

# Gráficos de controle multivariados para monitoramento de processos não lineares em bateladas

Danilo Marcondes Filho<sup>a,\*</sup>, Flávio Sanson Fogliatto<sup>b</sup>, Luiz Paulo Luna de Oliveira<sup>c</sup>

<sup>a,\*</sup>marcondes.filho@ufrgs.br, DEST/UFRGS, Brasil  
<sup>b</sup>ffogliatto@producao.ufrgs.br, PPGEP/UFRGS, Brasil  
<sup>c</sup>pluna@unisinis.br, PIPCA/UNISINOS, Brasil

## Resumo

Processos industriais em bateladas são empregados com frequência na produção de certos itens. Tais processos disponibilizam uma estrutura de dados peculiar; diante disso, existe um crescente interesse no desenvolvimento de gráficos de controle multivariados mais apropriados para seu monitoramento. Investiga-se aqui uma abordagem recente que utiliza gráficos de controle baseados no método *Statis*. O *Statis* constitui-se em uma técnica exploratória que permite avaliar similaridade entre matrizes de dados. Entretanto, essa técnica considera a similaridade em um contexto linear, investigando estruturas de correlação lineares nos dados. Propõe-se neste artigo a utilização de gráficos de controle baseados no *Statis* em conjunto com *kernels* para monitoramento de processos com presença de não linearidades fortes. Através dos *kernels*, definem-se funções não lineares dos dados para melhor representação da estrutura a ser caracterizada pelo método *Statis*. Essa nova abordagem, denominada *Kernel-Statis*, é desenvolvida e avaliada utilizando dados de um processo simulado.

## Palavras-chave

Controle multivariado da qualidade. Gráficos de controle. Processos em bateladas. *Kernel*. Método *Statis*.

## 1. Introdução

Processos industriais automatizados disponibilizam uma grande quantidade de informações sobre seu desempenho. Em tais processos são geradas medições simultâneas e em tempo real de diversas variáveis de desempenho. Obtêm-se, então, dados em quantidade suficiente para habilitar um monitoramento preciso do desempenho de operações industriais. Parte dessas indústrias conduz seus processos em bateladas.

Processos em bateladas tendem a apresentar características não lineares acentuadas (LEE et al., 2004), visto que: i) são não estacionários e, conseqüentemente, a média das variáveis de processo não é constante (isto é, as variáveis apresentam uma trajetória não linear ao longo da batelada); e ii) operam em diferentes estágios, aumentando assim a incidência de efeitos não lineares no sistema. Dadas essas características, foram propostos na literatura gráficos de controle (GCs) multivariados

que incorporam a estrutura de correlação não linear entre as variáveis desses processos.

Abordagens lineares de controle estatístico multivariado de processos (*MSPC* ou *Multivariate Statistical Process Control*) para processos em bateladas na literatura baseiam-se fundamentalmente na análise de componentes principais multidirecionais (*MPCA* ou *Multiway Principal Component Analysis*). A *MPCA* consiste em aplicar a Análise de Componentes Principais (*PCA* ou *Principal Component Analysis*) no arranjo de dados sugerido originalmente por Nomikos e MacGregor (1995). Através da *PCA* obtém-se uma representação resumida dos dados considerando correlações lineares entre as variáveis. As principais abordagens não lineares de controle baseiam-se em extensões não lineares da *PCA*, denominadas *Non-Linear PCA* (*NLPCA*). Os GCs baseados em *NLPCA* são obtidos a partir do uso da *PCA* em conjunto com modelos de

redes neurais, com o algoritmo de curvas principais ou mediante a utilização de *kernels*. Martin e Morris (1996), Lee, Yoo e Lee (2004) e Lee et al. (2004), apresentam uma discussão comparativa de GCs baseados em NLPCA.

Uma abordagem alternativa, denominada *Statis* (LAVIT et al., 1994), proposta recentemente, utiliza um arranjo de dados distinto em relação à MPCA. O *Statis* constitui-se em uma técnica exploratória que oferece uma representação sumária do grau de similaridade entre matrizes de dados através da utilização da PCA nesse arranjo. Os GCs baseados no método *Statis* foram propostos inicialmente por Scepi (2002) e formalizados para o monitoramento *online* e *off-line* de processos em bateladas por Fogliatto e Niang (2008, 2009). A caracterização dos dados oferecida pelo *Statis* traz um acréscimo em relação ao arranjo usado na MPCA, pois permite a construção de GCs para avaliar o desempenho do processo explicitamente a cada instante. Entretanto, assim como as demais abordagens lineares, a técnica avalia a similaridade no contexto linear, investigando apenas estruturas de correlação lineares nos dados.

Este artigo propõe o desenvolvimento de GCs baseados em uma modificação do *Statis* que incorpore também não linearidades presentes nos dados, e que será aqui denominada *Kernel-Statis* (K-S). Através do uso de *kernels*, definem-se funções não lineares dos dados para melhor representação da estrutura a ser caracterizada pelo método *Statis*. O desempenho dos GCs propostos é ilustrado utilizando dados simulados.

Os GCs aqui propostos trazem duas contribuições importantes na área de controle multivariado da qualidade. Primeiro, tais gráficos são capazes de absorver e representar de maneira resumida as informações oriundas de grandes massas de dados multivariados, considerando a complexa estrutura de correlação e autocorrelação não linear das variáveis de processo. Dessa forma, os gráficos podem oferecer sinalizações mais precisas, com menor número de alarmes falsos, sobre o estado do processo. Como decorrência, obtém-se um melhor conhecimento sobre o processo. Segundo, os GCs propostos são capazes de oferecer diagnósticos precisos que permitem uma rápida ação corretiva quando o processo estiver fora de controle. Os diagnósticos são úteis para a melhoria do processo e consequente melhoria da qualidade do que é produzido. Decorrem também outros benefícios, tais como a diminuição de custos de produção (decorrente de um menor desperdício de recursos) e a redução do impacto de resíduos químicos no meio ambiente.

O restante deste artigo está organizado da seguinte maneira. Na seção 2 revisam-se i) estratégias disponíveis na literatura para o controle multivariado de processos não lineares; e ii) o método *Statis*. Na seção 3 apresenta-se a versão não linear do método *Statis*, o *Kernel-Statis* (K-S). Na seção 4, gráficos de controle baseados no método K-S são derivados. Na seção 5, o método K-S é aplicado a um conjunto de dados simulados. Conclusões encerram o artigo na seção 6.

## 2. Fundamentação teórica

A presente seção divide-se em duas partes, claramente identificadas no texto. Primeiro, revisa-se o estado da arte sobre estratégias não lineares para MSPC. Na sequência, são apresentados os fundamentos analíticos do método *Statis*.

A aplicação de GCs para monitorar processos em bateladas de duração fixa foi inicialmente proposta por Jackson e Mudholkar (1979), sendo depois investigada por Nomikos e MacGregor (1995), Kourti e MacGregor (1996) e MacGregor (1997). Aplicações dos GCs propostos no monitoramento de processos em bateladas podem ser encontradas em Kourti (2003), entre outros autores.

De maneira resumida, o monitoramento de processos em bateladas utilizando GCs é realizado verificando o resultado de dois gráficos distintos. O primeiro é um gráfico do tipo  $T^2$  para os escores obtidos projetando bateladas futuras nos  $q$  componentes principais (CPs) retidos na representação da distribuição de referência. Tal distribuição é obtida realizando uma MPCA em dados oriundos de bateladas conformes (ao contrário da PCA tradicional, a MPCA permite a análise de matrizes tridimensionais de dados – para uma comparação entre os dois métodos ver JOLLIFFE, 2004, p. 397). O segundo é um gráfico do tipo  $Q$  para os resíduos do modelo de referência. O primeiro GC monitora o comportamento das fontes conhecidas de variabilidade do processo; o segundo GC permite detectar eventos atípicos que perturbem a estrutura de correlação e autocorrelação linear das variáveis de processo.

No contexto não linear, algumas modificações na MPCA foram propostas na literatura. Kramer (1991) desenvolveu um método de NLPCA através de redes neurais autoassociativas. A arquitetura da rede neural utilizada apresenta cinco camadas: i) de entrada, contendo os pontos amostrais para treinamento; ii) de mapeamento; iii) camada gargalo, de onde serão gerados os escores referentes aos PCs não lineares; iv) de mapeamento reverso; e

v) de saída, contendo os mesmos pontos amostrais usados na camada de entrada. Um algoritmo de gradiente conjugado é utilizado para treinar a rede, buscando reproduzir as observações da camada (i) na camada (iv) com mínimo erro. Como a dimensão da camada (iii) é menor do que a dimensão de (i) e (v), a rede é forçada a desenvolver uma representação compacta dos dados de entrada. O autor atinge esse objetivo introduzindo funções não lineares nos nodos das camadas de mapeamento e mapeamento reverso. Os CPs não lineares obtidos na camada (iii) são utilizados então para a geração do gráfico  $T^2$  para os escores, e do gráfico  $Q$  para os resíduos do modelo de referência. Entretanto, a rede proposta é de difícil treino, já que contém cinco camadas; além disso, a determinação do número de modos a ser usado em cada camada não é tarefa trivial.

Dong e McAvoy (1996) também propuseram uma abordagem de NLPKA, combinando curvas principais e redes neurais, para controle de processos não lineares contínuos e em bateladas. Os escores referentes aos CPs não lineares para os pontos amostrais de treinamento são obtidos pelo método da curva principal; o modelo de rede neural é então utilizado para mapear os dados originais em seus respectivos escores, os quais são então utilizados para obter os mesmos pontos amostrais da camada (i), com mínimo erro. Construindo a rede neural, uma estratégia de adaptação *on-line* pode ser desenvolvida para geração de gráficos  $T^2$  para os escores e  $Q$  para os resíduos. A abordagem de Dong e McAvoy (1996) apresenta duas limitações: i) o algoritmo da curva principal pressupõe que a função não linear possa ser aproximada por uma combinação linear de diversas funções univariadas (isto é, a função não linear pode ser decomposta como uma soma de funções das variáveis individuais), o que restringe a aplicação do algoritmo a estruturas que apresentem comportamento do tipo aditivo; e ii) deve-se resolver um problema de otimização não linear para calcular as curvas principais e treinar a rede neural e, para tanto o número de componentes principais deve ser especificado antes de treinar a rede; assim, sempre que o número de componentes for alterado, o procedimento de modelagem deverá ser rodado novamente.

Versões alternativas para a NLPKA baseadas em redes neurais foram também propostas por Hiden et al. (1999) e Jia, Martin e Morris (2001). Na abordagem de Hiden et al. (1999), as não linearidades presentes no sistema são explicitamente representadas em uma forma funcional, cuja natureza é otimizada usando um processo evolutivo baseado em programação genética. Jia, Martin e Morris (2001) propõem uma abordagem combinando

PCA e uma rede neural de entrada e treinamento (*ITNN - Input-training Neural Network*), de forma a considerar separadamente correlações lineares e não lineares presentes nos dados. Geng e Zhu (2005) reportam uma aplicação prática do método proposto por Jia, Martin e Morris (2001) no monitoramento de um processo químico.

Os trabalhos a seguir utilizam a *Kernel-PCA* (KPCA) como uma abordagem de NLPKA. A KPCA, originalmente proposta por Scholkopf, Smola e Muller (1998), é capaz de calcular componentes principais de forma eficiente em espaços característicos (*feature spaces*) de grandes dimensões através de operadores integrais e funções *kernel* não lineares. Em sua essência, a KPCA consiste de duas operações: i) o espaço de entrada (*input space*) é mapeado, através de funções não lineares, em um espaço característico; e ii) uma PCA linear é aplicada no espaço característico para obter componentes principais. Comparada a outros métodos não lineares, a KPCA apresenta a vantagem de não demandar um procedimento de otimização não linear; sua utilização envolve somente operações de álgebra linear, sendo de aplicação tão simples quanto a PCA padrão. A KPCA demanda a extração de autopares (autovalores e autovetores) do espaço característico, mas não requer que o número de componentes principais a ser extraído seja conhecido a priori. Como pode ser operacionalizada usando diferentes *kernels*, a KPCA pode ser eficiente na representação de diferentes tipos de não linearidades.

Lee, Yoo e Lee (2004) apresentam um procedimento para o monitoramento de processos contínuos no espaço característico obtido aplicando funções *kernel* sobre os dados de processo. Os autores ilustram o procedimento em um processo de tratamento de resíduos líquidos onde os dados de processo são mapeados no espaço característico através de uma função *kernel* de base radial. Uma vez disponível a representação dos dados de entrada no espaço característico, o monitoramento proposto é essencialmente o mesmo apresentado por Nomikos e McGregor (1995), utilizando PCA linear. O monitoramento proposto por Lee, Yoo e Lee (2004) somente permite o controle *on-line* de processos contínuos, já que sua operacionalização demanda, como amostra de entrada, a matriz completa de dados do processo de interesse, não disponível, no caso de processos em bateladas, antes de seu término.

Lee et al. (2004) estendem o procedimento em Lee, Yoo e Lee (2004) para o monitoramento *on-line* e *off-line* de processos em bateladas, em uma estratégia denominada *Kernel MPCA*. O

esquema proposto para o monitoramento *off-line* replica os desenvolvimentos propostos por Lee, Yoo e Lee (2004), já que o monitoramento *on-line* de processos contínuos e *off-line* de processos em bateladas se equivalem em termos metodológicos. Com relação ao monitoramento *on-line* de processos em bateladas, Lee et al. (2004) propõem completar a matriz de dados de processo proveniente da batelada em curso utilizando uma metodologia em que valores futuros são antecipados como uma média ponderada dos escores disponíveis até o tempo atual da batelada e dos escores previamente calculados na distribuição de referência. O procedimento é ilustrado em um processo de fermentação para produção de penicilina.

Cho et al. (2005) propõem um método para o diagnóstico de pontos fora de controle sinalizados nos gráficos  $T^2$  e  $Q$  desenvolvidos por Lee et al. (2004). A contribuição em Lee et al. (2004) limitou-se à fase de detecção, não trazendo propostas para o diagnóstico de eventuais pontos fora de controle. O método de diagnóstico em Cho et al. (2005) está baseado no cálculo do gradiente da função *kernel* utilizada no mapeamento dos dados de processo no espaço característico, sendo aplicável no diagnóstico de sinais registrados nos gráficos  $T^2$  e  $Q$ . O método é ilustrado usando dados simulados de dois processos contínuos, previamente analisados por Dong e McAvoy (1996).

Cui, Li e Wang (2008) também abordam o problema do diagnóstico de pontos fora de controle em gráficos baseados em KPCA, além de analisar estratégias para reduzir a dimensão da matriz *kernel* durante a fase de treinamento da KPCA. Com relação ao problema do diagnóstico, os autores propõem o uso conjunto da KPCA e da análise discriminante de Fisher (método para extração de características e redução dimensional de grandes amostras; ver HASTIE; TIBSHIRANI; FRIEDMAN, 2001, p. 84). Para reduzir a dimensão da matriz *kernel*, os autores propõem identificar subconjuntos de dados no banco completo de dados de processo suficientes para expressar todos os dados no espaço característico como uma combinação linear dos dados nos subconjuntos reduzidos. Os desenvolvimentos no artigo são ilustrados utilizando dados simulados de processos previamente analisados por Lee et al. (2004) e Cho et al. (2005).

Choi, Morris e Lee (2008) combinam as proposições em Lee et al. (2004) e Cho et al. (2005) para propor um novo esquema de monitoramento de processos não lineares. O artigo enfatiza o problema da detecção de eventos anormais ocorridos em escalas muito distintas. Em sua essência, os autores propõem substituir o método

de padronização de dados proposto por Scholkopf, Smola e Muller (1998), prévio à KPCA, pela utilização da transformação Wavelet. Na etapa de diagnóstico, os autores propõem a utilização da transformação Wavelet inversa para mapear dados do espaço característico no espaço de entrada.

O método *Statis* foi aplicado no MSPC por Scepti (2002) e estendido para o caso de processos em bateladas de tempo variável por Fogliatto e Niang (2008). Entretanto, a abordagem de controle via *Statis* proposta neste artigo está restrita ao caso de bateladas de tempo fixo, cabendo assim um desenvolvimento posterior que contemple o caso de bateladas de tempo variável. A seguir são apresentados, de maneira resumida, os fundamentos do método *Statis*.

## 2.1. Método Statis

O método *Statis* permite a análise de estruturas tridimensionais de dados, avaliando a similaridade entre matrizes bidimensionais em um plano de dimensões reduzidas (ESCOUFIER, 1987; LAVIT et al., 1994). Considerando que o interesse aqui é a análise de dados oriundos de processos em bateladas, a estrutura tridimensional nesse contexto traz dados de referência acerca de  $B$  matrizes bidimensionais  $X_b$ . Cada matriz  $X_b$ , de dimensão  $(T \times P)$ , contém vetores linha  $x_t^b$  padronizados (isto é, cada variável em  $X_b$  teve seus valores subtraídos da média e divididos pelo desvio padrão da sua coluna) que representam medições de  $P$  variáveis de processo durante  $T$  instantes de tempo, conforme esquematizado na Figura 1. Tem-se então uma estrutura com  $P$  variáveis  $\times T$  instantes de tempo  $\times B$  bateladas.

O método *Statis* foi aplicado no MSPC de processos por Scepti (2002) e estendido para o caso de processos em bateladas por Fogliatto e Niang (2008). A estruturação de dados apresentada na Figura 1 cumpre dois objetivos. Primeiro, representar

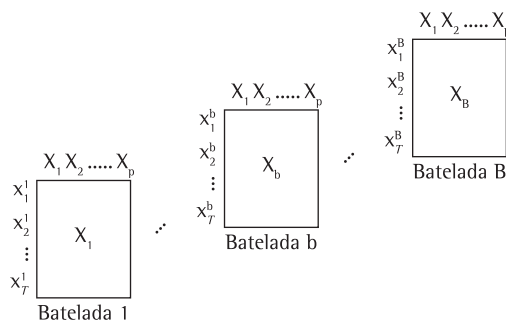


Figura 1. Arranjo de dados para análise via Statis.

em um espaço de dimensões reduzidas a correlação entre as  $P$  variáveis das matrizes  $X_b$  e  $X_{b'}$  no conjunto dos  $T$  instantes. Essa análise permite verificar o comportamento global das variáveis de uma nova batelada em relação à estrutura de referência capturada nas  $B$  bateladas conformes. Esse objetivo é alcançado através da análise da interestrutura. Segundo, representar em um espaço reduzido a correlação média (ou de compromisso) entre os  $T$  instantes, dois a dois, considerando todas as  $P$  variáveis de processo. Essa análise busca identificar, a cada instante de tempo transcorrido na nova batelada, possíveis desvios significativos em relação ao comportamento temporal de compromisso do conjunto das variáveis. Esse objetivo é alcançado através da análise da intraestrutura.

Na análise da interestrutura define-se, para cada matriz  $X_b$ , uma matriz  $W_b = X_b X_b'$  de dimensão  $(T \times T)$ , onde  $X_b'$  indica a transposta de  $X_b$ . Genericamente,  $W_b$  é dada por:

$$W_b = \begin{bmatrix} \ddots & & \ddots \\ & x_t^b, x_{t'}^b & \\ \ddots & & \ddots \end{bmatrix} \text{ para } t, t' = 1, \dots, T \text{ e } b = 1, \dots, B \quad (1)$$

Os elementos da matriz  $W_b$  correspondem ao produto interno entre os vetores de linha  $x_t^b$  e  $x_{t'}^b$  de  $X_b$ , com medições das  $P$  variáveis nos instantes  $t$  e  $t'$ , respectivamente. Considerando dados padronizados,  $W_b$  traz uma medida de correlação do conjunto das variáveis entre instantes de tempo na  $b$ -ésima batelada. A análise das informações da batelada por essa matriz prioriza os instantes de tempo.

A medida de similaridade entre pares de matrizes  $W_b$  utilizada no método *Statis* é o produto interno canônico de Hilbert-Schmidt (ESCOUFIER, 1973), dado por:

$$S_{bb'} = W_b |W_{b'}|_{HS} = Tr(DW_b DW_{b'}) \quad (2)$$

onde  $Tr(\bullet)$  representa o operador de traço matricial, e  $D$  é uma matriz diagonal, de dimensão  $(T \times T)$ , contendo pesos de importância  $D_t$  para os instantes de tempo. Neste artigo, consideram-se processos em que causas de variabilidade não desejadas incidem uniformemente nos instantes de tempo; logo,  $D = 1/T$ . Adicionalmente, consideram-se variáveis padronizadas e matrizes  $W_b$  normalizadas; consequentemente, utiliza-se  $W_b / \sqrt{Tr(DW_b DW_b)}$  no lugar de  $W_b$ . Nesse caso, a Equação 2 representa o coeficiente de correlação linear vetorial (coeficientes RV) entre as matrizes  $W_b$  e  $W_{b'}$ . Lavit et al. (1994) demonstram que o coeficiente RV corresponde ao quadrado do coeficiente de correlação linear de Pearson entre as variáveis das matrizes  $X_b$  e  $X_{b'}$ .

As correlações lineares vetoriais entre  $W_b$  e  $W_{b'}$  estão descritas na matriz:

$$S\Delta = \begin{pmatrix} \ddots & & \ddots \\ & \pi_b S_{bb'} & \\ \ddots & & \ddots \end{pmatrix} \quad (3)$$

onde  $\Delta$  é a matriz de pesos de importância  $\pi_b$  ( $b = 1, \dots, B$ ) atribuídos às  $B$  bateladas. Para bateladas igualmente ponderadas,  $\Delta = 1/B$ .

Para obter uma caracterização resumida da estrutura de correlação entre as  $B$  bateladas, aplica-se a PCA na matriz  $S\Delta$ , retendo-se os autopares que promovem, em conjunto, uma representação eficiente do conjunto de dados analisados. A representação das  $B$  bateladas nos novos eixos ortogonais obtidos a partir da PCA é realizada utilizando os autovetores  $u_i$  retidos na análise. Cada elemento  $u_{i,b}$  de  $u_i$ , ponderado pelo desvio padrão do CP correspondente (dado pela raiz quadrada do  $i$ -ésimo autovalor  $\lambda_i$ ), representa a posição da  $b$ -ésima batelada no  $i$ -ésimo eixo ortogonal; tem-se então:

$$a_{i,b} = \sqrt{\lambda_i} u_{i,b}, \quad (4)$$

onde  $a_{i,b}$  é a coordenada que representa essa posição. A Figura 2 apresenta de maneira esquemática o roteiro de análise da interestrutura descrita acima.

A análise da intraestrutura é realizada a partir da obtenção de uma matriz de compromisso  $W$ , que representa a estrutura de correlação média (considerando as  $B$  bateladas de referência) entre os  $T$  instantes de tempo. Entende-se que a matriz  $W$  apresenta o comportamento temporal do conjunto das variáveis nos instantes de uma batelada ideal.  $W$  é obtida através da soma ponderada das  $B$  matrizes de referência, utilizando a expressão

$$W = \sum_{b=1}^B \alpha_b W_b \text{ ou, de forma genérica:}$$

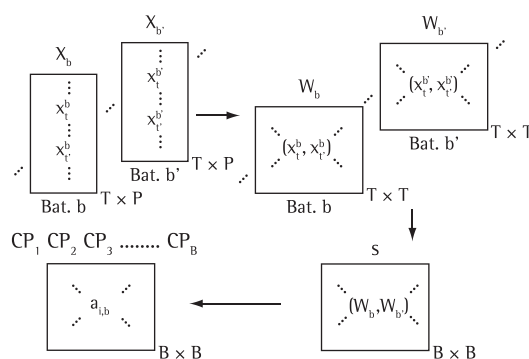


Figura 2. Esquema geral da análise da interestrutura.

$$W = \begin{bmatrix} \cdot & & \cdot \\ \cdot & \sum_{b=1}^B \alpha_b \langle x_t^b, x_{t'}^b \rangle & \cdot \\ \cdot & & \cdot \end{bmatrix}, \text{ para } t, t' = 1, \dots, T \quad (5)$$

Lavit et al. (1994) demonstram que a combinação linear que relaciona melhor as matrizes  $W_b$  com  $W$  está associada ao maior autovalor ( $\lambda_1$ ) extraído de  $S\Delta$  e ao seu autovetor correspondente  $u_1$ . Assim, os pesos  $\alpha_b$  são dados por:

$$\alpha_b = \frac{1}{\sqrt{\lambda_1}} \frac{1}{B} u_{1,b} \quad (6)$$

onde  $u_{b,1}$  representa o b-ésimo elemento do vetor  $u_1$  referente à b-ésima batelada.

Para obter uma caracterização resumida da estrutura de correlação de compromisso das  $P$  variáveis nos  $T$  instantes de tempo, a exemplo do que foi feito na análise da interestrutura, aplica-se uma PCA na matriz  $WD$ . Isso é feito através da sua diagonalização para seleção dos maiores autovalores  $\delta_i$  e respectivos autovetores  $\varepsilon_i$  ( $i = 1, \dots, T$ ), que descrevem a posição das observações  $x_t^b$  médias em um número reduzido de eixos, derivados das principais direções ortogonais de variabilidade comum em  $WD$ .

A representação da estrutura de compromisso nos novos eixos ortogonais é realizada utilizando os autovetores  $\varepsilon_i$ . Cada elemento  $\varepsilon_{i,t}$  de  $\varepsilon_i$ , ponderado pelo desvio padrão do CP correspondente, representa a posição do t-ésimo instante de tempo de compromisso no i-ésimo eixo ortogonal; tem-se então:

$$z_{i,t} = \sqrt{\delta_i} \varepsilon_{i,t} \quad (7)$$

onde  $z_{i,t}$  é a coordenada que representa essa posição.

Para comparar em cada instante o comportamento do conjunto das  $P$  variáveis da b-ésima batelada  $W_b$  em relação à batelada de compromisso  $W$ , obtém-se a representação de cada matriz  $W_b$  nos novos eixos ortogonais, através da seguinte expressão:

$$z_{i,t}^b = \frac{1}{\sqrt{\delta_i}} \frac{1}{T} w_t^b \varepsilon_i \quad (8)$$

onde  $\varepsilon_i'$  é o transposto de  $\varepsilon_i$ ,  $w_t^b$  representa a t-ésima linha de  $W_b$  e  $z_{i,t}$  é o valor que representa a posição no i-ésimo eixo ortogonal da b-ésima batelada no t-ésimo instante de tempo. A Figura 3 ilustra o esquema geral da análise da intraestrutura.

### 3. Kernel-Statís

Nesta seção propõe-se uma generalização do método *Statís* que permita obter medida de correlação não linear entre matrizes  $X_b$ . Isso será feito usando a teoria de *kernels* (SCHOLKOPF; SMOLA, 2002), através da imersão das matrizes  $X_b$  em um espaço estendido (denominado espaço característico) contendo funções não lineares  $\Phi(x^b)$  das observações originais  $x_t^b$  de  $X_b$ . Tal espaço pode, por exemplo, conter produtos de segunda ordem dos elementos de  $x_t^b$ ; nesse caso, pode-se avaliar a correlação não linear (quadrática) entre as matrizes  $X_b$ . Esse objetivo é atingido com utilização de uma função *kernel* apropriada das observações  $x_t^b$ .

Considere dois vetores de observações  $x_t^b$  e  $x_{t'}^b$ , e uma função  $k(x_t^b, x_{t'}^b)$ , denominada *kernel*, que gera um número real que representa a similaridade entre os vetores, tal que  $k(x_t^b, x_{t'}^b) = k(x_{t'}^b, x_t^b)$ . As medidas de similaridade comumente usadas em análise multivariada de dados utilizam o produto interno canônico, definido através do *kernel*  $k(x_t^b, x_{t'}^b) = x_t^b \cdot x_{t'}^b$ . Essa medida permite trabalhar com construções geométricas formuladas em termos de ângulos, comprimentos e distâncias.

Ressalta-se que técnicas estatísticas lineares de análise multivariada, tais como a PCA, análise discriminante e o método *Statís*, constituem-se em algoritmos que podem ser escritos em função de produtos internos canônicos entre observações. Versões não lineares desses algoritmos podem ser obtidas através de medidas de similaridade mais gerais, representadas por produtos internos modificados dos dados via *kernels*. Esse resultado é conhecido como *kernel trick*, sendo detalhado em Scholkopf e Smola (2002).

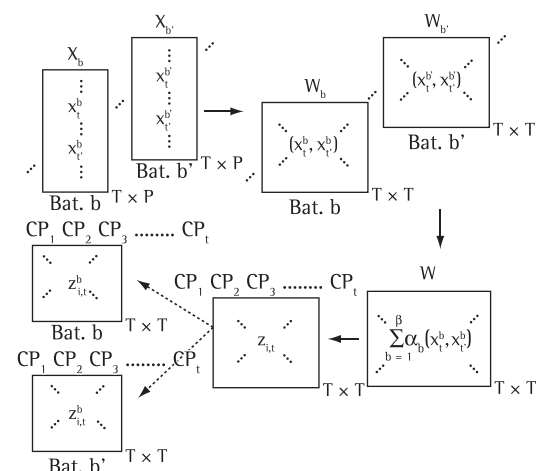


Figura 3. Esquema geral da metodologia de análise da intraestrutura.

Medidas de similaridade podem ser obtidas a partir da construção de um mapa não linear  $\Phi$ :

$$\begin{aligned} \Phi: \mathfrak{R}^p &\rightarrow F \\ x_t^b &\rightarrow \Phi(x_t^b) \end{aligned} \quad (9)$$

onde  $F$  representa o espaço dos produtos internos ou espaço dos atributos, de dimensão  $N_F$  maior ou igual ao espaço de entrada  $\mathfrak{R}^p$ . Note que o vetor  $\Phi(x_t^b)$  traz a representação do vetor  $x_t^b$  no espaço dos atributos  $F$ .

A escolha do mapa não linear  $\Phi$  permite construir uma série de medidas alternativas de similaridade (tais como correlações e distâncias em um contexto não linear), trabalhando apenas no espaço de entrada  $\mathfrak{R}^p$ , utilizando  $k(x_t^b, x_t^b) = \Phi(x_t^b), \Phi(x_t^b)$ . Mesmo que a dimensionalidade aumente demais, o espaço dos atributos não é acessado, pois através de um *kernel* apropriado pode-se construir um produto interno modificado no  $\mathfrak{R}^p$  que corresponda ao produto interno entre os mapas  $\Phi$  no espaço  $F$  (como exemplificado mais adiante nesta seção utilizando um *kernel* polinomial).

Uma questão relevante é a definição de uma classe de *kernels* que permita a representação  $k(x_t^b, x_t^b) = \Phi(x_t^b), \Phi(x_t^b)$ . Em outras palavras, dado um *kernel*, deve-se identificar um mapa  $\Phi$  no espaço dos atributos  $F$ , tal que a igualdade acima se verifique. O teorema de análise funcional denominado teorema de Mercer (BOSER; GUYON; VAPNIK, 1992) demonstra que se o *kernel*  $k$  é positivo definido, então existe um mapa  $\Phi$  no qual  $k(x_t^b, x_t^b) = \Phi(x_t^b), \Phi(x_t^b)$ . Um *kernel* é positivo definido se gerar uma matriz simétrica  $K$ , com elementos  $K_{ij} = k(x_i, x_j)$ , positiva definida, isto é, que possua todos os autovalores não negativos. Esse resultado respalda as versões não lineares via *kernels* dos algoritmos lineares, como a PCA e o *Statist*.

Boa parte das abordagens de MSPC está baseada na utilização da PCA. A PCA linear consiste na diagonalização da matriz de covariâncias  $S$ , que por sua vez pode ser escrita como uma matriz do tipo  $K$ . Tal modificação permite a definição de diferentes *kernels* positivo definidos, representados por produtos internos não canônicos, para extração de estruturas não lineares nos dados. Essa abordagem consagrada denomina-se KPCA (SCHOLKOPF; SMOLA; MULLER, 1998).

A análise *Statist* descrita na seção 2 consiste em diagonalizar as matrizes  $S\Delta$  (Equação 3) e  $WD$  (Equação 5, acrescida da matriz  $D$ ), originalmente do tipo  $K$ . Dessa forma, a exemplo da PCA, pode-se obter modificações não lineares nas estatísticas descritas na análise da interestrutura e da intraestrutura,

utilizadas no *Statist*. Particularmente, neste artigo propõe-se a inclusão de um *kernel* polinomial de segunda ordem nas estatísticas definidas no método *Statist* para capturar correlações não lineares quadráticas entre as variáveis de processo. Esse *kernel* é descrito na sequência.

Sem perda de generalização, considere um vetor de observações  $w = ([x]_1, [x]_2)$ , onde  $x \in \mathfrak{R}^2$  e um vetor  $\Phi(x) = ([x]_1^2, [x]_2^2, [x]_1[x]_2, [x]_2[x]_1)$ , onde  $\Phi(x) \in \mathfrak{R}^3$  é uma função de  $x$  cujos elementos representam todos os monômios de segunda ordem obtidos para os elementos em  $x$ . Através dessa função, as informações do vetor  $x$  passam a ser analisadas no espaço dos produtos de ordem 2 de seus elementos. No espaço  $F = \mathfrak{R}^3$ , utilizar o produto interno canônico entre vetores  $\Phi(x_i)$  e  $\Phi(x_j)$  equivale a multiplicar monômios de segunda ordem nos dados de entrada. Entretanto, podem-se computar esses produtos implicitamente definindo um *kernel* apropriado, tal que  $\Phi(x_i), \Phi(x_j) = k(x_i, x_j)$ . Para monômios desse tipo, o produto interno no espaço dos atributos

$F = \mathfrak{R}^3$  entre os vetores  $\Phi(x_i)$  e  $\Phi(x_j)$  é dado por

$$\Phi(x_i), \Phi(x_j) = [x_i]_1^2 [x_j]_1^2 + [x_i]_2^2 [x_j]_2^2 + 2[x_i]_1$$

$[x_j]_2 [x_i]_1 [x_j]_2 = x_i, x_j^2 = k(x_i, x_j)$ . Obtêm-se então os produtos internos entre monômios de segunda ordem no espaço dos atributos, sem utilizar  $\Phi(x)$  diretamente, através do *kernel* que calcula o quadrado do produto interno canônico entre as observações  $x$  originais, possibilitando investigar correlações não lineares quadráticas entre variáveis.

Esse resultado pode ser generalizado para o espaço  $\mathfrak{R}^N$  dos monômios ordenados de ordem  $d$ , aplicado aos elementos de  $x_i$  e  $x_j$ . Nesse caso, o *kernel* que calcula o produto interno entre os vetores  $\Phi(x_i)$  e  $\Phi(x_j)$  em  $F$  é dado por (POGGIO, 1975):

$$k(x_i, x_j) = \Phi(x_i), \Phi(x_j) = x_i, x_j^d \quad (10)$$

É importante notar que para determinada escolha de  $N$  e  $d$ , o espaço  $F$  dos atributos terá dimensão muito superior ao espaço original  $\mathfrak{R}^N$ , sendo dada por  $N_F = \binom{d+N-1}{d} = \frac{(d+N-1)!}{d!(N-1)!}$ . Entretanto, ressalta-se que através da utilização do produto interno modificado da Equação 10, o espaço  $F$  não é acessado diretamente.

### 3.1. Análise da interestrutura utilizando um kernel polinomial

Na seção 2 foi apresentada a base teórica do método *Statis* para a análise da interestrutura em um contexto de correlações lineares entre variáveis. Propõe-se agora uma modificação naquela análise redefinindo as estatísticas, que passam a ser descritas em função de um *kernel* polinomial. Esse procedimento será denominado *Kernel-Statis* (K-S).

Considere novamente os dados referentes a  $B$  bateladas. Têm-se então  $B$  matrizes  $X_b$ , de dimensão  $(T \times P)$ , compostas por vetores linha  $x_t^b$  de dados padronizados, de dimensão  $(1 \times P)$ , representando medições das  $P$  variáveis de processo no  $t$ -ésimo instante da  $b$ -ésima batelada, conforme apresentado na Figura 1. Define-se um *kernel* polinomial de ordem  $d$  (Equação 10) a partir do mapa não linear  $\Phi$  apresentado na Equação 9. Têm-se então vetores  $\Phi(x_t^b)$ , de dimensão  $(1 \times N_p)$ , cujos elementos representam todos os monômios de ordem  $d$  dos elementos dos vetores  $x_t^b$ .

Decorre disso que, no espaço  $F$ , a matriz  $X_b$  passa a ter dimensão  $(T \times N_p)$ , contendo assim  $T$  vetores linha  $\Phi(x_t^b)$ . Através da Equação 1 obtém-se a matriz quadrada  $W_b$ , de dimensão  $(T \times T)$ , onde cada elemento representa o produto interno canônico entre os vetores  $\Phi(x_t^b)$ , dois a dois. Através do *kernel* apresentado na Equação 10, define-se a matriz *kernel*/ $W_b$  (designada por  $\ddot{W}_b$ ) utilizando as matrizes  $X_b$  no espaço original das observações  $x_t^b$  [isto é,  $x_t^b \in \mathcal{R}^p$  e, assim,  $X_b$  novamente tem dimensão  $(T \times P)$ ]. A matriz apresentada na Equação 1 é então reescrita como:

$$\ddot{W}_b = \begin{bmatrix} \ddots & & \ddots \\ \cdot & k(x_t^b, x_{t'}^b) = \langle x_t^b, x_{t'}^b \rangle^d & \cdot \\ \cdot & & \ddots \\ \ddots & & \ddots \end{bmatrix} \text{ para } t, t' = 1, \dots, T \text{ e } (11)$$

$b = 1, \dots, B$

Cada elemento de  $\ddot{W}_b$  descreve uma medida de correlação temporal de ordem  $d$  entre os instantes  $t$  e  $t'$ , do conjunto das  $P$  variáveis de processo na  $b$ -ésima batelada. As matrizes  $\ddot{W}_b$  são normalizadas para utilização posterior, através da expressão  $\ddot{W}_b / \sqrt{\text{Tr}(D \times \ddot{W}_b \times D \times \ddot{W}_b)}$ . Na sequência, obtém-se a medida de similaridade entre as matrizes  $\ddot{W}_b$  e  $\ddot{W}_{b'}$ , reescrevendo a Equação 2 em função do produto modificado, resultando em  $\ddot{S}_{bb'} = \text{Tr}(D \times \ddot{W}_b \times D \times \ddot{W}_{b'})$ .

Conforme discutido na seção 2, quando se utiliza o produto interno usual  $k(x_t^b, x_{t'}^b) = \langle x_t^b, x_{t'}^b \rangle$

nos elementos de  $\ddot{W}_b$ ,  $\ddot{S}_{bb'}$  representa o quadrado do coeficiente de correlação linear de Pearson entre as  $P$  variáveis das matrizes  $X_b$  e  $X_{b'}$ . Já no caso da matriz  $\ddot{W}_b$ , definida na Equação 11, obtém-se uma medida  $\ddot{S}_{bb'}$  que representa o quadrado da correlação não linear (de ordem  $d$ ) entre as variáveis das matrizes  $X_b$  e  $X_{b'}$ . Em outras palavras, o valor de  $\ddot{S}_{bb'}$  indica o grau de semelhança, em um contexto não linear, entre os perfis das  $P$  variáveis, no conjunto dos instantes, em  $X_b$  e  $X_{b'}$ .

O resultado da análise da interestrutura das  $B$  matrizes  $X_b$  é armazenado na matriz  $S\Delta$  (Equação 3), reescrita da seguinte forma:

$$S\Delta = \begin{pmatrix} \ddots & & \ddots \\ \cdot & \frac{1}{B} \times \ddot{S}_{bb'} & \cdot \\ \cdot & & \ddots \\ \ddots & & \ddots \end{pmatrix}, \text{ onde } \sum_{b=1}^B 1/B = 1 \quad (12)$$

Na etapa seguinte, obtém-se uma representação resumida da interestrutura aplicando a PCA na matriz  $S\Delta$ . Analogamente ao que é feito no *Statis* linear em relação à matriz  $S\Delta$ , a posição das  $B$  bateladas no primeiro plano fatorial é definida pelos autovetores  $u_i$  associados aos dois maiores autovalores  $\lambda_i$  de  $S\Delta$ , adaptando a Equação 4.

### 3.2. Análise da intraestrutura utilizando um kernel polinomial

A análise da intraestrutura é realizada a partir da obtenção de uma matriz de compromisso  $W$  que fornece uma medida de correlação temporal entre os instantes  $t$  e  $t'$ , do conjunto das  $P$  variáveis de processo dentro da batelada padrão. Essa matriz é obtida através de uma combinação linear das matrizes  $W_b$  (Equação 5). A partir das matrizes  $\ddot{W}_b$ , definidas na Equação 11,  $W$  pode ser reescrita em função do *kernel* polinomial como  $\ddot{W} = \sum_{b=1}^B \alpha_b \times \ddot{W}_b$ , ou mais explicitamente como:

$$\ddot{W} = \begin{pmatrix} \ddots & & \ddots \\ \cdot & \sum_{b=1}^B \alpha_b k(x_t^b, x_{t'}^b) = \sum_{b=1}^B \alpha_b \langle x_t^b, x_{t'}^b \rangle^d & \cdot \\ \cdot & & \ddots \\ \ddots & & \ddots \end{pmatrix}, \quad (13)$$

$\text{para } t, t' = 1, \dots, T$

onde  $\ddot{W}$  é a matriz *kernel* de compromisso e  $\alpha_b$  são os pesos da combinação linear, dados pelos respectivos elementos do autovetor  $u_i$  (referente ao maior autovalor  $\lambda_i$ ) da matriz  $S\Delta$ .

A etapa seguinte consiste em buscar uma representação resumida da intraestrutura, aplicando



a PCA na matriz  $\ddot{W}$ . A correlação temporal não linear entre os instantes  $t$  e  $t'$  da batelada padrão é sumarizada no primeiro plano fatorial através dos autovetores  $\varepsilon_i$  associados aos dois maiores autovalores  $\delta_i$ , mediante adaptação da Equação 7.

#### 4. Gráficos de controle IS e $CO_t$ utilizando resultados da análise *Kernel-Statís*

A viabilização do monitoramento de novas bateladas utilizando a análise *Statís* se dá através do GC IS (derivado da análise da interestrutura) e dos GCs  $CO_t$  (derivados da análise da intraestrutura), conforme proposto por Fogliatto e Niang (2008).

O GC IS é obtido através da Equação 4, a partir das matrizes  $W_b$  e  $S\Delta$  descritas nas Equações 1 e 3, respectivamente. Esse gráfico permite verificar se a estrutura de correlação linear entre as  $P$  variáveis da nova batelada segue a estrutura de correlação linear padrão, capturada nas  $B$  bateladas de referência. O GC *kIS* (derivado do *Kernel-Statís*) realiza a mesma comparação levando em conta as correlações não lineares nos dados através das matrizes  $\ddot{W}_b$  e  $\ddot{S}\Delta$ , descritas nas Equações 11 e 12, respectivamente (no caso, correlações quadráticas através do *kernel* polinomial de segunda ordem).

Os GCs  $CO_t$  são obtidos através da Equação 8, a partir das matrizes  $W_b$  e  $W$  descritas nas Equações 1 e 5, respectivamente, e permitem verificar o comportamento temporal do conjunto das  $P$  variáveis de uma nova batelada em relação ao comportamento temporal esperado em função das  $B$  bateladas de referência. Analogamente ao GC *kIS*, os GCs  $kCO_t$  realizam esse monitoramento temporal considerando uma estrutura não linear nos dados, representada pelas matrizes  $\ddot{W}_b$  e  $\ddot{W}$  descritas nas Equações 11 e 13, respectivamente.

O primeiro passo consiste em obter uma região de controle para os GCs. Ao contrário do procedimento tradicional para GCs multivariados, a região de controle será determinada através de uma abordagem não paramétrica. O procedimento apresentado a seguir é uma adaptação proposta por Fogliatto e Niang (2008), para o contexto de GCs IS e  $CO_t$ , do procedimento em Zani, Riani e Corbellini (1998).

Considerando a representação sumária das  $B$  bateladas de referência no primeiro plano fatorial dos GCs, têm-se  $B$  pares ordenados  $(a_{1,b}, a_{2,b})$  para o GC *kIS*, e  $(z_{1,t}^b, z_{2,t}^b)$  para os GCs  $kCO_t$ . Para exemplificar o procedimento, utilizam-se os vetores  $(z_{1,t}^b, z_{2,t}^b)$  na determinação da região de controle dos GCs  $kCO_t$ . Inicialmente, calcula-se o ponto

que representa o vetor de média  $\begin{pmatrix} -b \\ z_{1,t}, z_{2,t} \end{pmatrix}$ . A seguir, estima-se a distância de Mahalanobis entre os vetores  $z_t^b = (z_{1,t}^b, z_{2,t}^b)$  e  $\bar{z}_t = \begin{pmatrix} -b \\ z_{1,t}, z_{2,t} \end{pmatrix}$ , dada por  $D_b = \left( z_t^b - \bar{z}_t \right) H^{-1} \left( z_t^b - \bar{z}_t \right)$  para  $(b = 1, \dots, B)$ , onde  $\begin{pmatrix} z_t^b - \bar{z}_t \end{pmatrix}$  representa o vetor linha de diferenças entre os vetores  $z_t^b$  e  $\bar{z}_t$ , cujo transposto é dado por  $\begin{pmatrix} z_t^b - \bar{z}_t \end{pmatrix}$ , e  $H^{-1}$  é a matriz inversa da matriz  $H$  de covariâncias entre vetores.

A seguir, as  $B$  distâncias  $D_b$  são ordenadas em ordem crescente e 50% das menores distâncias são retidas. Os vetores  $z_t^b$  correspondentes formarão o *convex hull* (polígono) de abrangência 50% no primeiro plano fatorial. Obtém-se então a expansão da região formada pelo *convex hull* a partir de um fator de escala. Para tanto, define-se um múltiplo  $l$  da distância  $D_b$  entre o centróide (representado pelo vetor  $\bar{z}_t$ , obtido a partir dos vetores representados pelos pontos internos do polígono) e os pontos limítrofes do polígono de abrangência 50%. O valor de  $l$  é determinado a partir da probabilidade de alarme falso  $\alpha$  (ou erro do tipo I) desejada para o GC, supondo vetores  $z_t^b$  no interior do polígono de abrangência 50% distribuídos segundo uma normal bivariada. Finalmente, uma curva do tipo *B-spline* (ver HASTIE; TIBSHIRANI; FRIEDMAN, 2001) é ajustada aos novos pontos limítrofes, que formam as arestas do polígono externo. A definição da região de controle do GC *kIS* ocorre de maneira análoga, a partir dos pontos  $(a_{1,b}, a_{2,b})$ .

Finalmente, a determinação do erro relativo da representação sumária da interestrutura e da intraestrutura das  $B$  bateladas de referência no primeiro plano fatorial é dada pelas respectivas funções perda:  $FP_{INTER} = \left( \sum_{i=3}^B \lambda_i / \sum_{i=1}^B \lambda_i \right) \times 100$  e  $FP_{INTRA} = \left( \sum_{i=3}^T \delta_i / \sum_{i=1}^T \delta_i \right) \times 100$ .

##### 4.1. Controle on-line

O controle *on-line* de uma nova batelada  $(B+1)$  é realizado durante sua progressão no tempo a cada instante  $t^* \leq T$ , onde  $t^*$  corresponde ao mais recente instante de tempo no qual as  $P$  variáveis de processo foram amostradas. Assim, apenas uma parte da matriz de dados  $X_{B+1}$  estará disponível e as informações faltantes deverão ser estimadas. Para tanto, o procedimento adotado é análogo àquele proposto por Fogliatto e Niang (2008) para os GCs IS e  $CO_t$ .

Inicialmente, define-se a matriz incompleta de dados da seguinte forma:

$$X_{B+1} = \begin{bmatrix} x_1^{B+1} \\ \vdots \\ x_g^{B+1} \\ \vdots \\ x_{t^*}^{B+1} \\ x_{t^*+1}^{B+1} \\ \vdots \\ x_q^{B+1} \\ \vdots \\ x_T^{B+1} \end{bmatrix} \quad (14)$$

onde os vetores linha  $x_g^{B+1}$ , de dimensão  $(1 \times P)$ , para  $1 \leq g \leq t^*$ , contêm as observações disponíveis das  $P$  variáveis até o instante atual  $t^*$  da batelada, e  $x_q^{B+1}$ , para  $t^* + 1 \leq q \leq T$ , representam vetores de observações ainda não disponíveis, relativas aos instantes futuros da batelada em curso.

No passo seguinte, obtém-se a matriz  $\ddot{W}_{B+1} = X_{B+1}(X_{B+1})^T$ , de dimensão  $(T \times T)$ , contendo os produtos internos entre os vetores linha de  $X_{B+1}$  dados pelo *kernel* selecionado. Entretanto, diferentemente da matriz  $\ddot{W}_b$ , obtida a partir da matriz completa  $X_b$ , a matriz  $\ddot{W}_{B+1}$  apresenta apenas uma submatriz, de dimensão  $(t^* \times t^*)$ , trazendo as informações disponíveis até a  $t^*$ -ésima linha da matriz  $X_{B+1}$ . Os elementos das demais submatrizes de  $\ddot{W}_{B+1}$  dependem de  $x_q^{B+1}$  e são estimados a partir da matriz de compromisso  $\ddot{W}$  na Equação 13. A ideia é verificar o comportamento temporal da batelada nova até o instante atual  $t^*$ , assumindo que nos próximos instantes da batelada as variáveis sigam o comportamento apresentado nas  $B$  bateladas de referência. Explicitamente, pode-se escrever a matriz  $\ddot{W}_{B+1}$  particionada como se segue:

$$W_{b+1} = \begin{pmatrix} \ddots & & & & \ddots \\ & k(x_g^{B+1}, x_g^{B+1}) & & & \sum_{b=1}^B \alpha_b k(x_g^b, x_q^b) \\ & \vdots & & & \vdots \\ & & \sum_{b=1}^B \alpha_b k(x_q^b, x_g^b) & & \vdots \\ & & \vdots & & \sum_{b=1}^B \alpha_b k(x_q^b, x_q^b) \\ \ddots & & & & \ddots \end{pmatrix} \quad (15)$$

Observa-se que os elementos das submatrizes, que deveriam ser obtidos pelos produtos internos  $k(x_g^{B+1}, x_q^{B+1})$ ,  $k(x_q^{B+1}, x_g^{B+1})$  e  $k(x_q^{B+1}, x_q^{B+1})$ , não estão disponíveis, sendo assim estimados pela média dos produtos internos  $k(x_g^{B+1}, x_q^{B+1})$  e

$k(x_q^{B+1}, x_q^{B+1})$ , respectivamente, descritos na matriz de compromisso  $\ddot{W}$ . A exemplo das matrizes  $\ddot{W}_b$ , a matriz  $\ddot{W}_{B+1}$  é normalizada, sendo reescrita como  $\ddot{W}_{B+1} / \sqrt{\text{Tr}(D \times \ddot{W}_{B+1} \times D \times \ddot{W}_{B+1})}$ .

A operacionalização do monitoramento da nova batelada  $X_{B+1}$  através do GC  $kIS$  é realizada utilizando a matriz  $\ddot{S}\Delta_{B+1}$ , que corresponde à matriz  $\ddot{S}\Delta$  (Equação 12) adicionada da coluna  $(B + 1)$  e da linha  $(B + 1)$ , contendo as similaridades entre os pares de matrizes  $\ddot{W}_b$  e  $\ddot{W}_{B+1}$ . Após realizar a PCA em  $\ddot{S}\Delta_{B+1}$ , o par de valores que representa a projeção da nova batelada no GC  $kIS$  é dado por:

$$a_{B+1} = (a_{1,B+1}, a_{2,B+1}) = (\sqrt{\lambda_1} u_{1,B+1}, \sqrt{\lambda_2} u_{2,B+1}) \quad (16)$$

onde  $u_{1,B+1}$  e  $u_{2,B+1}$  representam, respectivamente, o  $(B+1)$ -ésimo elemento dos autovetores  $u_1$  e  $u_2$ , e  $\lambda_1$  e  $\lambda_2$  os dois maiores autovalores associados. Monitora-se, dessa forma, a correlação não linear entre as  $P$  variáveis da batelada nova  $X_{B+1}$  e das bateladas de referência até o instante  $t^*$ .

O monitoramento utilizando os GCs  $kCO_t$  é realizado utilizando os vetores linha  $w_t^{B+1}$  da matriz  $\ddot{W}_{B+1}$ . Obtém-se as coordenadas para representação da nova batelada  $X_{B+1}$  no  $t$ -ésimo instante reescrevendo a Equação 8 como:

$$z_t^{B+1} = (z_{1,t}^{B+1}, z_{2,t}^{B+1}) = \left( \frac{1}{\sqrt{\delta_1}} \frac{1}{T} w_t^{B+1} \varepsilon_1, \frac{1}{\sqrt{\delta_2}} \frac{1}{T} w_t^{B+1} \varepsilon_2 \right), \quad (17)$$

para  $t = 1, \dots, T$

onde  $\varepsilon_1$  e  $\varepsilon_2$  representam os autovetores gerados da PCA realizada na matriz *kernel* de compromisso  $\ddot{W}$ , associados aos dois maiores autovalores  $\delta_1$  e  $\delta_2$ , respectivamente. Monitora-se dessa forma o comportamento temporal não linear das  $P$  variáveis no transcurso da batelada até o instante  $t^*$ .

## 4.2. Resumo da metodologia de controle via GCs $kIS$ e $kCO_t$

Esta seção apresenta os procedimentos para construção dos GCs  $kIS$  e  $IS$  para o monitoramento *on-line* de novas bateladas.

### 4.2.1. Construção dos GCs $kIS$ e $kCO_t$

- Obtenção de dados de um conjunto de  $B$  bateladas de referência completadas  $X_b$ .
- Padronização dos dados em  $X_b$ , subtraindo cada valor da matriz pela média da sua coluna e dividindo pelo desvio padrão da coluna.

- Obter as matrizes  $\ddot{W}_b$  utilizando o *kernel* apropriado (neste artigo, propõe-se um *kernel* polinomial de ordem  $d$ ).
- Obter as matrizes  $\ddot{W}_b$  padronizadas (isto é,  $\ddot{W}_b / \sqrt{\text{Tr}(D \times \ddot{W}_b \times D \times \ddot{W}_b)}$ ).
- Obter a matriz  $\Delta$  de correlações  $S_{bb}$ , entre os dados de  $X_b$  e  $X_b$ .
- Aplicar PCA em  $\Delta$  e armazenar os dois maiores autovalores  $\lambda_i$  e os autovetores associados  $u_i$  (análise *Statis* da interestrutura).
- Obter a posição das  $B$  bateladas  $X_b$  no primeiro plano fatorial do GC  $kIS$  através dos pares ordenados  $(a_{1,b}, a_{2,b})$  obtidos através da Equação 4.
- Obter a matriz de compromisso  $\ddot{W} = \sum_{b=1}^B \alpha_b \times \ddot{W}_b$ , onde  $\alpha_b = \frac{1}{\sqrt{\lambda_1}} \frac{1}{B} u_{1,b}$ .
- Aplicar a PCA em  $\ddot{W}$  e armazenar os dois maiores autovalores  $\varepsilon_i$  e os autovetores associados  $\delta_i$  (análise *Statis* da intraestrutura).
- Obter a posição das  $B$  bateladas  $X_b$  no primeiro plano fatorial dos GCs  $kCO_t$  através dos pares ordenados  $(z_{1,t}^b, z_{2,t}^b)$ , obtidos na Equação 7.
- Determinar a região de controle para os GCs  $kIS$  e  $kCO_t$ , ajustando uma curva do tipo B-spline a partir das  $B$  coordenadas  $(a_{1,b}, a_{2,b})$  e  $(z_{1,t}^b, z_{2,t}^b)$ , respectivamente.

#### 4.2.2. Monitoramento *on-line* de uma nova batelada

- Obtenção de dados da nova batelada  $X_{B+1}$  em progressão no instante  $t^*$ .
- Obter a matriz  $\ddot{W}_{B+1}$  através da Equação 15 e normalizá-la através da operação  $\ddot{W}_{B+1} / \sqrt{\text{Tr}(D \times \ddot{W}_{B+1} \times D \times \ddot{W}_{B+1})}$ .
- Obter a matriz  $\ddot{\Delta}_{B+1}$  que inclui as correlações  $S_{b,B+1}$  entre os dados de  $X_b$  e  $X_{B+1}$ .
- Aplicar PCA em  $\ddot{\Delta}_{B+1}$  e armazenar os dois maiores autovalores  $\lambda_i$  e os autovetores associados  $u_i$ .
- Obter a posição da nova batelada  $X_{B+1}$  no GC  $kIS$  através dos pares ordenados  $a_{B+1} = (a_{1,B+1}, a_{2,B+1})$ , obtidos na Equação 16.
- Obter a posição da nova batelada  $X_{B+1}$  no GC  $kCO_t$  através dos pares ordenados  $\ddot{W} z_t^{B+1} = (z_{1,t}^{B+1}, z_{2,t}^{B+1})$ , obtidos na Equação 17.

### 5. Aplicação dos gráficos propostos em um conjunto simulado de dados

Considere um processo industrial em bateladas com desempenho determinado pelo comportamento

de duas variáveis correlacionadas  $X_1$  e  $X_2$ . Suponha que as leis físicas que regem esse processo sejam descritas pelo seguinte sistema de equações diferenciais:

$$\begin{cases} \dot{x}_1 = a(x_1 - c_1) - b(x_2 - c_2) \\ \dot{x}_2 = b(x_1 - c_1) + a(x_2 - c_2) + nl(x_1 - c_1)(x_2 - c_2) \end{cases} \quad (18)$$

onde  $a, b$  e  $nl$  são constantes reais e os pontos sobre as variáveis denotam derivadas temporais de  $X_1$  e  $X_2$ . O sistema na Equação 18 é uma perturbação não linear do seguinte sistema linear:

$$\begin{cases} \dot{x}_1 = a(x_1 - c_1) - b(x_2 - c_2) \\ \dot{x}_2 = b(x_1 - c_1) + a(x_2 - c_2) \end{cases} \quad (19)$$

O sistema na Equação 19 tem o ponto  $(c_1, c_2)$  como ponto de equilíbrio. Os dois autovalores associados são números complexos; i.e.,  $\lambda_{1,2} = a \pm ib$ . Assim, tem-se um comportamento oscilatório em torno do ponto de equilíbrio  $(c_1, c_2)$ , que é estável se  $a < 0$  e instável se  $a > 0$ . O coeficiente  $nl$  define o grau de perturbação imposto na não linearidade.

Para transformar a Equação 19 em uma forma iterativa, adotou-se o esquema de Euler (PATEL, 1993), o que as transforma em:

$$\begin{cases} x_{1t+1} = x_{1t} + [a(x_{1t} - c_1) - b(x_{2t} - c_2)] \Delta t \\ x_{2t+1} = x_{2t} + [b(x_{1t} - c_1) + a(x_{2t} - c_2) + nl(x_{1t} - c_1)(x_{2t} - c_2)] \Delta t \end{cases} \quad (20)$$

As simulações dos dados através do sistema apresentado na Equação 20, bem como os cálculos necessários para construção dos GCs  $kIS$  e  $IS$  detalhados na seção 4.2, foram realizados por meio de rotinas desenvolvidas no *software Matlab*, as quais encontram-se disponíveis no Apêndice 1.

Para as simulações das bateladas de referência, foram adotados os seguintes valores para os coeficientes da Equação 20:  $a = -1, b = 2, c_1 = 10, c_2 = 20$  e diferentes valores para  $nl$ , descritos mais adiante. Neste estudo,  $\Delta t$  é suficientemente pequeno tal que a Equação 20 seja uma aproximação do sistema contínuo na Equação 18. Essa configuração gerou duas variáveis de processo com trajetórias similares àquelas observadas em processos industriais reais.

As trajetórias de  $X_1$  e  $X_2$  foram amostradas em  $T = 20$  instantes, igualmente espaçados, em bateladas distintas, a partir do sistema descrito na Equação 20. Pequenas variações foram impostas nas condições iniciais, obtendo-se assim bateladas representativas do processo sob controle estatístico. A Figura 4 apresenta as séries temporais das

trajetórias das variáveis em 100 bateladas simuladas ( $B = 100$ ), para  $nl = 2$ . A partir dos dados gerados, a estratégia de simulação utilizada busca atingir dois objetivos:

- Verificar se o *Kernel-Statist* oferece melhor caracterização da estrutura de correlação não linear (quadrática) existente entre as variáveis, se comparada ao *Statist* tradicional. Visto que a análise da interestrutura investiga diretamente as correlações entre as variáveis no conjunto dos instantes, utilizam-se os GCs *kIS* e *IS* para realizar essa comparação.
- Analisar o desempenho do *Kernel-Statist* em relação ao *Statist* tradicional na detecção de bateladas novas com descontroles não lineares impostos sobre o sistema, a cada instante de tempo. Esse objetivo é alcançado comparando os GCs  $kCO_t$  e  $CO_t$ , oriundos da análise da intraestrutura, visto que esta investiga a correlação temporal do conjunto das variáveis.

Inicia-se apresentando o estudo comparativo entre GCs *kIS* e *IS*, relacionado ao primeiro objetivo acima listado. O GC *IS* (derivado da análise da interestrutura tradicional) descreve resumidamente o quadrado da estrutura de correlação linear entre as variáveis de diferentes pares de bateladas. A correlação não linear quadrática presente nos dados é representada nos GCs *kIS*, construídos com a utilização do *kernel* polinomial de segunda ordem da Equação 10, com  $d = 2$ .

Os dados gerados pelo sistema da Equação 20 foram analisados utilizando a abordagem *Statist* tradicional e a abordagem que utiliza o *kernel* polinomial de segunda ordem. Para cada valor de  $nl$ , foram simuladas 100 bateladas, mantendo os valores definidos acima para os demais parâmetros. A seguir, foram construídos os GCs *kIS*, com probabilidade nominal de alarme falso  $\alpha = 0,01$ . No passo seguinte, essas mesmas 100 foram consideradas como novas bateladas  $X_{B+1}$  (Equação 14) sob monitoramento e projetadas nos respectivos GCs *kIS*, através da Equação 16. Esse procedimento foi repetido para a análise *Statist* tradicional. Analogamente, para cada  $nl$ , partindo das 100 bateladas de referência, o GC *IS* foi obtido a partir das matrizes  $W_b$  e  $S\Delta$ , utilizando o produto interno usual. Depois, para cada  $nl$ , as bateladas foram consideradas como novas bateladas  $X_{B+1}$  e projetadas nos respectivos GCs *IS* através da Equação 16, sem a utilização de *kernels*. A Tabela 1 apresenta o quadro comparativo do desempenho das duas abordagens.

Sabe-se que o valor de  $nl$  representa o peso dado ao termo não linear quadrático do sistema na Equação 20. Através da Tabela 1, nota-se que para  $nl = 0$  (sistema estritamente linear) a frequência

relativa de bateladas conformes mal classificadas não difere muito do valor nominal ( $\alpha = 0,01$ ), tanto nos GCs *kIS* (1/100) quanto nos GCs *IS* (3/100), embora a ela seja maior nesse último grupo. Em contrapartida, para  $nl > 0$  pode-se observar um aumento na frequência relativa de alarme falso nos GCs *IS*, enquanto nos GCs *kIS* mantém-se no mesmo patamar. Isso evidencia a caracterização mais completa dos dados através da inclusão do *kernel* polinomial de segunda ordem na análise *Statist* tradicional. Mais especificamente, a inclusão dos batimentos (produtos) de ordem 2 das variáveis  $X_1$  e  $X_2$  capturam, em algum grau, as não linearidades geradas pelo termo de batimento  $nlx_{1t}x_{2t}$  presente no sistema. Adicionalmente, à medida que o termo de batimento se faz mais presente nos dados com o aumento no valor de  $nl$ , o ganho de precisão do GC *kIS* em relação ao GC *IS* torna-se mais destacado, conforme evidenciado pela Função Perda.

A seguir, apresenta-se o estudo comparativo entre os GCs  $kCO_t$  e  $CO_t$ , relacionado ao segundo objetivo desse estudo simulado; tais GCs são derivados da análise da intraestrutura dos dados. Sabe-se que através dos GCs  $CO_t$  obtém-se uma caracterização resumida do comportamento temporal do conjunto das variáveis. Os GCs  $kCO_t$  (utilizando novamente o *kernel* polinomial de segunda ordem na análise) descrevem resumidamente o comportamento temporal não linear do conjunto das variáveis.

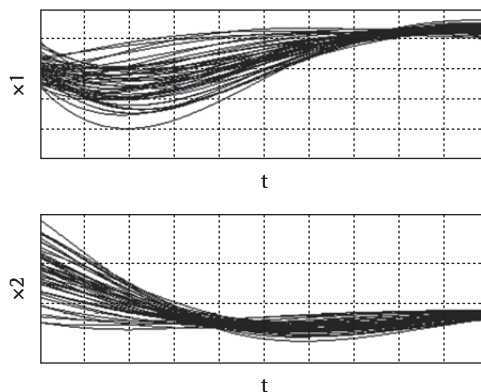


Figura 4. Trajetórias das duas variáveis de processo amostradas em 100 bateladas de referência.

Tabela 1. Quadro comparativo do desempenho dos GCs *kIS* e *IS* para  $\alpha=0,01$ .

nl	GCs <i>kIS</i>		GCs <i>IS</i>	
	Bateladas mal classificadas	Função Perda (FP <sub>inter</sub> )	Bateladas mal classificadas	Função Perda (FP <sub>inter</sub> )
0	1/100	0,0015	3/100	0,0025
1	1/100	0,0021	5/100	0,0035
2	2/100	0,0026	8/100	0,0057
3	1/100	0,0031	9/100	0,0069

Utilizou-se a abordagem tradicional e via *kernel* polinomial para análise dos dados. Novamente, 10 bateladas de referência foram geradas, mantendo a mesma configuração de parâmetros descrita no início da seção 5, com o termo não linear  $nl=0$ . A partir daí foram construídos os GCs  $kCO_t$  utilizando a abordagem descrita na seção 4. De modo análogo, os GCs  $CO_t$  foram obtidos através das matrizes  $W_b$  e  $W$ . A seguir, 10 bateladas foram simuladas com perturbações impostas na não linearidade, do instante 10 ao instante 14. Durante esses instantes, o valor de  $nl=0$  foi substituído por  $nl=3$ . As bateladas foram projetadas nos GCs  $kCO_t$  e  $CO_t$  através da Equação 17. A Figura 5 apresenta o monitoramento simultâneo das 10 bateladas utilizando os GCs  $CO_t$  e  $kCO_t$ , respectivamente.

Os GCs  $CO_t$  (Figura 5a) e  $kCO_t$  (Figura 5b) apresentam a projeção das bateladas novas em ordem cronológica, ao longo das linhas. Observa-se que, em ambas as abordagens, o descontrole é acusado corretamente a partir do instante 10. Entretanto, observa-se a pouca precisão nos GCs  $CO_t$  para detectar que o processo retornou ao estado sob controle no instante 15, visto que as bateladas aparecem em sua maioria fora da região de controle após esse instante. Os GCs  $kCO_t$ , por sua vez, identificam que o processo está sob controle a partir do instante 15 em todas as bateladas verificadas (quando, de fato, cessaram as perturbações), exceto no último instante quando gerou alguns alarmes falsos. Esses resultados evidenciam novamente um ganho na caracterização do sistema com a utilização do *kernel* polinomial. A Função Perda associada ratifica os resultados obtidos, pois  $FP_{INTRA} = 0,0024$  para os GCs  $kCO_t$  ao passo que  $FP_{INTRA} = 0,0059$  para os GCs  $CO_t$ .

### 5.1. Considerações sobre a metodologia Kernel-*Statis*

A seção 5 apresentou um exemplo simulado de aplicação dos GCs baseados no *Kernel-*Statis**. Entretanto, a consolidação e posterior utilização prática dessa metodologia estão vinculadas a pesquisas suplementares incluindo os seguintes pontos: i) metodologias que permitam a escolha de um *kernel* conveniente a partir da investigação a priori do tipo de não linearidade presente nos dados do processo; ii) modificação na metodologia proposta permitindo a construção de GCs para processos em bateladas de tempo variável; iii) desenvolvimento de ferramentas de diagnóstico, que permitam melhor compreensão de sinalizações de descontrole apontadas nos gráficos; e iv) desenvolvimento de *software* que incorpore a teoria proposta. Uma breve discussão sobre essas questões é apresentada a seguir.

No caso do exemplo apresentado na seção 5, simulou-se um processo cujas leis físicas estavam descritas por um sistema onde foram impostas não linearidades do tipo quadráticas. Nesse caso, utilizou-se um *kernel polinomial de segunda ordem* (em combinação com o *Statis*) para caracterizar a estrutura dos dados gerados. A utilização de uma função não linear polinomial se baseia na teoria das expansões de funções em séries de Taylor. É natural, assim, considerar os termos quadráticos como aqueles não lineares mais relevantes, seguidos dos termos cúbicos e de mais alto grau. Entretanto, em um processo real, não se sabe a priori qual tipo de estrutura não linear está presente nos dados. Dessa forma, necessita-se de uma etapa de pré-processamento para investigar os dados e, a partir disso, propor um *kernel* adequado.

A metodologia proposta se aplica a processo em bateladas de tempo fixo. Acredita-se que pequenas modificações permitirão sua aplicação no controle de bateladas de tempo variável. Fogliatto e Niang (2008) apresentaram um procedimento de emparelhamento das bateladas para posterior aplicação dos GC via *Statis*. A abordagem proposta cumpre uma etapa de pré-processamento dos dados estimando as observações faltantes das bateladas em relação àquela que durou mais tempo através de modelos de séries temporais. Tal procedimento pode ser implementado na metodologia *Kernel-*Statis**.

Através dos GCs propostos foi atingido o primeiro objetivo do MSPC, que consiste em sinalizar o quanto antes desvios significativos das variáveis em torno de suas trajetórias de referência. Entretanto, em um segundo momento, necessita-se diagnosticar as causas dos pontos fora de controle assinalados pelos GCs, identificando as variáveis que apresentaram maiores desvios em torno dos seus padrões. Esse segundo objetivo não foi contemplado no presente artigo; porém os autores estão desenvolvendo uma proposta baseada no uso do *Statis Dual*. A abordagem, ainda não validada, basicamente utiliza a análise *Statis* descrita na seção 2 a partir de matrizes  $V_b = X_b'X_b$  no lugar de  $W_b = X_bX_b'$ . A matriz  $V_b$  representa a correlação das variáveis no conjunto dos instantes e dessa forma prioriza uma descrição do comportamento de cada variável sobre todos os instantes. As matrizes  $V_b$  permitem a construção de GCs  $CO_p$  (assim como os GCs  $CO_t$ , oriundos das matrizes do tipo  $W_b$ ). A metodologia completa utilizará então os GCs  $kIS$  e  $kCO_t$  para monitoramento, em conjunto com os GCs  $CO_p$  para fins de diagnóstico.

O Apêndice traz a rotina escrita no *software Matlab* para construção dos GCs  $kIS$  e  $kCO_t$ , e o monitoramento *on-line* de uma batelada nova. Entretanto, para o método proposto ser totalmente

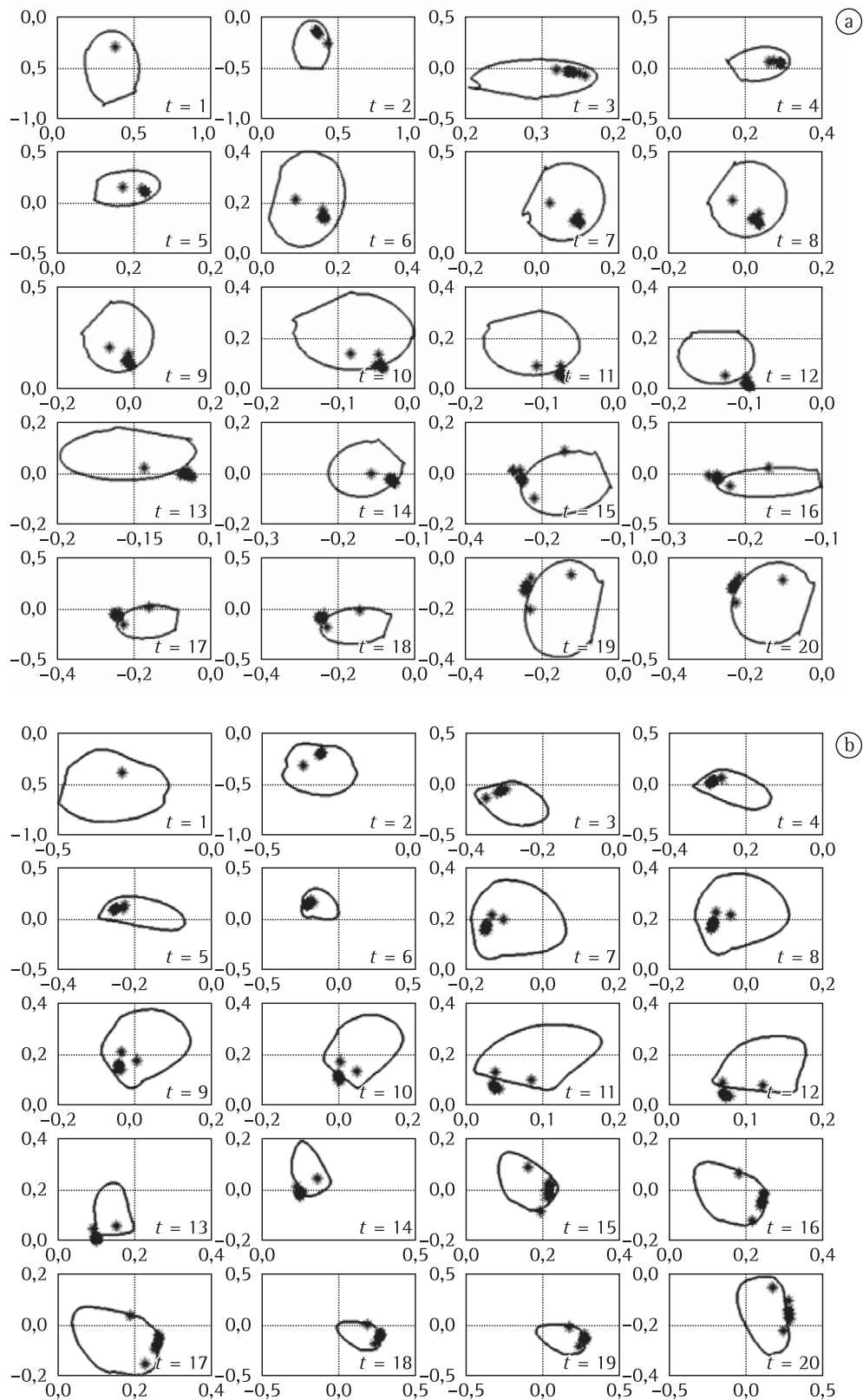


Figura 5. a) GCs  $CO_2$ ; e b) GCs  $kCO_2$ .

aplicável é essencial uma automação plena através do desenvolvimento de uma plataforma computacional que ofereça uma interface amigável com operadores, apresentando os GCs e permitindo atualizações constantes de dados de bateladas em progressão.

## 6. Conclusões

Neste artigo, foram propostos gráficos de controle multivariados baseados no *Kernel-Statís* para monitoramento de processos em bateladas, com variáveis apresentando correlações não lineares. Esses gráficos trazem duas contribuições para a área de controle multivariado da qualidade: i) incorporam e representam de maneira resumida as informações oriundas de grandes massas de dados multivariados, considerando a estrutura de correlação e autocorrelação não linear das variáveis de processo; e ii) oferecem diagnósticos precisos que permitem uma rápida ação corretiva quando o processo estiver fora de controle.

Os desenvolvimentos propostos no artigo são aplicados a um processo, com dados simulados a partir de um sistema não linear de duas variáveis. Através do estudo, validou-se o *Kernel-Statís* e verificou-se o ganho de precisão de tal procedimento em relação ao *Statís* usual no monitoramento de bateladas futuras. Verificou-se que os GCs  $kIS$  e  $kCO_2$  ofereceram uma caracterização do processo superior àquela apresentada pelos GCs  $IS$  e  $CO_2$ , na medida em que não linearidades quadráticas tornaram-se pronunciadas com mais intensidade no sistema proposto.

Extensões naturais da pesquisa aqui apresentada incluiriam: i) metodologias que permitam a escolha de um *kernel* conveniente a partir da investigação a priori do tipo de não linearidade presente nos dados do processo; e ii) desenvolvimento de ferramentas de diagnóstico, que permitam melhor compreensão de sinalizações de descontrole apontadas nos gráficos.

## Referências

- BOSER, B. E.; GUYON, I. M.; VAPNIK, V. A training algorithm for optimal margin classifiers. In: ANNUAL WORKSHOP ON COMPUTATIONAL LEARNING THEORY, 15., 1992, Pittsburgh. *Anais...*
- CHO, J. H. et al. Fault identification for process monitoring using kernel principal component analysis. *Chemical Engineering Science*, v. 60, p. 279-288, 2005.
- CHOI, S. W.; MORRIS, J.; LEE, I.-B. *Nonlinear* multiscale modelling for fault detection and identification. *Chemical Engineering Science*, in press. doi: 10.1016/j.ces.2008.01.022, 2008.
- CUI, P.; LI, J.; WANG, G. Improved kernel principal component analysis for fault detection. *Expert Systems with Applications*, v. 34, p. 1210-1219, 2008.
- DONG, D.; MCAVOY, T. J. *Nonlinear* principal components analysis based on principal curves and neural networks. *Computers and Chemical Engineering*, v. 20, p. 65-78, 1996.
- ESCOUFIER, Y. Three-mode data analysis: the STATIS method. In: FICHET, B.; LAURO, N. C. (Eds.). *Methods for multidimensional data analysis*. ECAS, 1987. p. 259-272.
- FOGLIATTO, F. S.; NIANG, N. Controle multivariado de processos em batelada com duração variada. *Revista Produção*, v. 18, p. 240-259, 2008.
- FOGLIATTO, F. S.; NIANG, N. Multivariate statistical control of batch processes with variable duration. In: INTERNATIONAL CONFERENCE ON INDUSTRIAL ENGINEERING AND ENGINEERING MANAGEMENT - IEEM, 2009, Hong Kong. *Anais...*
- GENG, Z. Q.; ZHU, Q. X. Multiscale *nonlinear* principal component analysis (NLPCA) and its application for chemical process monitoring. *Industrial and Engineering Chemistry Research*, v. 44, p. 3585-3593, 2005.
- HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. *The elements of statistical learning - Data mining, inference, and prediction*. New York: Springer Science, 2001.
- HIDEN, H. G. et al. *Nonlinear* principal components analysis using genetic programming. *Computers and Chemical Engineering*, v. 23, p. 413-425, 1999.
- JACKSON, J. E.; MUDHOLKAR, G. S. Control Procedures for Residuals Associated with Principal Component Analysis. *Technometrics*, v. 21, n. 3, p. 341-349, 1979.
- JIA, F.; MARTIN, E. B.; MORRIS, A. J. *Non-linear* principal components analysis with application to process fault detection. *International Journal of Systems Science*, v. 31, p. 1473-1487, 2001.
- JOLLIFFE, I. T. *Principal component analysis*. 2<sup>nd</sup> ed. New York: Springer, 2004.
- KOURTI, T. Multivariate dynamic data modeling for analysis and statistical process control of batch process, start-ups and grade transitions. *Journal of Chemometrics*, v. 17, p. 93-109, 2003.
- KOURTI, T.; MacGREGOR, J. F. Multivariate SPC Methods for Process and Product Monitoring. *Journal of Quality Technology*, v. 28, n. 4, p. 409-428, 1996.
- KRAMER, M. A. *Nonlinear* principal component analysis using autoassociative neural networks. *AIChE Journal*, v. 37, p. 233-243, 1991.
- LAVIT, C. et al. The ACT (STATIS method). *Computational Statistics & Data Analysis*, v. 19, p. 97-119, 1994.
- LEE, J.; YOO, C. K.; LEE, I.-B. *Nonlinear* process monitoring using kernel principal component analysis. *Chemical Engineering Science*, v. 59, p. 223-234, 2004.
- LEE, J.-M. et al. Fault detection of batch processes using multiway kernel principal component analysis. *Computers & Chemical Engineering*, v. 28, p. 1837-1847, 2004.
- MacGREGOR, J. F. Using *on-line* process data to improve quality: challenges for statisticians. *International Statistical Review*, v. 65, p. 309-323, 1997.
- MARTIN, E. B.; MORRIS, A. J. An overview of multivariate statistical process control in continuous and batch process performance monitoring. *Transactions of the Institute of Measurement and Control*, v. 18, n. 1, p. 51-60, 1996.
- NOMIKOS, P.; MCGREGOR, J. F. Multivariate SPC charts for monitoring batch processes. *Technometrics*, v. 37, p. 41-59, 1995.
- PATEL, V. A. A. *Numerical Analysis*. New York: Saunders College Publishing, 1993.
- POGGIO, T. On optimal *nonlinear* associative recall. *Biological Cybernetics*, v. 19, p. 201-209, 1975.
- SCEPI, G. Parametric and non parametric multivariate quality control charts. In: LAURO, C. et al. (Eds.). *Multivariate Total Quality Control*. Physica-Verlag, 2002. p. 163-189.
- SCHOLKOPF, B.; SMOLA, A. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. Cambridge: MIT Press, 2002.
- SCHOLKOPF, B.; SMOLA, A.; MULLER, K. L. *Nonlinear* component analysis as a kernel eigenvalue problem. *Neural Computation*, v. 10, p. 1299-1399, 1998.
- ZANI, S.; RIANI, M.; CORBELLINI, A. Robust bivariate boxplots and multiple outlier detection. *Computational Statistics & Data Analysis*, v. 28, p. 257-270, 1998.

## Multivariate control charts for monitoring non-linear batch processes

### Abstract

Industrial batch processes are widely used in the production of certain items. Such processes provide a peculiar data structure; therefore there is a growing interest in the development of customized multivariate control charts for their monitoring. We investigate a recent approach that uses control charts based on the Statis method. Statis is an exploratory technique for measuring similarities between data matrices. However, the technique only assesses similarities in a linear context, i.e. investigating structures of linear correlation in the data. In this paper we propose control charts based on the Statis method in conjunction with a kernel for monitoring processes in the presence of strong nonlinearities. Through kernels we define nonlinear functions of data for better representing the structure to be characterized by the Statis method. The new approach, named Kernel-Statis, is developed and illustrated using simulated data.

### Keywords

Multivariate quality control. Control charts. Batch processes. Kernel. Statis method.

Apêndice 1. Rotina escrita no *software Matlab* para os GCs IS e CO<sub>1</sub> e para o monitoramento *on-line*.

Lê os dados simulados, realiza a análise *kernel-statis* e salva as coordenadas num arquivo

```
Dados1 =load('KernelReferencias.dat');
Ntotal=size(Dados1,1); DeltaN=floor(Ntotal/NInstantes);
Dados2=Dados1(1:DeltaN:Ntotal,:);
Nbat=size(Dados2,2)/vAnalise;
save KernelReferenciasSelDat Dados2 -ascii
MediasVet=mean(Dados2);
DesviosVet=std(Dados2);
for jj=1:NInstantes+1
    MediasMat(jj,:)=MediasVet;
    DesviosMat(jj,:)=DesviosVet;
end
Dados=(Dados2-MediasMat)./DesviosMat;
Delta1=eye(Nbat)/Nbat;
Delta2=eye(N)/N;
for ii=1:Nbat
    X=Dados(:,vAnalise*(ii-1)+1:vAnalise*ii);
    W=X*X';
    W=W/sqrt(trace(Delta2*W*Delta2*W));
    Wpond=eye(N)*W/N;
    for n=1:N
        w(n,(ii-1)*N+1:N*ii)=W(n,:);
    end
    for jj=1:Nbat
        Y=Dados(:,vAnalise*(jj-1)+1:vAnalise*jj);
        WW=Y*Y';
        WW=WW/sqrt(trace(Delta2*WW*Delta2*WW));
        WWpond=eye(N)*WW/N;
        S(ii,jj)=trace(Wpond*WWpond);
    end
end
[Svectors,Svalues] = eig(S*Delta1);
Svalues = max(Svalues);
[Svalue1, IndcS] = max(Svalues);
SvectorPrincipal = Svectors(:,IndcS);
Wtotal = zeros(N,N);
for bat = 1:Nbat
    Svector1 = SvectorPrincipal(bat,1);
    alfa(bat) = Svector1/(sqrt(Svalue1)*Nbat);
    Wtotal = Wtotal+alfa(bat)*w(:,(bat-1)*N+1:N*bat);
end
Wtotal = Wtotal/sqrt(trace(Delta2*Wtotal*Delta2*Wtotal));
[Wvectors,Wvalues] = eig(Delta2*Wtotal);
Wvalues = max(Wvalues);
[Wvalue1, IndcW] = max(Wvalues);
Wvector1 = Wvectors(:,IndcW);
Wvalues(IndcW) = 0;
[Wvalue2, IndcW] = max(Wvalues);
Wvector2 = Wvectors(:,IndcW);
for bat = 1:Nbat
    for n = 1:N
        Z(bat,2*n-1) = w(n,(bat-1)*N+1:N*bat)*Wvector1/(sqrt(Wvalue1)*N);
        Z(bat,2*n) = -w(n,(bat-1)*N+1:N*bat)*Wvector2/(sqrt(Wvalue2)*N);
    end
end
save Kerneldados_referencia.dat Z -ascii -double
save Kernelparametros Wvalue1 Wvalue2 Wvector1 Wvector2 w Wtotal
```



**Apêndice 1. Rotina escrita no *software Matlab* para os GCs IS e CO<sub>2</sub> e para o monitoramento *on-line*.**

Lê o arquivo com as coordenadas e faz o <i>B-SPLINE</i> para os GCs
<pre> load ParGerais alpha = 1.68; xall = load('Kerneldados_referencia.dat'); % arquivo com os dados referencia Tc = 1; s = size(xall); Nt = NInstantes+1; Na = s(1); figure(5);clf for t = 1:Nt;     x = xall(:,(2*t-1):(2*t));     c = cov(x);     ic = inv(c);     m = mean(x);     for j = 1:Na;         d(j) = (x(j,:)-m)*ic*(x(j,:)-m)';     end;     [mm,ind1] = sort(d);     ind = ind1(1:ceil(Tc*Na));     xfilt = x(ind,:);     mr = mean(xfilt);     k = convhull(xfilt(:,1),xfilt(:,2));     xhull = xfilt(k,:);     s = size(xhull);     clear yy     for i = 1:s(1);         tmp = xhull(i,:)-mr;         r = norm(tmp);         phi = angle(tmp(1)+sqrt(-1)*tmp(2));         yy(i,:) = mr+r*alpha*[cos(phi) sin(phi)];     end;     subplot(5,4,t);     hull{t} = yy;     t0 = 1:size(yy,1);     tt = 1.:size(yy,1);     xh = interp1(t0,yy(:,1),tt,'spline');     yh = interp1(t0,yy(:,2),tt,'spline');     plot(xh,yh);     hold on;     grid end </pre>
Lê o arquivo com os dados da batelada nova e os projeta <i>on-line</i> nos GCs
<pre> DadosDesc1 = load('KernelDescontrole.dat'); load Kernelparametros load ParGerais NbatNova = 1; DadosDesc2 = load('KernelDescontrole.dat') MediasDescVet = mean(DadosDesc2); DesviosDescVet = std(DadosDesc2); for jj = 1:NInstantes+1     MediasDescMat(jj,:) = MediasDescVet;     DesviosDescMat(jj,:) = DesviosDescVet; end DadosDesc = (DadosDesc2-MediasDescMat)./DesviosDescMat; N = NInstantes+1; WDesc = DadosDesc*DadosDesc'; Delta2 = eye(N)/N; WDesc = WDesc/sqrt(trace(Delta2*WDesc*Delta2*WDesc)); WDescOn = Wtotal; figure(5) hold on     for nova = 1:NbatNova         for n = 1:N             if n &gt; 1                 WDescOn(1:n,1:n) = WDesc(1:n,1:n);                 WDescOn(1:n-1,1:n-1) = Wtotal(1:n-1,1:n-1);             end             ZNovaOn(nova,2*n-1) = WDescOn(n,(nova-1)*N+1:N*nova)*Wvector1/sqrt(Wvalue1)/N;             ZNovaOn(nova,2*n) = -WDescOn(n,(nova-1)*N+1:N*nova)*Wvector2/sqrt(Wvalue2)/N;         end         subplot(5,4,n)         plot(ZNovaOn(nova,2*n-1),ZNovaOn(nova,2*n),'m*')     end end save Kernelpontos_descontrole.dat ZNovaOn -ascii </pre>