutorial. *Journal of Chemometrics*, v. 24, p. 728-737, 2010. http://dx.doi.org/10.1002/cem.1360

ANZANELLO, M. J.; ALBIN, S. L.; CHAOVALITWONGSE, W. A. Selecting the best variables for classifying production batches into two quality levels. *Chemometrics Intelligent Laboratory Systems*, v. 97, p. 111-117, 2009. http://dx.doi.org/10.1016/j.chemolab.2009.03.004

ANZANELLO, M. J.; ALBIN, S. L.; CHAOVALITWONGSE, W. Multicriteria variable selection for classification of production batches. *European Journal of Operational Research*, v. 218, p. 97-105, 2012. http://dx.doi.org/10.1016/j.ejor.2011.10.015

CHIANG, L. H.; PELL, R. J. Genetic algorithms combined with discriminant analysis for key variable identification. *Journal of Process Control*, v. 14, p. 143-155, 2004. http://dx.doi.org/10.1016/S0959-1524(03)00029-5

CHONG, I.-G.; JUN, C.-H. Performance of some variable selection methods when multicollinearity is present. *Chemometrics Intelligent Laboratory Systems*, v. 78, p. 103-112, 2005. http://dx.doi.org/10.1016/j.chemolab.2004.12.011

DENHAM, M. C. Choosing the number of factors in partial least squares regression: estimating and minimizing the mean squared error of prediction. *Journal of Chemometrics*, v. 14, p. 351-361, 2000. http://dx.doi.org/10.1002/1099-128X(200007/08)14:4<351::AID-CEM598>3.0.CO;2-Q

ERIKSSON, L.; WOLD, S. A graphical index of separation (GIOS) in multivariate modeling. *Journal of Chemometrics*, *Bognor Regis*, v. 24, p. 779-789, 2010.

ESPOSITO-VINZI, V. et al. *Handbook of Partial Least Squares*: Concepts, Methods and Applications in Marketing and Related Fields. Berlin: Springer, 2007. 850 p.

FERRER, A. et al. PLS: A versatile tool for industrial process improvement and optimization. *Applied Stochastic Models in Business and Industry*, v. 24, p. 551-567, 2008. http://dx.doi.org/10.1002/asmb.716

GAUCHI, J. P.; CHAGNON, P. Comparison of selection methods of explanatory variables in PLS regression with application to manufacturing process data. *Chemometrics Intelligent Laboratory Systems*, v. 58, p. 171-193, 2001. http://dx.doi.org/10.1016/S0169-7439(01)00158-7

GELADI, P.; KOWALSKI, B. Partial least-squares regression: a tutorial. Analytica Chimica Acta, v. 185, p. 1-17, 1986. http://dx.doi.org/10.1016/0003-2670(86)80028-9

HÖSKULDSSON, A. Variable and subset selection in PLS regression. *Chemometrics and Intelligent Laboratory Systems*, v. 55, p. 23-38, 2001. http://dx.doi.org/10.1016/S0169-7439(00)00113-1

KONDYLIS, A.; WHITTAKER, J. Adaptively preconditioned Krylov spaces to identify irrelevant predictors. *Chemometrics and Intelligent Laboratory Systems*, v. 104, p. 205-213, 2010. http://dx.doi.org/10.1016/j.chemolab.2010.08.010

KOURTI, T.; MacGREGOR, J. F. Process analysis, monitoring and diagnosis, using multivariate projection methods. *Chemometrics Intelligent Laboratory Systems*, v. 28, p. 3-21, 1995.

LAZRAQ, A.; CLÉROUX, R. The PLS multivariate regression model: testing the significance of successive PLS components. *Journal of Chemometrics*, v. 15, p. 523-536, 2001. http://dx.doi.org/10.1002/cem.641

LAZRAQ, A.; CLÉROUX, R.; GAUCHI, J.-P. Selecting both latent and explanatory variables in the PLS1 regression model. *Chemometrics Intelligent Laboratory Systems*, v. 66, p. 117-126, 2003. http://dx.doi.org/10.1016/S0169-7439(03)00027-3

MARTIN, E. B.; MORRIS, A. J.; KIPARISSIDES, C. Manufacturing performance enhancement through multivariate statistical process control. *Annual Reviews in Control*, v. 23, p. 35-44, 1999.

MONTGOMERY, D. C. *Introdução ao controle estatístico da qualidade*. 4. ed. Rio de Janeiro: LTC – Livros Técnicos e Científicos Editora S.A., 2004. 513 p.

MONTGOMERY, D. C.; RUNGER, G. C. *Estatística aplicada e probabilidade para engenheiros*. 4. ed. Rio de Janeiro: LTC – Livros Técnicos e Científicos Editora S.A., 2009. 493 p.

PIERNA, J. A. F. et al. A Backward Variable Selection method for PLS regression (BVSPLS). *Analytica Chimica Acta*, v. 642, p. 89-93, 2009. PMid:19427462. http://dx.doi.org/10.1016/j.aca.2008.12.002

WOLD, S.; SJÖSTRÖM, M.; ERIKSSON, L. PLS-regression: a basic tool of chemometrics. *Chemometrics Intelligent*

Laboratory Systems, v. 58, p. 109-130, 2001. http://dx.doi.org/10.1016/S0169-7439(01)00155-1

XIAOBO, Z. et al. Independent component analysis in information extraction from visible/near-infrared hyperspectral imaging data of cucumber leaves. *Chemometrics and Intelligent Laboratory Systems*, v. 104, p. 265-270, 2010. http://dx.doi.org/10.1016/j.chemolab.2010.08.019

ZHAI, H. L.; CHEN, X. G.; HU, Z. D. A new approach for the identification of important variables. *Chemometrics Intelligent Laboratory Systems*, v. 80, p. 130-135, 2006. http://dx.doi.org/10.1016/j.chemolab.2005.09.002

A new framework for predictive variable selection based on variable importance indices

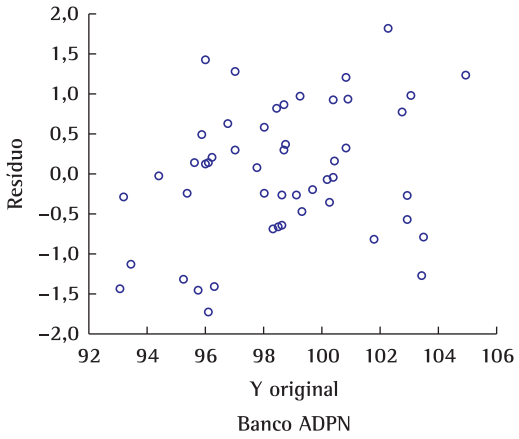
Abstract

The large volume of process variables collected from manufacturing applications has jeopardized process control activities. The Partial Least Squares (PLS) regression has been widely used for variable selection due to its ability to handle a large number of correlated and noisy variables. This paper presents a method for selecting the most relevant variables aimed at predicting product variables. For that matter, variable importance indices are developed based on PLS parameters and used to guide the elimination of noisy and irrelevant variables. Variables are then systematically removed from the dataset and the performance of the predictive model evaluated. When applied to five manufacturing datasets, the proposed method retained 31% of the original variables and yielded 6% more accurate predictions than using all original variables. Further, the proposed method outperformed the traditional Stepwise method regarding prediction accuracy.

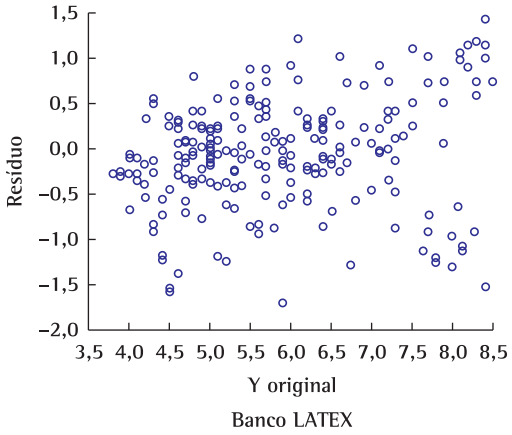
Keywords

Variable selection. PLS regression. Variable importance índices.

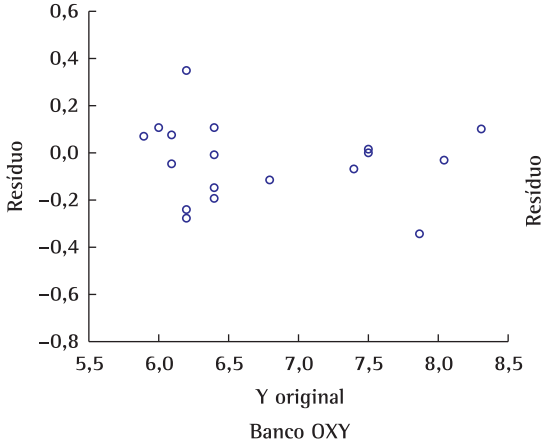
Apêndice 1. Gráficos de resíduos dos modelos.



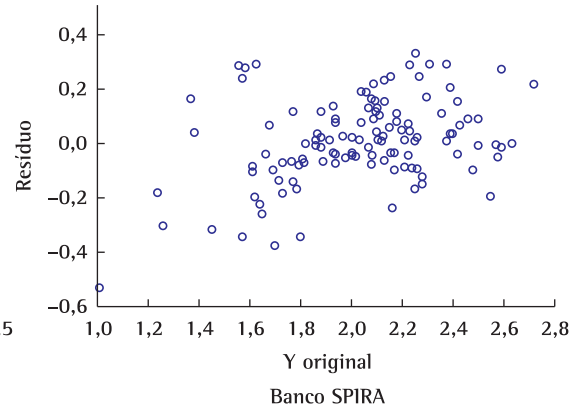
Banco ADPN



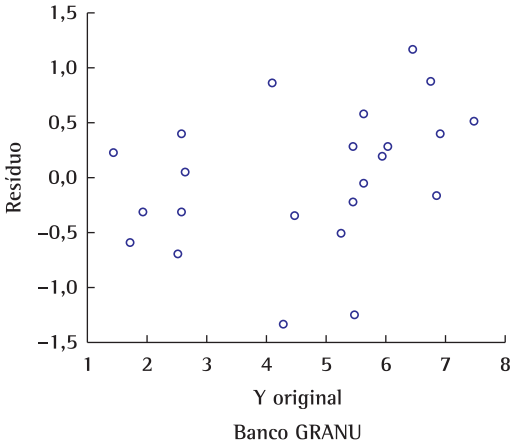
Banco LATEX



Banco OXY



Banco SPIRA



Banco GRANU