AIChE

A Bidirectional Between-Set Statistical Analysis Method and Its Applications

Chunhui Zhao and Furong Gao
Dept. of Chemical and Biomolecular Engineering, The Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong SAR

DOI 10.1002/aic.12339
Published online July 26, 2010 in Wiley Online Library (wileyonlinelibrary.com).

In this work, a bidirectional statistical modeling and analysis approach is developed to relate two data tables ($\mathbf{X}_1$ and $\mathbf{X}_2$) under the supervision of each other. Different from quality prediction where the interest was to interpret one set of variables by another set, the current task lies in modeling simultaneously both data spaces in bidirectional fashion ($\mathbf{X}_1 \leftrightarrow \mathbf{X}_2$) responding to different between-set relationships. It is performed in two steps. The first step aims at a bidirectional latent variable (Bi-LV) extraction and preparation, by which the between-set covarying relationship is preliminarily set up. In the second step, where a joint postprocessing is performed on the Bi-LV modeling result (here termed Bi-JPLV algorithm), different types of systematic variations are decomposed in each space. Correlated and unique variations are discriminated and evaluated in specific model parameters separately revealing between-set similarity and dissimilarity, respectively. The proposed method gives a good interpretation of the underlying information within each data space from a bidirectional viewpoint, revealing practical application potential. The feasibility and performance of the proposed method are illustrated with both numerical and real industrial cases. © 2010 American Institute of Chemical Engineers *AIChE J,* 57: 1233–1249, 2011

Keywords: between-set relationship, bidirectional latent variable extraction, joint postprocessing, canonical correlation analysis, correlated and unique variation information

Introduction

In the fields of chemometrics and statistics, the common problem is to analyze the measured or calculated variables for a set of observations collected in a data table. As data get more and more complex, several types of data sets may be collected in a specific problem, e.g., industrial processes. The data tables may come from different sources. In chemistry, it may contain different properties for a set of molecules or the state of some chemical process. The easiest way to handle different data spaces is to merge them into a single one and then use conventional multivariate statistical analysis method. This is simple and straightforward to summarize the general systematic information in two data spaces. However, their information is mixed together without discrimination and the problem is that the result may be hard to interpret owing to the missing view of each specific data space. In practical application, it is often of considerable interest to perform between-set statistical analysis. For example, in industrial processes, the measured monitoring variables can be classified into two groups by distinguishing some variables representing operating conditions such as a feed flow rate and a set-point, from those affected by them such as compound compositions. Thus, for the specific monitoring purpose, it transfers the uniform view on all variables to the dual insight into two different spaces. In this way, more information can be explored by extracting different types of variations, covering both between-set similarity and dissimilarity. Then, they can be monitored, providing a better analysis platform for fault detection and diagnosis.

Correspondence concerning this article should be addressed to F. Gao at kefgao@ust.hk.

Further, today process information representing both physical and chemical characteristics is available. In practice, with the possibility of interactions between chemical and physical factors, a change of absorbance at one wavelength period, revealing certain chemical variation, is usually accompanied by corresponding changes in physical state such as temperature and pressure. Besides the correlations, sometimes some fluctuations are only related with physical factors, whereas others are only affiliated with chemical factors. The challenge is to how to make better use of different types of data information under the supervision of each other. It has been demonstrated that a combination of spectroscopic and process data can provide a deeper knowledge of the studied process. The benefits should extend to other cases where there are more complex relationships between chemistry and physics, not only the between-set related but also their unique information. By relating two different types of variables and focusing on their different between-set relationships, it can provide a better characterization of the process and thus improved process understanding along with monitoring performance can be achieved.

Data-based multivariate calibration methods have been widely used to establish a quantitative regression relationship between two data tables with the purpose of either prediction or interpretation. The popular and conventional tools[1–8] include multiple linear regression, principal component regression, partial least squares (PLS), canonical correlation analysis (CCA), and so forth. Among them, the latent variable (LV) based methods, such as PLS and CCA, have a dominating role, which allow shrinking of the original data space into a lower dimensional feature subspace. However, conventional regression algorithms, taking example for PLS,[3] generally regress one data space on the other space. Thus, it only models the variations in $\mathbf{X}_1$-space and then uses these variations to predict $\mathbf{X}_2$ but does not model the variations in $\mathbf{X}_2$-space, so it overlooks the explanative contribution of $\mathbf{X}_2$ to $\mathbf{X}_1$. For comparison, it should be called as unidirectional regression algorithm here. Comparatively, unlike PLS, CCA[9–12] is not limited to a prediction technique but rather a technique for portraying the correlations between two sets of multivariate data. It can be implemented in a bidirectional fashion, in which, it models the variations in both spaces and can interpret them from both ways, $\mathbf{X}_1 \leftrightarrow \mathbf{X}_2$. Moreover, in PLS, the existence of $\mathbf{X}_2$-irrelevant variations in $\mathbf{X}_1$-space may weaken the $\mathbf{X}_1 \rightarrow \mathbf{X}_2$ correspondence, and thus lead to complex regression model structures. Unlike PLS, CCA inherently ignores the between-set irrelevant variation in each data space and directly maximizes the mutual correlation relationship, which thus improves the $\mathbf{X}_1 \leftrightarrow \mathbf{X}_2$ prediction correspondence. However, as the measurement variables are often high dimensional and collinear, directly applying CCA to the raw input space will lead to an ill-conditioned problem because it involves the calculation of $(\mathbf{X}_1^{\mathrm{T}}\mathbf{X}_1)^{-1}$ and $(\mathbf{X}_2^{\mathrm{T}}\mathbf{X}_2)^{-1}$ in the model estimation. That is why CCA is not as popular as PLS in practical applications.[13] Yu and MacGregor[13] developed a PLS-CCA algorithm, in which, as a postprocessing, CCA was implemented on PLS $\mathbf{X}_1$-LVs and $\mathbf{X}_2$ measurement variables to directly maximize their correlation. In this way, it avoided the rank-deficiency problem existing in single CCA algorithm and got rid of the pseudo quality-relevant variations so that a parsimonious regression model was obtained with the same prediction ability as the standard PLS model. However, resulting from the effect of first-step PLS modeling, it was also clarified in a unidirectional fashion, and thus limited to the specific regression analysis. It paid more attention to refining PLS $\mathbf{X}_1$-LVs under the influence of $\mathbf{X}_2$ but overlooked describing the $\mathbf{X}_2$-LVs ($\mathbf{U}$) in specific model parameters. Actually, in all regression modeling algorithms, the major interest was to explain one data set by the other set. Thus, their focus was on variation exploration of only one data space although both data spaces reveal interesting underlying information for process analysis and understanding. Besides, another common problem in the aforementioned calibration methods is that they only model the between-set correlated information but ignore their respective unique information. This part of information is also an important index in the evaluation of a complete between-set relationship. At the same time, their intensity should be quantified in specific model structure.

Various work have been developed on the basis of PLS algorithm to more comprehensively decompose the underlying variations although they were performed in a unidirectional fashion. Orthogonal signal correction (OSC) methods[14–16] are generally used as a preprocessing step to remove systematic variations in $\mathbf{X}_1$ that are definitely unrelated to $\mathbf{X}_2$. This is achieved by imposing mathematical orthogonality to a given $\mathbf{X}_2$-matrix or as close to orthogonal as possible. After the removal, from the corrected $\mathbf{X}_1$, the extracted PLS LVs may be more $\mathbf{X}_2$-related. Trygg and Wold[17] put forward an orthogonal PLS (O-PLS) algorithm by directly integrating OSC into the regular PLS algorithm, which could prior exclude the quality-orthogonal systematic variation from the regular PLS LVs without a fully estimated PLS model. The obvious advantages with O-PLS are more parsimonious regression model and easier interpretation, as well as consistent quality prediction performance. Zhou et al.[18] proposed a total PLS algorithm, which focused their attention on the decomposition of the whole $\mathbf{X}_1$-space under the supervision of $\mathbf{X}_2$. It was divided into four different subspaces, involving $\mathbf{X}_2$-informative, $\mathbf{X}_2$-orthogonal, other systematic information, and the residual. In this way, the underlying information in $\mathbf{X}_1$ was given a more comprehensive description. However, all the above methods were performed unidirectionally, that is, they only focused on the $\mathbf{X}_1$-space decomposition under the supervision of $\mathbf{X}_2$. For any two data tables, the relationship is interactional, which means the underlying variations in both spaces ($\mathbf{X}_1$ and $\mathbf{X}_2$) can be decomposed simultaneously under the mutual supervision. A comprehensive between-set relationship evaluation will then be obtained by making full use of these different variations. O2-PLS method was presented by Trygg and Trygg and Wold,[19,20] which could model and predict both $\mathbf{X}_1$ and $\mathbf{X}_2$ and had a dual OSC filter that could separate structured noises from $\mathbf{X}_1$ to $\mathbf{X}_2$ covariation in each data space. However, similar to O-PLS algorithm, the OSC filter was necessary only if structured noise was present in either $\mathbf{X}_1$-LVs or $\mathbf{X}_2$-LVs, that is, not all unique information in the original $\mathbf{X}_1$ and $\mathbf{X}_2$-spaces were figured out. This directly resulted from its specific objective of regression prediction, which, however, failed to gain an insight into the between-set dissimilarity.

In this study, a bidirectional modeling and analysis framework are designed for a comprehensive exploration of

between-set relationship ($\mathbf{X}_1 \leftrightarrow \mathbf{X}_2$). It tries to explore the underlying structures of each data space under the supervision of each other. In each data space, both correlated and unique variations can be structured in specific LVs and model parameters. At the same time, one would be able to quantify their respective intensity to fully evaluate the between-set relationship. To achieve this purpose, it constructs and ties together first-step bidirectional LV (Bi-LV) extraction and second-step joint postprocessing procedure, i.e., a two-step modeling strategy (here termed Bi-JPLV) is designed. Two problems of particular interest are addressed as below:

(1) Construct a Bi-LV extraction solution for preliminary LV preparation in both ways $\mathbf{X}_1 \leftrightarrow \mathbf{X}_2$, which can model $\mathbf{X}_1$ from $\mathbf{X}_1$, and $\mathbf{X}_2$ from $\mathbf{X}_2$, too. The LVs stand for the between-set covarying systematic variation and prepare an analysis platform for the following postprocessing. The corresponding calculation method is designed by solving a simple optimization algorithm.

(2) Perform a joint postprocessing on first-step LV extraction result. Common and specific information are separated from each other in each data space and made full use to quantitatively evaluate the between-set relationship. In this way, the underlying variation information in each data space can be more comprehensively explored and understood.

This article is organized as follows. First, the conventional PLS algorithm is analyzed, revealing its unidirectional limitation as well as the underlying reason and thus indicating the necessity of bidirectional extension. The proposed algorithm is then described in detail and its underlying principle is also explained. Its practical applications to process monitoring are also depicted. Simulation examples are presented to illustrate the performance of the proposed method. Discussion is conducted based on the results, highlighting the feasibility of the proposed method and also pointing out its future directions and possible improvements. Finally, conclusions are drawn in the last section.

Methodology

Theory analysis

As analyzed before, for between-set analysis, some problems in conventional regression methods should be paid special attention too. In the following, taking example for conventional PLS algorithm,[2,3] the existing problems will be further analyzed, revealing its limitation and the underlying reason, which can thus provide the basis for potential solution.

PLS is one of the most common regression methods for analyzing multivariate data where a quantitative prediction relationship between a descriptor matrix $\mathbf{X}_1$ and a response matrix $\mathbf{X}_2$ is sought. The underlying assumption is that the system or process under consideration is driven by a small number of LVs so that a LV model is first derived from one data block sharing a covarying relationship with the other block. However, as analyzed before, PLS is limited to the unidirectional quality regression and prediction. For a clear understanding, a typical PLS calculation procedure is briefly presented as below:

(1) Mean-center and scale the $\mathbf{X}_1$ and $\mathbf{X}_2$ matrices;

(2) Compute the one-component statistics $\mathbf{w}_i$, $\mathbf{t}_i$, and $\mathbf{u}_i$ using either the NIPALS algorithm[3] or a kernel algorithm[21];

(3) Deflation:

$$\mathbf{p}_i = \mathbf{X}_{1,i}^{\mathrm{T}}\mathbf{t}_i \Big/ \mathbf{t}_i^{\mathrm{T}}\mathbf{t}_i, \quad \mathbf{X}_{1,i+1} = \mathbf{X}_{1,i} - \mathbf{t}_i\mathbf{p}_i^{\mathrm{T}}$$

$$\mathbf{q}_i = \mathbf{X}_{2,i}^{\mathrm{T}}\mathbf{t}_i \Big/ \mathbf{t}_i^{\mathrm{T}}\mathbf{t}_i, \quad \mathbf{X}_{2,i+1} = \mathbf{X}_{2,i} - \mathbf{t}_i\mathbf{q}_i^{\mathrm{T}}$$

(4) Return to step (2) for the next LV.

Although many successful applications[22–25] about PLS have been reported, for the specific purpose of between-set analysis, there are following major problems which should be paid special attention to:

First, as analyzed in Introduction section, to give a complete description of the between-set relationship, both data spaces should be considered and modeled at the same time. Common and unique information should be distinguished to evaluate between-set similarity and dissimilarity, respectively. However, PLS, as one specific regress modeling strategy, was implemented in a unidirectional fashion, where, it only stressed the descriptor ($\mathbf{X}_1$) variations and used them to explain both spaces. Moreover, the unique-to-set information was fully disregarded. It was also conflicted in such a way that real correlation could not be obtained by pursuing the maximization of $\mathbf{X}_1$–$\mathbf{X}_2$ covariance in the extraction of PLS LVs. Both $\mathbf{X}_2$-related and irrelevant variations were mixed up in PLS LVs, thus leading to interpretation problem.

Second, in contrast to CCA, which incorporated orthogonality constraints for determining subsequent score vectors, PLS used a deflation procedure. It was implemented in a sequential fashion where only one $\mathbf{X}_1$-LV ($\mathbf{t}_i$) was computed at a time and then both $\mathbf{X}_1$ and $\mathbf{X}_2$ matrices were deflated to remove the information enclosed by $\mathbf{t}_i$ before the calculation of the next LV ($\mathbf{t}_{i+1}$); $\mathbf{X}_{1,i+1} = \mathbf{X}_{1,i} - \mathbf{t}_i\mathbf{p}_i^{\mathrm{T}}$, $\mathbf{X}_{2,i+1} = \mathbf{X}_{2,i} - \mathbf{t}_i\mathbf{q}_i^{\mathrm{T}}$. Deflation inherently took care of the orthogonality of the $\mathbf{X}_1$-LVs so that information could not repetitively appear in subsequent LVs. However, such a deflation method will result in "information mixing" problem in PLS. Clearly, by calculating the respective contributions of $\mathbf{X}_1$-LV ($\mathbf{t}_i$) toward $\mathbf{X}_1$ and $\mathbf{X}_2$ matrices, each data space was actually projected onto the space spanned by the $\mathbf{X}_1$-LV. When $\mathbf{X}_1$ was deflated, the updated data still stayed within the original column space of $\mathbf{X}_1$. However, the space spanned by the deflated $\mathbf{X}_2$-columns did not necessarily lie in the original $\mathbf{X}_2$-space. This implied that PLS only modeled $\mathbf{X}_2$ within its original space in the first $\mathbf{X}_2$-LV ($\mathbf{u}_1$). The other $\mathbf{X}_2$-LVs after the first dimension were not guaranteed because information outside the $\mathbf{X}_2$-space was introduced into its corrected form, which thus led to false model interpretation of the $\mathbf{X}_2$-variations.

Then a question may naturally arise: Why are only $\mathbf{X}_1$-LVs used for deflation in both spaces? Actually, each pair of LVs ($\mathbf{t}_i - \mathbf{u}_i$) represents a single-input, single-output (SISO) model between two variable sets as pointed out by Sharma et al.[26] It means that some variations in one $\mathbf{X}_2$-LV ($\mathbf{u}_i$) cannot be predicted by the corresponding $\mathbf{X}_1$-LV ($\mathbf{t}_i$). If $\mathbf{u}_i$ is removed from $\mathbf{X}_2$, it cannot be attained from the future $\mathbf{X}_2$-LVs either and thus shows no chance to be explained by other $\mathbf{X}_1$-LVs, leading to inferior prediction of $\mathbf{X}_2$. Here it is called "information missing" problem. Therefore, only the contribution predicted by the SISO models can be used to deflate both $\mathbf{X}_1$ and $\mathbf{X}_2$. That is, only the information

encapsulated in the $\mathbf{X}_1$-LVs ($\mathbf{T}$) are used to reconstruct the used $\mathbf{X}_1$ variation and the predicted $\mathbf{X}_2$ variation.

For the current specific study of between-set relationship, the aforementioned problems should be considered. In the following subsection, a two-step statistical modeling strategy is designed. In the first step, a Bi-LV extraction algorithm is formulated, which can preliminarily prepare a series of LV correspondence pairs from both ways ($\mathbf{X}_1 \leftrightarrow \mathbf{X}_2$) for the following postprocessing. Discussions are conducted to reveal its difference compared with conventional PLS regression method. In the second step, a joint postprocessing procedure is then developed based on Bi-LV extraction result, in which different types of variations within each data space can be distinguished and constructed in specific model parameters. They will be made of full use for a more comprehensive and quantitative comprehension of between-set relationship.

First-step Bi-LV extraction

Here, focusing on two data sets $\mathbf{X}_1$ ($N \times J_1$) and $\mathbf{X}_2$ ($N \times J_2$), a Bi-LV extraction strategy is designed, by which, the between-set correspondence is preliminarily prepared. We try to define different linear combinations of either variable set, i.e., the sub-latent variables (subLVs) in each data space, $\mathbf{X}_1\mathbf{a}_1$ and $\mathbf{X}_2\mathbf{a}_2$, and relate them via one super latent variable (supLV, $\mathbf{t}_g$). The cost function can be formulated with certain constraints as below:

$$\max R^2 = \max\left(\left(\mathbf{t}_g^T\mathbf{X}_1\mathbf{a}_1\right)^2 + \left(\mathbf{t}_g^T\mathbf{X}_2\mathbf{a}_2\right)^2\right) \quad \text{s.t.} \begin{cases} \mathbf{t}_g^T\mathbf{t}_g = 1 \\ \mathbf{a}_1^T\mathbf{a}_1 = 1 \\ \mathbf{a}_2^T\mathbf{a}_2 = 1 \end{cases} \tag{1}$$

For each data table, the means of each column are subtracted to approximately eliminate the main nonlinearity. Each variable is scaled to unit variance to handle different measurement units, thus giving each equal weight. The combination coefficient vector $\mathbf{a}_i$ ($i = 1,2$) is set to unit length. Therefore, the subLV $\mathbf{X}_i\mathbf{a}_i$ carries the associated variance information and $(\mathbf{t}_g^T\mathbf{X}_i\mathbf{a}_i)^2$ actually models the covariance information between each subLV ($\mathbf{X}_i\mathbf{a}_i$) and supLV ($\mathbf{t}_g$) rather than their pure correlation. SubLVs in two data spaces are both related to $\mathbf{t}_g$ instead of directly maximizing their covariance. That is, they are indirectly related by means of a third-party $\mathbf{t}_g$, in which $\mathbf{t}_g$ actually plays the role of connection bridge. By such an optimization objective, it is expected that the extracted subLVs should carry as many variations as possible within each data space, whereas being correlated as closely as possible through the connection of $\mathbf{t}_g$.

Using a Lagrange operator, the initial objective function can be expressed as an unconstrained extremum problem:

$$\mathrm{F}(\mathbf{t}_g, \mathbf{a}_i, \lambda) = \left(\mathbf{t}_g^T\mathbf{X}_1\mathbf{a}_1\right)^2 + \left(\mathbf{t}_g^T\mathbf{X}_2\mathbf{a}_2\right)^2 - \lambda_g\left(\mathbf{t}_g^T\mathbf{t}_g - 1\right) - \lambda_1\left(\mathbf{a}_1^T\mathbf{a}_1 - 1\right) - \lambda_2\left(\mathbf{a}_2^T\mathbf{a}_2 - 1\right) \tag{2}$$

where λ_g and $\lambda_i (i = 1,2)$ are constant scalars.

Calculating the derivatives of $F(\mathbf{t}_g, \mathbf{a}_i, \lambda_g, \lambda)$ with respect to $\mathbf{t}_g$, $\mathbf{a}_i$, λ_g, and λ_i, and setting all equal to zero, the following mathematical expressions can be obtained:

$$\frac{\partial \mathrm{F}}{\partial \mathbf{t}_g} = 2\left(\mathbf{t}_g^T\mathbf{X}_1\mathbf{a}_1\right)\mathbf{X}_1\mathbf{a}_1 + 2\left(\mathbf{t}_g^T\mathbf{X}_2\mathbf{a}_2\right)\mathbf{X}_2\mathbf{a}_2 - 2\lambda_g\mathbf{t}_g = 0 \tag{3}$$

$$\frac{\partial F}{\partial \mathbf{a}_i} = 2\left(\mathbf{t}_g^T\mathbf{X}_i\mathbf{a}_i\right)\mathbf{X}_i^T\mathbf{t}_g - 2\lambda_i\mathbf{a}_i = 0 \tag{4}$$

$$\mathbf{t}_g^T\mathbf{t}_g - 1 = 0 \tag{5}$$

$$\mathbf{a}_i^T\mathbf{a}_i - 1 = 0 \tag{6}$$

Left multiply Eq. 3 with $\mathbf{t}_g^T$ and Eq. 4 with $\mathbf{a}_i^T$, and the following relations can be derived in combination with Eqs. 5 and 6:

$$\begin{cases} \left(\mathbf{t}_g^T\mathbf{X}_1\mathbf{a}_1\right)^2 + \left(\mathbf{t}_g^T\mathbf{X}_2\mathbf{a}_2\right)^2 = \lambda_g \\ \left(\mathbf{t}_g^T\mathbf{X}_i\mathbf{a}_i\right)^2 = \lambda_i \end{cases} \tag{7}$$

Therefore, it tells us that λ_g is actually the desired optimization objective and λ_i denotes the suboptimal objective parameter, meanwhile satisfying this equality $\lambda_g = \lambda_1 + \lambda_2$. It should be noted that these suboptimal objective values λ_i are not necessarily the same, that is, the covariance information is not necessarily equivalently distributed between the two different sets.

Correspondingly, Eqs. 3 and 4 can be modified

$$\sqrt{\lambda_1}\mathbf{X}_1\mathbf{a}_1 + \sqrt{\lambda_2}\mathbf{X}_2\mathbf{a}_2 = \lambda_g\mathbf{t}_g \tag{8}$$

$$\frac{1}{\sqrt{\lambda_i}}\mathbf{X}_i^T\mathbf{t}_g = \mathbf{a}_i \tag{9}$$

From Eq. 8, the supLV is actually the weighted average of the associated subLVs each with $\sqrt{\lambda_i}/\lambda_g$ as attached weight: $\mathbf{t}_g = \frac{\sqrt{\lambda_1}}{\lambda_g}\mathbf{t}_1 + \frac{\sqrt{\lambda_2}}{\lambda_g}\mathbf{t}_2$.

Input Eq. 9 into Eq. 8:

$$\begin{aligned} &\left(\mathbf{X}_1\mathbf{X}_1^T + \mathbf{X}_2\mathbf{X}_2^T\right)\mathbf{t}_g = \lambda_g\mathbf{t}_g \\ &\text{i.e., } \mathbf{Q}\mathbf{t}_g = \lambda_g\mathbf{t}_g \end{aligned} \tag{10}$$

This is a standard algebra problem. At the request of the maximal objective function value, i.e., the largest λ_g, analytically, the solution leads to the eigenvalue decomposition of the sum of Gram matrices, $\mathbf{X}_1\mathbf{X}_1^T$ and $\mathbf{X}_2\mathbf{X}_2^T$.

Inputting Eq. 9 into Eq. 7, the suboptimal objective parameter λ_i can be calculated:

$$\mathbf{t}_g^T\mathbf{X}_i\mathbf{X}_i^T\mathbf{t}_g = \lambda_i \tag{11}$$

Then combining Eqs. 9 and 11, the subLVs can be calculated, revealing an obvious conversion relationship associated with the supLV:

$$\begin{aligned} \mathbf{t}_i &= \mathbf{X}_i\mathbf{a}_i \\ &= \sqrt{\frac{1}{\lambda_i}}\mathbf{X}_i\mathbf{X}_i^T\mathbf{t}_g \\ &= \sqrt{\frac{1}{\mathbf{t}_g^T\mathbf{X}_i\mathbf{X}_i^T\mathbf{t}_g}}\mathbf{X}_i\mathbf{X}_i^T\mathbf{t}_g \end{aligned} \tag{12}$$

In order, R supLVs ($\mathbf{T}_g$) can be derived by Eq. 10 in accord with the descending λ_g. Correspondingly, the same

number of resulting subLVs can be directly calculated using Eq. 12 in each dataset, which can be formulated as:

$$\mathbf{T}_i = \mathbf{X}_i \mathbf{X}_i^{\mathrm{T}} \mathbf{T}_{\mathrm{g}} \Lambda_i \tag{13}$$

where Λ_i is a diagonal matrix with a series of $\sqrt{\frac{1}{\lambda_i}}$ as its elements.

As shown in Eq. 10, the supLVs ($\mathbf{T}_{\mathrm{g}}$) can describe the general systematic variation in the concatenation data space $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2]$ as $\mathbf{T}_{\mathrm{g}}$ actually converges to the principal components (PCs) in $\mathbf{X}$-space. Then under its supervision, the corresponding subLVs ($\overline{\mathbf{T}}_i$) are extracted by regressing each data table on each supLV ($\mathbf{t}_{\mathrm{g}}$), respectively, in turn as shown in Eq. 12. In this way, the subLVs are defined to capture the local systematic variations in each specific data space, which are related with the general systematic information represented by supLVs $\mathbf{T}_{\mathrm{g}}$.

However, the subLVs calculated as shown in Eq. 12 may have collinearity problem in each data space because each data table is repetitively used. Moreover, they are not sorted according to the covariance of between-set subLV pairs. For the successful implementation of the second-step postprocessing, it is necessary to make some modification on the original subLVs. On one hand, it is expected that the subLVs can be automatically ordered in accord with the descending between-set covariances. On the other hand, the subLVs are supposed to be orthogonal with each other. Here, to achieve the above purposes, as shown in Appendix, a modification strategy is developed, by which, modified subLVs (MsubLVs) in two spaces ($\mathbf{T}_1$ and $\mathbf{T}_2$) can be readily figured out. When compared with the original subLVs, the orthogonality constraint is imposed by deflating variation carried by previous subLV in each step and they are generally automatically ordered following descending between-set covariances. The modified Bi-LV extraction results are then formulated as below:

$$\begin{aligned} \mathbf{T}_1 &= \mathbf{X}_1 \mathbf{R}_1 \\ \mathbf{P}_1^{\mathrm{T}} &= \left(\mathbf{T}_1^{\mathrm{T}} \mathbf{T}_1\right)^{-1} \mathbf{T}_1^{\mathrm{T}} \mathbf{X}_1 \\ \mathbf{T}_2 &= \mathbf{X}_2 \mathbf{R}_2 \\ \mathbf{P}_2^{\mathrm{T}} &= \left(\mathbf{T}_2^{\mathrm{T}} \mathbf{T}_2\right)^{-1} \mathbf{T}_2^{\mathrm{T}} \mathbf{X}_2 \end{aligned} \tag{14}$$

where $\mathbf{R}_1$ and $\mathbf{R}_2$ are the Bi-LV weights to directly calculate the LVs from the original data spaces. $\mathbf{P}_1$ and $\mathbf{P}_2$ are the corresponding loadings, respectively. For simplicity, in the following presentation, subLVs uniformly mean modified orthogonal subLVs unless otherwise noted.

Then each data space ($\mathbf{X}_i$) can be preliminarily separated into two parts, where the associated systematic subspace is enclosed by $\mathbf{T}_i$:

$$\begin{aligned} \mathbf{X}_1 &= \tilde{\mathbf{X}}_1 + \mathbf{E}_1 = \mathbf{T}_1 \mathbf{P}_1^{\mathrm{T}} + \mathbf{E}_1 = \mathbf{G}_{\mathbf{T}_1} \mathbf{X}_1 + \mathbf{H}_{\mathbf{T}_1} \mathbf{X}_1 \\ \mathbf{X}_2 &= \tilde{\mathbf{X}}_2 + \mathbf{E}_2 = \mathbf{T}_2 \mathbf{P}_2^{\mathrm{T}} + \mathbf{E}_2 = \mathbf{G}_{\mathbf{T}_2} \mathbf{X}_2 + \mathbf{H}_{\mathbf{T}_2} \mathbf{X}_2 \end{aligned} \tag{15}$$

where $\mathbf{G}_{\mathbf{T}_1}$ and $\mathbf{G}_{\mathbf{T}_2}$ are the orthogonal projectors onto the column spaces of $\mathbf{T}_1$ and $\mathbf{T}_2$, respectively, which guarantee that both the systematic variations $\tilde{\mathbf{X}}_1$ and $\tilde{\mathbf{X}}_1$ lie well in their own spaces. $\mathbf{H}_{\mathbf{T}_1}$ and $\mathbf{H}_{\mathbf{T}_2}$ are the orthogonal complement of $\mathbf{G}_{\mathbf{T}_1}$ and $\mathbf{G}_{\mathbf{T}_2}$, respectively. $\mathbf{E}_1$ and $\mathbf{E}_2$ are the Bi-LV residuals.

The complete Bi-LV modeling procedure is summarized in Appendix.

Property analysis and discussion

In detail, here, the Bi-LV extraction algorithm will be further analyzed and interpreted with reference to the optimal objective and the resulting LV structure.

(1) From the calculation shown in Eq. 10, the extracted supLVs actually converge to the principal scores in the concatenation data space $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2]$ and thus reveal the general systematic variability when taking into account the correlation between the two data spaces $\mathbf{X}_1$ and $\mathbf{X}_2$. Some important variations isolated in $\mathbf{X}_1$ might no longer be of importance if they do not exist in $\mathbf{X}_2$ and vice versa (although the case is not absolute). Moreover, the orthogonality of $\mathbf{T}_{\mathrm{g}}$ is readily achieved with no deflation procedure, which can prevent the same general systematic variations from being repeatedly explored.

On the other hand, the extracted supLVs serve as the connecting bridge between two data spaces, $\mathbf{X}_1 \rightarrow \mathbf{T}_{\mathrm{g}}$ and $\mathbf{X}_2 \rightarrow \mathbf{T}_{\mathrm{g}}$. Under its conduct, the MsubLVs ($\mathbf{T}_1$ and $\mathbf{T}_2$) are extracted from both data tables as shown in Appendix. Seen from Eqs. A1 and A2, deflation is carried out by subtracting the MsubLVs in their respective data spaces, which guarantees no information mixing. Moreover, supLVs have fixed the amount of general systematic variations covered in the joint data space $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2]$. Therefore, under their supervision, the deflation will not result in the "information missing" problem as analyzed before in conventional PLS. In this way, the initial correspondence of Bi-LV pairs are set up in both ways through the bridge of $\mathbf{T}_{\mathrm{g}}$, $\mathbf{X}_1 \rightarrow \mathbf{T}_{\mathrm{g}} \leftarrow \mathbf{X}_2$.

(2) The Bi-LV extraction is graphically illustrated in Figure 1a. Generally, Bi-LV is to make preliminary preparations for between-set similarity in either data space. This leads to two modeling levels: the upper level, where the two data tables are put together and viewed simultaneously so that the supLVs ($\mathbf{T}_{\mathrm{g}}$) summarize the joint systematic variability; and the lower level, where the MsubLVs ($\mathbf{T}_i$) are extracted by associating each data space with $\mathbf{T}_{\mathrm{g}}$ so that they depict the local information in each specific data table in the presence of the other table. From this point, the Bi-LV method basically shares certain common characteristics with conventional PLS as the other data set plays a central role in determining the LVs in one data space. SupLVs ($\mathbf{T}_{\mathrm{g}}$)work as a third-party data space, which makes the interpretation of MsubLVs easier. Although the MsubLVs are not guaranteed to be between-set closely related, however, they provide a proper bidirectional statistical analysis platform for the second-step postprocessing. The underlying variations will be further separated based on different between-set relationships, involving both correlated and orthogonal variations.

(3) Essentially, both Bi-LV and typical PLS algorithms seek the covarying correspondence between two data tables, which, however, are performed in bidirectional and unidirectional fashions, respectively. Moreover, they are achieved by different objective functions and calculation procedures. As shown in Eq. 1, instead of the "direct" covariance index ($\left((\mathbf{X}_1\mathbf{a}_1)^{\mathrm{T}}(\mathbf{X}_2\mathbf{a}_2)\right)^2$) used by 2PLS, Bi-LV$_2$uses the "indirect" covariance index ($\left(\mathbf{t}_{\mathrm{g}}^{\mathrm{T}}\mathbf{X}_1\mathbf{a}_1\right) + \left(\mathbf{t}_{\mathrm{g}}^{\mathrm{T}}\mathbf{X}_2\mathbf{a}_2\right)$) by means of the bridge of a third party ($\mathbf{t}_{\mathrm{g}}$). When comparing the two different optimization indices, "indirect" index allows one term to be zero, i.e., bias optimization issue as one common dilemma in multioptimization problem.

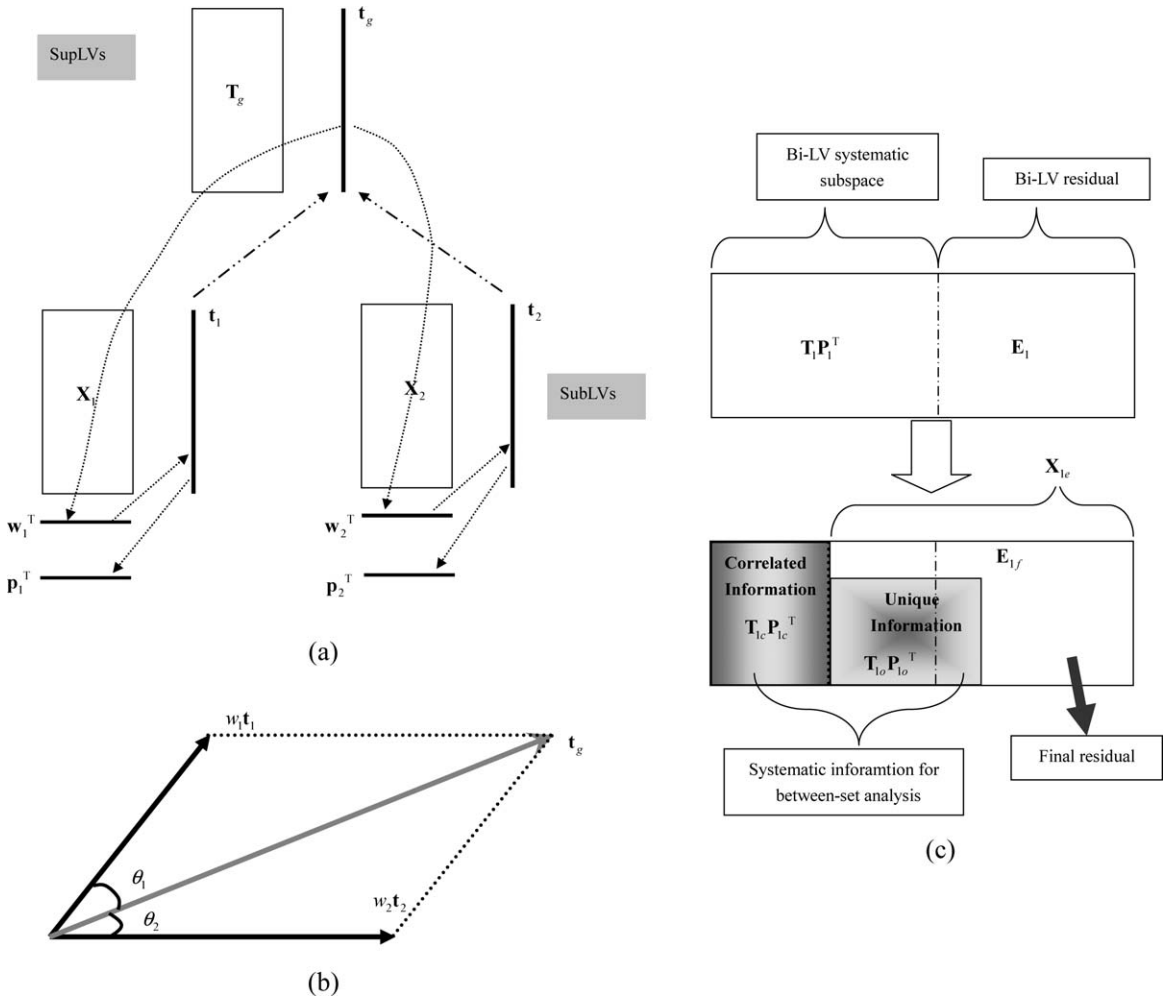

Figure 1. (a) Schematic overview of the Bi-LV model structures, (b) graphical representation of subLVs and supLV, and (c) summary of the subspace separation.

In detail, as shown in Eq. 8, the calculated supLV is the weighted average of the associated original subLVs each with λ_i/λ_g as attached weights (w_i): $\mathbf{t}_g = \frac{\sqrt{\lambda_1}}{\lambda_g}\mathbf{t}_1 + \frac{\sqrt{\lambda_2}}{\lambda_g}\mathbf{t}_2 = \boldsymbol{w}_1\mathbf{t}_1 + \boldsymbol{w}_2\mathbf{t}_2$. Their relationship can be geometrically illustrated in Figure 1b. The weighted subLVs are the sides of a parallelogram, the supLV locates along the diagonal line. Resulting from $\mathbf{t}_g^T\,\mathbf{t}_g = 1$, the square inner product $(\mathbf{t}_g^T\,\mathbf{t}_i)^2 = |\mathbf{t}_g|^2\,|\mathbf{t}_i|^2\cos^2\theta_i = |\mathbf{t}_i|^2\cos^2\theta_i$ is determined by the length of each subLV ($\mathbf{t}_i$) and the $\mathbf{t}_g - \mathbf{t}_i$ angle. From the geometric figure, generally, if the two subLVs are more related with each other, i.e., a smaller angle ($\theta_1 + \theta_2$) between them, they may both approach the supLV, resulting in a smaller angel (θ_i) and larger cos θ_i value. On the contrary, if the two subLVs are less related, the angle between them ($\theta_1 + \theta_2$) is larger, which means at least one subLV ($\mathbf{t}_i$) will be far from the supLV (i.e., larger θ_i). As shown by the objective function $|\mathbf{t}_1|^2\cos^2\theta_1 + |\mathbf{t}_2|^2\cos^2\theta_2$, it is possible that there are extreme cases that the objective value is only dominated by one term, whereas the other term may approach zero (e.g., $\theta_1 = 0$). It means that the supLV may only have a close relationship with one subLV, whereas has less relationship with the other one. For conventional PLS, the LV pairs are directly related with each other, which can be geometrically expressed as $|\mathbf{t}_1||\mathbf{t}_2|\cos(\theta_1 + \theta_2)$ based on the same notation as shown in Figure 1b. It guarantees that subLVs are related with each other to some extent at least, i.e., the angle ($\theta_1 + \theta_2$) is not too large. However, as analyzed before, in conventional PLS, only the first pair of LVs is guaranteed to locate in their respective data space, whereas the bidirectional correspondence of sequential LV pairs are deteriorated resulting from the information mixing problem.

(4) As analyzed before, the supLVs ($\mathbf{T}_g$) capture the underlying systematic variation information in the concatenation data space jointed by $\mathbf{X}_1$ and $\mathbf{X}_2$. Therefore, the retained number of supLVs (R) can be determined by the cumulative explained variance rate,[27] defined by $R = \sum_{m=1}^{R}\lambda_{g,m}/.\sum\lambda_{g,m} \geq \beta$ (where λ_g is obtained by Eq. 10 and β is the threshold value). In this way, the major systematic information in the concatenation space can be generally retrieved. Moreover, their respective contributions to each specific data space can also be counted as $X_i = \mathbf{T}_g\left(\mathbf{T}_g^T\mathbf{T}_g\right)^{-1}\mathbf{T}_g^T\mathbf{X}_i$. Then by regressing each data space on $\mathbf{T}_g$, orthogonal MsubLVs ($\mathbf{T}_i$) are extracted. The amount

of variability taking part in $\mathbf{T}_g$ interpretation in each specific data space can be counted as $\tilde{\mathbf{X}}_i = \mathbf{T}_i(\mathbf{T}_i^T\mathbf{T}_i)^{-1}\mathbf{T}_i^T\mathbf{X}_i$. Correspondingly, the number of MsubLVs ($\mathbf{T}_i$) in each specific data space can be determined by checking whether the amount of variations ($\underset{\sim i}{X}$) contributed by $\mathbf{T}_g$ has been reconstructed sufficiently by $\mathbf{T}_i$. An evaluation index is thus defined by quantitatively comparing them, $\text{Ratio}_i = \frac{\sum \tilde{\mathbf{x}}_i^2}{\sum \underset{\sim i}{X}^2}$. A value not less than 1 means a sufficient reconstruction. Then, the number of $\mathbf{T}_i$ can be determined in each data space ($\mathbf{R}_1$ and $\mathbf{R}_2$), respectively.

Second-step joint postprocessing

Here, in the second modeling step, a joint postprocessing procedure is designed and performed on the first-step Bi-LV extraction result, which can further decompose the underlying information within each data space under the supervision of each other. According to their mutual relationship, two major systematic subspaces can be separated. The first part is exploited by CCA on the Bi-LV systematic subspace, revealing their closely related information. The second part is the extracted orthogonal components by OSC from CCA residuals, which will describe the unique systematic information in each data space.

The complete modeling procedure is summarized as below:

(1) Run Bi-LV extraction (as shown in Appendix) on $\{\mathbf{X}_1, \mathbf{X}_2\}$:

$$\begin{aligned}
&\mathbf{T}_1 = \mathbf{X}_1\mathbf{R}_1\\
&\mathbf{T}_2 = \mathbf{X}_2\mathbf{R}_2\\
&\mathbf{X}_1 = \tilde{\mathbf{X}}_1 + \mathbf{E}_1 = \mathbf{T}_1\mathbf{P}_1^T + \mathbf{E}_1\\
&\mathbf{X}_2 = \tilde{\mathbf{X}}_2 + \mathbf{E}_2 = \mathbf{T}_2\mathbf{P}_2^T + \mathbf{E}_2
\end{aligned} \tag{16}$$

where $\mathbf{R}_1$ $\mathbf{X}_1$- LVs ($\mathbf{T}_1$) and $\mathbf{R}_2$ $\mathbf{X}_2$- LVs ($\mathbf{T}_2$) that are extracted from their respective spaces preliminarily prepare the between-set related systematic information.

(2) Focusing on $\{T_1, \mathbf{T}_2\}$, CCA is performed to extract R_c pairs of close correlated canonical components ($\mathbf{T}_{1c}$ and $\mathbf{T}_{2c}$). Then the associated subspace decomposition in $\mathbf{X}_1$ and $\mathbf{X}_2$ can be updated, respectively:

$$\begin{aligned}
&\mathbf{T}_{1c} = \mathbf{T}_1\mathbf{W}_{1c}\\
&\mathbf{P}_{1c}^T = (\mathbf{T}_{1c}^T\mathbf{T}_{1c})^{-1}\mathbf{T}_{1c}^T\mathbf{X}_1 = \mathbf{T}_{1c}^T\mathbf{X}_1\\
&\mathbf{X}_1 = \mathbf{X}_{1c} + \mathbf{X}_{1e} = \mathbf{T}_{1c}\mathbf{P}_{1c}^T + \mathbf{X}_{1e}\\
&\mathbf{T}_{2c} = \mathbf{T}_2\mathbf{W}_{2c}\\
&\mathbf{P}_{2c}^T = (\mathbf{T}_{2c}^T\mathbf{T}_{2c})^{-1}\mathbf{T}_{2c}^T\mathbf{X}_2 = \mathbf{T}_{2c}^T\mathbf{X}_2\\
&\mathbf{X}_2 = \mathbf{X}_{2c} + \mathbf{X}_{2e} = \mathbf{T}_{2c}\mathbf{P}_{2c}^T + \mathbf{X}_{2e}
\end{aligned} \tag{17}$$

where canonical components ($\mathbf{T}_{1c}$ and $\mathbf{T}_{2c}$) are ordered descendingly according to their correlation values. The number R_c can be up to min (R_1, R_2). Considering that the closely related information may be centralized in the first several pairs of canonical components, the last few pairs can be excluded if they have low correlations and do not carry much variation information. Modeled by $\mathbf{T}_{1c}$ and $\mathbf{T}_{2c}$ the between-set close correlated systematic information are retrieved in their own space in a bidirectional manner, respectively.

Moreover, based on the extraction of $\mathbf{T}_{1c}$ and $\mathbf{T}_{2c}$, their predictions in their opposite data spaces can also be modeled, respectively:

$$\begin{aligned}
&\mathbf{Q}_2^T = (\mathbf{T}_{1c}^T\mathbf{T}_{1c})^{-1}\mathbf{T}_{1c}^T\mathbf{X}_2 = \mathbf{T}_{1c}^T\mathbf{X}_2\\
&\hat{\mathbf{X}}_2 = \mathbf{T}_{1c}\mathbf{Q}_2^T = \mathbf{T}_{1c}\mathbf{T}_{1c}^T\mathbf{X}_2\\
&\mathbf{Q}_1^T = (\mathbf{T}_{2c}^T\mathbf{T}_{2c})^{-1}\mathbf{T}_{2c}^T\mathbf{X}_1 = \mathbf{T}_{2c}^T\mathbf{X}_1\\
&\hat{\mathbf{X}}_1 = \mathbf{T}_{2c}\mathbf{Q}_1^T = \mathbf{T}_{2c}\mathbf{T}_{2c}^T\mathbf{X}_1
\end{aligned} \tag{18}$$

Here it should be noted that $\hat{\mathbf{X}}_1$ and $\hat{\mathbf{X}}_2$ in Eq. 18 are different from $\mathbf{X}_{1c}$ and $\mathbf{X}_{2c}$ in Eq. 17, respectively. For example, in $\mathbf{X}_1$-space, $\mathbf{X}_{1c}$ is modeled by $\mathbf{T}_{1c}$, representing the information that can be used to predict $\mathbf{X}_2$; differently, $\hat{\mathbf{X}}_1$ is modeled by $\mathbf{T}_{2c}$, representing the information predicted by $\mathbf{X}_2$. They have different statistical meanings and may be of different values too. This will be further quantitatively illustrated in Simulation section.

(3) After the extraction of closely related systematic variations, the major between-set irrelevant information can be extracted using Fearn's OSC algorithm[15] ($\mathbf{X}_{1e} \xrightarrow{\text{osc}} \mathbf{T}_{2c}$, $\mathbf{X}_{2e} \xrightarrow{\text{osc}} \mathbf{T}_{1c}$):

$$\begin{aligned}
&\mathbf{T}_{1o} = \mathbf{X}_{1e}\mathbf{W}_{1o}\\
&\mathbf{P}_{1o}^T = (\mathbf{T}_{1o}^T\mathbf{T}_{1o})^{-1}\mathbf{T}_{1o}^T\mathbf{X}_{1e}\\
&\mathbf{X}_{1e} = \mathbf{X}_{1o} + \mathbf{E}_{1f} = \mathbf{T}_{1o}\mathbf{P}_{1o}^T + \mathbf{E}_{1f}\\
&\mathbf{T}_{2o} = \mathbf{X}_{2e}\mathbf{W}_{2o}\\
&\mathbf{P}_{2o}^T = (\mathbf{T}_{2o}^T\mathbf{T}_{2o})^{-1}\mathbf{T}_{2o}\mathrm{T}\mathbf{X}_{2e}\\
&\mathbf{X}_{2e} = \mathbf{X}_{2o} + \mathbf{E}_{2f} = \mathbf{T}_{2o}\mathbf{P}_{2o}^T + \mathbf{E}_{2f}
\end{aligned} \tag{19}$$

where the orthogonal components $\mathbf{T}_{1o}$ and $\mathbf{T}_{2o}$ reveal the systematic information, which are orthogonal to $\mathbf{T}_{2c}$ and $\mathbf{T}_{1c}$, respectively, i.e., the systematic variations that do not participate in the mutual prediction. The retained number ($\mathbf{R}_{1o}$ and $\mathbf{R}_{2o}$) can be determined by their respective variances. $\mathbf{E}_{1f}$ and $\mathbf{E}_{2f}$ are the final residuals in each data space, which cover no persistent systematic information useful for interpreting between-set relationship.

Summarily, three orthogonal subspaces are decomposed according to different between-set relationships:

$$\begin{aligned}
&\mathbf{X}_1 = \mathbf{X}_{1c} + \mathbf{X}_{1o} + \mathbf{E}_{1f} = \mathbf{T}_{1c}\mathbf{P}_{1c}^T + \mathbf{T}_{1o}\mathbf{P}_{1o}^T + \mathbf{E}_{1f}\\
&\mathbf{X}_2 = \mathbf{X}_{2c} + \mathbf{X}_{2o} + \mathbf{E}_{2f} = \mathbf{T}_{2c}\mathbf{P}_{2c}^T + \mathbf{T}_{2o}\mathbf{P}_{2o}^T + \mathbf{E}_{2f}
\end{aligned} \tag{20}$$

where the first part ($\mathbf{T}_{1c}$ $\mathbf{P}_{1c}^T$ and $\mathbf{T}_{2c}$ $\mathbf{P}_{2c}^T$) reveals the systematic information which are really related with each other and can be used to predict each other, revealing the between-set similarity; the second part ($\mathbf{T}_{1o}$ $\mathbf{P}_{1o}^T$ and $\mathbf{T}_{2o}$ $\mathbf{P}_{2o}^T$) reveals the unique systematic information within each data space, which evaluates the between-set dissimilarity.

Property analysis and discussion

(1) By bidirectional modeling manner, both canonical and orthogonal components can be directly associated with their original measurement space, that is, they lie well in the original space with no outside information introduced. This is demonstrated taking example for the $\mathbf{X}_1$-space as below:

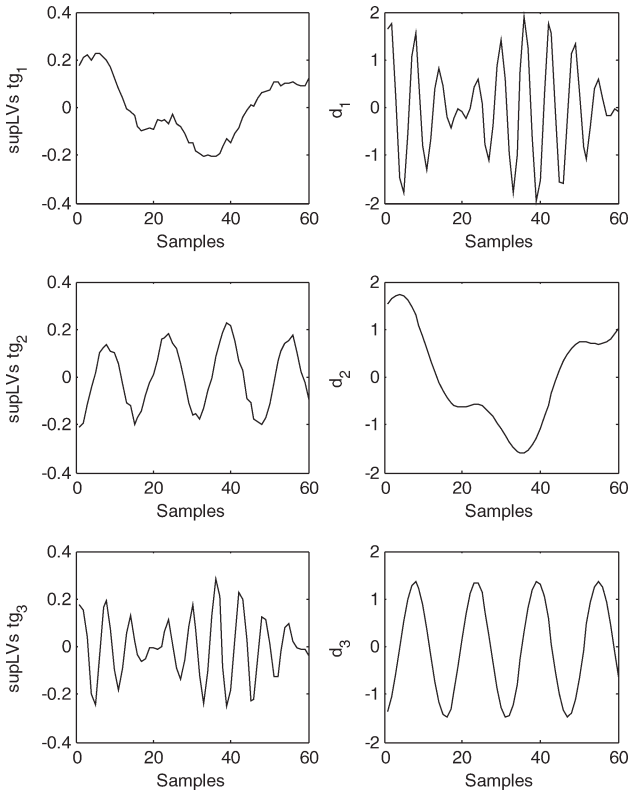

Figure 2. SupLV extraction result and the original source signals.

$$\begin{aligned}\mathbf{T}_{1c} &= \mathbf{T}_1\mathbf{W}_{1c} = \mathbf{X}_1\mathbf{R}_1\mathbf{W}_{1c} = \mathbf{X}_1\theta_{1c}\\ \mathbf{T}_{1o} &= \mathbf{X}_{1e}\mathbf{W}_{1o} = \left(\mathbf{X}_1 - \mathbf{T}_{1c}\mathbf{P}_{1c}^{\mathrm{T}}\right)\mathbf{W}_{1o}\\ &= \left(\mathbf{X}_1 - \mathbf{X}_1\theta_{1c}\mathbf{P}_{1c}^{\mathrm{T}}\right)\mathbf{W}_{1o} = \mathbf{X}_1\theta_{1o}\end{aligned} \tag{21}$$

Moreover, it is easy to realize that the three separated subspaces are actually orthogonal with each other. The basic orthogonal properties are listed as below:

$$\begin{aligned}&\mathbf{T}_{1c}^{\mathrm{T}}\mathbf{T}_{1o} = \mathbf{0}, \quad \mathbf{T}_{1c}^{\mathrm{T}}\mathbf{E}_{1f} = \mathbf{0}, \quad \mathbf{T}_{1o}^{\mathrm{T}}\mathbf{E}_{1f} = \mathbf{0}, \quad \mathbf{T}_{1o}^{\mathrm{T}}\mathbf{T}_{2c} = \mathbf{0}\\ &\mathbf{T}_{2c}^{\mathrm{T}}\mathbf{T}_{2o} = \mathbf{0}, \quad \mathbf{T}_{2c}^{\mathrm{T}}\mathbf{E}_{2f} = \mathbf{0}, \quad \mathbf{T}_{2o}^{\mathrm{T}}\mathbf{E}_{2f} = \mathbf{0}, \quad \mathbf{T}_{2o}^{\mathrm{T}}\mathbf{T}_{1c} = \mathbf{0}\end{aligned} \tag{22}$$

(2) Within each data space, the part reconstructed by canonical components ($\mathbf{T}_{1c}\,\mathbf{P}_{1c}^{\mathrm{T}}$ and $\mathbf{T}_{2c}\,\mathbf{P}_{2c}^{\mathrm{T}}$) is part of the original Bi-LV systematic subspace ($\tilde{\mathbf{X}}_1$ and $\tilde{\mathbf{X}}_2$). To prove this point, it is equivalent to proving that $\mathbf{T}_{1c}\,\mathbf{P}_{1c}^{\mathrm{T}}$ and $\mathbf{T}_{2c}\,\mathbf{P}_{2c}^{\mathrm{T}}$ are orthogonal to the original Bi-LV residual subspaces $\mathbf{E}_1$ and $\mathbf{E}_2$, respectively:

$$\begin{aligned}\mathbf{E}_1^{\mathrm{T}}\mathbf{T}_{1c}\mathbf{P}_{1c}^{\mathrm{T}} &= \mathbf{E}_1^{\mathrm{T}}\mathbf{T}_1\mathbf{W}_{1c}\mathbf{P}_{1c}^{\mathrm{T}} = \mathbf{0}\\ \mathbf{E}_2^{\mathrm{T}}\mathbf{T}_{2c}\mathbf{P}_{2c}^{\mathrm{T}} &= \mathbf{E}_2^{\mathrm{T}}\mathbf{T}_2\mathbf{W}_{2c}\mathbf{P}_{2c}^{\mathrm{T}} = \mathbf{0}\end{aligned} \tag{23}$$

(3) The subspace decomposition is illustratively shown in Figure 1c taking example for the $\mathbf{X}_1$-space. By Bi-LV, the original $\mathbf{X}_1$-space can be preliminarily divided into two parts under the supervision of $\mathbf{X}_2$, the Bi-LV systematic subspace ($\tilde{\mathbf{X}}_1$), and the Bi-LV residual ($\mathbf{E}_1$). Then the close-related variation subspace ($\mathbf{T}_{1c}\,\mathbf{P}_{1c}^{\mathrm{T}}$) is separated from $\mathbf{X}_1$. The new residual ($\mathbf{X}_{1e}$) covers the entire Bi-LV residual ($\mathbf{E}_1$) and part of the Bi-LV systematic subspace ($\tilde{\mathbf{X}}_1$), from which, the orthogonal components can be uncovered, revealing the unique-to-set systematic information ($\mathbf{T}_{1o}\,\mathbf{P}_{1o}^{\mathrm{T}}$). Conclusively, by Bi-JPLV, each data space can be separated into different subspaces, which play different roles in evaluating the between-set relationship.

Bi-JPLV for Process Monitoring

In previous section, the core algorithm has been formulated along with its property analysis. Here, on the basis of the extracted LV models, one of its potential applications with respect to process monitoring is considered.

Modern industrial processes often possess a large number of measured variables, such as flow rates, concentrations, temperatures, and pressures. As analyzed in Introduction, variables that are measured for monitoring can be classified into two groups. The first group consists of variables representing operation conditions such as feed flow rate and a setpoint and so on. The second group consists of variables affected by the operating conditions, such as variables describing composition property. Changes of variable correlations in different groups reflect different operation characteristics and give different status indications. They may also influence each other. By performing bidirectional between-set analysis, one attractive feature is a more comprehensive decomposition of the underlying variations in each data space as well as resulting more specific monitoring behaviors.

It should be noted that after the two-step Bi-JPLV modeling, the information left in the final residual ($\mathbf{E}_{i,\mathrm{f}}$, where subscripts i denotes data space) may still cover some systematic variations although they are useless for between-set relationship interpretation. For the purpose of process monitoring, sometimes PCA modeling on the residual matrix ($\mathbf{E}_{i,\mathrm{f}}$) may be required, which should be first judged.

First, the rate of variations (Rv) can be calculated to evaluate the systematic variations modeled by between-set correlated and orthogonal parts: $\mathrm{Rv}_{i,\mathrm{c+o}}(\%) = \frac{\sum\left(\mathbf{X}_{i,\mathrm{c}}^2 + \mathbf{X}_{i,\mathrm{o}}^2\right)}{\sum \mathbf{X}_i^{\,2}} \times 100$. Moreover, perform PCA on the residual in each data space and calculate the rate of variations counted by each of the first several PCA components: $\mathrm{Rv}_{i,\mathrm{p},j}(\%) = \frac{\sum\left(\mathbf{t}_{i,\mathrm{p},j}\mathbf{p}_{i,\mathrm{p},j}^{\mathrm{T}}\right)^2}{\sum \mathbf{X}_i^{\,2}} \times 100$ (where subscripts p and j denote PCA method and the jth PC, respectively). On the one hand, if $R_{i,\mathrm{c+o}}(\%)$ is larger than a certain value, such as 90%, it means that most of the systematic information in $\mathbf{X}_i$ has been

Table 1. Bi-JPLV Modeling and Statistical Analysis Comparison for Case Study 1: Correlation Analysis Between supLVs and the Original Signals

Correlation Coefficient	$\mathbf{d}_1$	$\mathbf{d}_2$	$\mathbf{d}_3$
$\mathbf{t}_{\mathrm{g},1}$	0.0435	0.9907	0.1783
$\mathbf{t}_{\mathrm{g},2}$	−0.1021	−0.0994	0.9767
$\mathbf{t}_{\mathrm{g},3}$	0.9844	−0.0079	0.0396

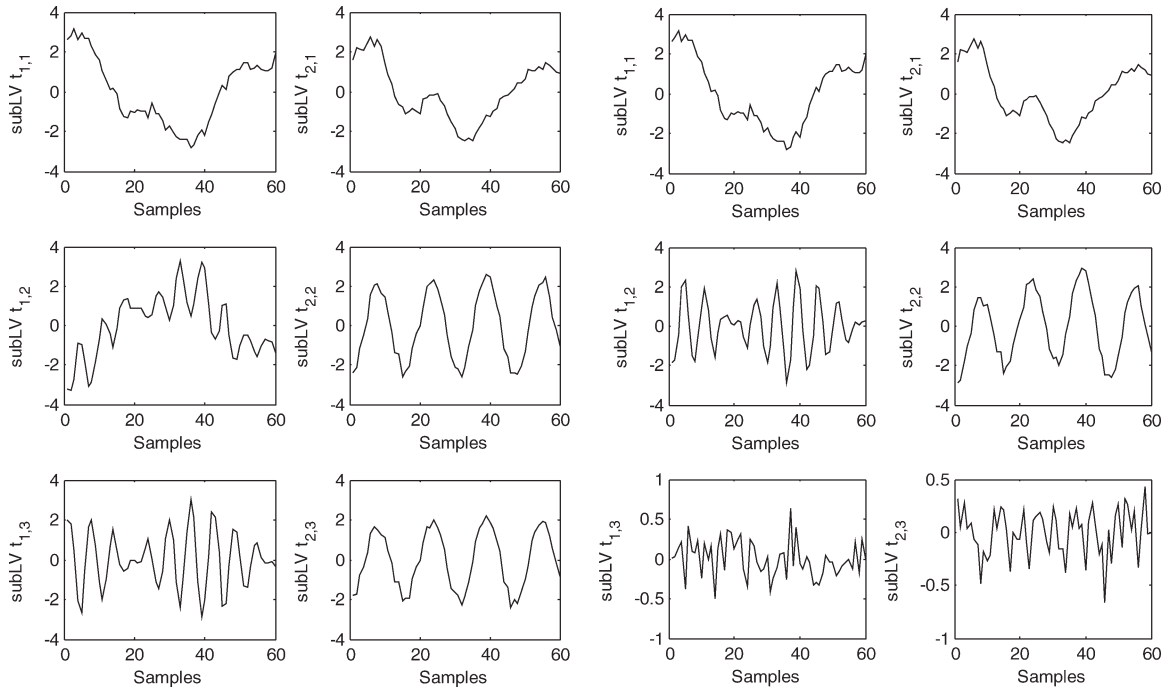

Figure 3. SubLV extraction result.

(a) The original subLVs and (b) the modified subLVs.

counted. On the other hand, by checking $R_{i,\mathrm{p},j}$, one can know whether the PCs can count important systematic information. The two rate indices can be combined to determine whether it is necessary to perform PCA on the residuals $\mathbf{E}_{1\mathrm{f}}$ and $\mathbf{E}_{2\mathrm{f}}$.

If PCA modeling is necessary, the corresponding model structures are formulated in each data space as below:

$$\begin{aligned}\mathbf{E}_{1\mathrm{f}} &= \mathbf{T}_{1\mathrm{p}}\mathbf{P}_{1\mathrm{p}}^{\mathrm{T}} + \mathbf{E}_{1\mathrm{p}} \\ \mathbf{E}_{2\mathrm{f}} &= \mathbf{T}_{2\mathrm{p}}\mathbf{P}_{2\mathrm{p}}^{\mathrm{T}} + \mathbf{E}_{2\mathrm{p}}\end{aligned} \tag{24}$$

where $\mathbf{T}_{1\mathrm{p}}$ and $\mathbf{T}_{1\mathrm{p}}$ are PCA scores, $\mathbf{P}_{1\mathrm{p}}$ and $\mathbf{P}_{2\mathrm{p}}$ are PCA loadings, and $\mathbf{E}_{1\mathrm{p}}$ and $\mathbf{E}_{1\mathrm{p}}$ are PCA residuals.

Then the monitoring system design is described as below. For monitoring purpose, two types of statistics are commonly calculated: the T^2-statistic, which describes the systematic part captured by monitoring models; the Q-statistic, which represents the residual part unoccupied by monitoring models. In each data space, $\mathbf{T}_{i,\mathrm{c}}$, $\mathbf{T}_{i,\mathrm{o}}$, and $\mathbf{T}_{i,\mathrm{p}}$ contain the systematic variation information and will be used for T^2-statistics, whereas the residual, $\mathbf{E}_{i,\mathrm{p}}$, is suitable for Q-statistic:

$$\begin{aligned}T_{i,\mathrm{c,n}}^2 &= \left(\mathbf{t}_{i,\mathrm{c,n}} - \mathbf{t}_{i,\mathrm{c}}\right)^{\mathrm{T}}\mathbf{S}_{i,\mathrm{c}}^{-1}\left(\mathbf{t}_{i,\mathrm{c,n}} - \mathbf{t}_{i,\mathrm{c}}\right) = \mathbf{t}_{i,\mathrm{c,n}}^{\mathrm{T}}\mathbf{t}_{i,\mathrm{c,n}}(N-1) \\ T_{i,\mathrm{o,n}}^2 &= \left(\mathbf{t}_{i,\mathrm{o,n}} - \mathbf{t}_{i,\mathrm{o}}\right)^{\mathrm{T}}\mathbf{S}_{i,\mathrm{o}}^{-1}\left(\mathbf{t}_{i,\mathrm{o,n}} - \mathbf{t}_{i,\mathrm{o}}\right) = \mathbf{t}_{i,\mathrm{o,n}}^{\mathrm{T}}\mathbf{S}_{i,\mathrm{o}}^{-1}\mathbf{t}_{i,\mathrm{o,n}} \\ T_{i,\mathrm{p,n}}^2 &= \left(\mathbf{t}_{i,\mathrm{p,n}} - \mathbf{t}_{i,\mathrm{p}}\right)^{\mathrm{T}}\mathbf{S}_{i,\mathrm{p}}^{-1}\left(\mathbf{t}_{i,\mathrm{p,n}} - \mathbf{t}_{i,\mathrm{p}}\right) = \mathbf{t}_{i,\mathrm{p,n}}^{\mathrm{T}}\mathbf{S}_{i,\mathrm{p}}^{-1}\mathbf{t}_{i,\mathrm{p,n}} \\ \mathrm{SPE}_{i,\mathrm{n}} &= \mathbf{e}_{i,\mathrm{p,n}}^{\mathrm{T}}\mathbf{e}_{i,\mathrm{p,n}}\end{aligned} \tag{25}$$

where subscripts i and n denote data space and observation, respectively. $\mathbf{t}_{i,\mathrm{c,n}}(R_{\mathrm{c}} \times 1)$, $\mathbf{t}_{i,\mathrm{o,n}}(R_{i,\mathrm{o}} \times 1)$, $\mathbf{t}_{i,\mathrm{p,n}}(R_{i,\mathrm{p}} \times 1)$ and $\mathbf{e}_{i,\mathrm{p,n}}(J_i \times 1)$ are the canonical variate, OSC component, principal component, and residual vector of the nth observation in each data space ($\mathbf{X}_i$), respectively, $\bar{\mathbf{t}}_{i,\mathrm{c}}$, $\bar{\mathbf{t}}_{i,\mathrm{o}}$, and $\bar{\mathbf{t}}_{i,\mathrm{p}}$ denote the corresponding mean vectors, which are all zero ones due to the use of mean-centering in data preprocessing procedure, $\mathbf{S}_{i,\mathrm{c}}$, $\mathbf{S}_{i,\mathrm{o}}$, and $\mathbf{S}_{i,\mathrm{p}}$ are the variance-covariance matrices of components, respectively, in which, $\mathbf{S}_{i,\mathrm{c}}$ actually

Table 2. Bi-JPLV Modeling and Statistical Analysis Comparison for Case Study 1: Analysis of Original subLVs

SubLVs		Variance	Correlation Coefficient $\mathbf{d}_1$	$\mathbf{d}_2$	$\mathbf{d}_3$
$\mathbf{X}_1$-space	$\mathbf{t}_{1,1}$	2.8911	0.0452	**0.9928**	0.0809
	$\mathbf{t}_{1,2}$	2.4495	−0.5292	**−0.8612**	−0.0428
	$\mathbf{t}_{1,3}$	1.9626	**0.9890**	−0.0087	−0.0409
$\mathbf{X}_2$-space	$\mathbf{t}_{2,1}$	2.0511	0.0401	**0.9641**	0.3124
	$\mathbf{t}_{2,2}$	2.8790	−0.0432	−0.0029	**0.9899**
	$\mathbf{t}_{2,3}$	2.0471	−0.0201	0.0094	**0.9873**

Table 3. Bi-JPLV Modeling and Statistical Analysis Comparison for Case Study 1: Analysis of Modified subLVs (MsubLVs)

MsubLVs		Variance	Correlation Coefficient $\mathbf{d}_1$	$\mathbf{d}_2$	$\mathbf{d}_3$
$\mathbf{X}_1$-space	$\mathbf{t}_{1,1}$	2.8911	0.0452	**0.9928**	0.0809
	$\mathbf{t}_{1,2}$	1.6628	**−0.9869**	0.0013	0.0551
	$\mathbf{t}_{1,3}$	0.0532	0.0712	−0.0016	0.2222
$\mathbf{X}_2$-space	$\mathbf{t}_{2,1}$	2.0511	0.0401	**0.9641**	0.3124
	$\mathbf{t}_{2,2}$	2.7614	−0.0540	−0.2333	**0.9431**
	$\mathbf{t}_{2,3}$	0.0504	0.2618	0.0300	0.0030

Table 4. Bi-JPLV Modeling and Statistical Analysis Comparison for Case Study 1: Correlation Analysis for subLVs and MsubLVs, Respectively

	SubLVs			MsubLVs		
	$\mathbf{t}_{1,1}$	$\mathbf{t}_{1,2}$	$\mathbf{t}_{1,3}$	$\mathbf{t}_{1,1}$	$\mathbf{t}_{1,2}$	$\mathbf{t}_{1,3}$
$\mathbf{t}_{2,1}$	**0.9595**	0.0004	0.0149	**0.9595**	0.2287	0.0560
$\mathbf{t}_{2,2}$	**0.8285**	0.0201	0.0036	0.0088	0.0401	0.2600
$\mathbf{t}_{2,3}$	0.0110	0.0261	0.0034	0.0737	0.2125	0.1297

converges to an diagonal matrix with all elements to be $N-1$ resulting from the unit length of canonical vectors. Here it should be pointed out when PCA modeling is not needed, the number of monitoring statistics will shrink to three in each data space.

Gaussian-distribution premise provides an important basis for deriving the confidence limits of monitoring statistics. The T^2 control limit can be defined by F-distribution,[28,29] whereas the confidence limit of SPE can be approximated by a weighted Chi-squared distribution[29] $g\chi^2_h$.

Process monitoring is conducted by continuously comparing all the statistics with the predetermined control limits. If they stay well within the predefined normal regions, the current operation can be deemed to be normal. Otherwise the statistics will go beyond the control limits, indicating the occurrence of abnormal behaviors. From the between-set viewpoint, the monitoring charts in response to different types of variations can jointly reveal more abundant process information and thus a more specific tracking of operation status could be achieved.

Simulations and Discussions

Case Study 1

In this simple numerical example, we conclude with an analysis of the algorithm itself. It will first demonstrate how the proposed method models the underlying variations in each data space under the influence of each other. The following two data tables of five variables are considered, where each data space contains a strong direction that is not available in the other space:

$$\begin{aligned} \mathbf{X}_1 &= [\mathbf{d}_1\ \mathbf{d}_1\ \mathbf{d}_2\ \mathbf{d}_2\ \mathbf{d}_2] \\ \mathbf{X}_2 &= [\mathbf{d}_2\ \mathbf{d}_2\ \mathbf{d}_3\ \mathbf{d}_3\ \mathbf{d}_3] \\ d_1(k) &= 2cos(0.08k)sin(0.006k) \\ d_2(k) &= sin(03k) + 3cos(0.1k) \\ d_3(k) &= (2sin(0.2k))^2 \end{aligned} \tag{26}$$

where $\mathbf{d}_i$ ($i = 1,2,3$) denote major distribution directions in each data space, which are irrelevant with each other. Each data space includes two major directions, sharing one common direction $\mathbf{d}_2$ and behaving differently from each other in the other direction (as indicated by $\mathbf{d}_1$ and $\mathbf{d}_3$). Moreover, each direction in each data block is added by 5% normally distributed random noises.

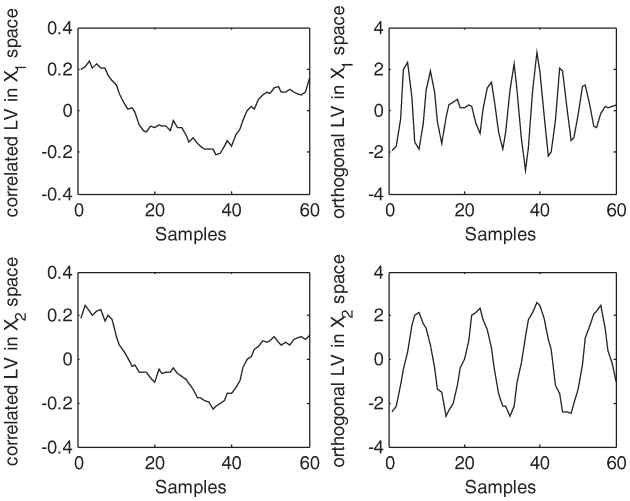

Figure 4. Postprocessing result for Bi-LVs.

Each data set has 100 observations, in which, 60 samples are used for model training, and 40 samples are used for validation. All variables in both data tables are mean-centered and scaled to unit variance. Based on the indication of cumulative explained variance rate,[27] three supLVs ($\mathbf{t}_g$) are chosen to explain the general systematic information in the joint data space [$\mathbf{X}_1$, $\mathbf{X}_2$] which account for more than 90% of overall variability. The recovered three supLVs are shown in Figure 2 in comparison with the original signals, $\mathbf{d}_1$, $\mathbf{d}_2$, and $\mathbf{d}_3$. Clearly, the supLVs actually recover the three major directions and are ordered in accord with the associated variances. The first supLV $\mathbf{t}_{g,1}$ is more similar to $\mathbf{d}_2$, the common direction in $\mathbf{X}_1$ and $\mathbf{X}_2$ spaces. The second supLV corresponds to $\mathbf{d}_3$ and the third supLV leads to $\mathbf{d}_1$. This can be also demonstrated by the correlation analysis result shown in Table 1. Moreover, corresponding to each supLV, the associated subLVs are extracted in both data spaces. The original subLVs and MsubLVs are comparatively shown in Figures 3a, b. Moreover, their variances and correlations with original signals are compared in Tables 2 and 3. For the three original subLVs, each shows indispensible variance and meanwhile reveals a close relationship with certain $\mathbf{d}_i$ (as indicated by the bold). For the MsubLVs, the variances are integrated in the first 2 ones, and only the first 2 MsubLVs in each space have close relationships with certain $\mathbf{d}_i$ (marked in bold). This tells us that by original subLVs, as calculated by Eq. 13, the systematic information in each specific data space is repetitively extracted. By the modification

Table 5. Bi-JPLV Modeling and Statistical Analysis Comparison for Case Study 1: Modeling Variations (%)

Bi-LV Modeling Result		Bi-JPLV Modeling Result				Prediction	
$R^2\tilde{\mathbf{X}}_1$	$R^2\tilde{\mathbf{X}}_2$	$R^2\mathbf{X}_{1c}$	$R^2\mathbf{X}_{1o}$	$R^2\mathbf{X}_{2c}$	$R^2\mathbf{X}_{2o}$	$R^2\hat{\mathbf{X}}_1$	$R^2\hat{\mathbf{X}}_2$
96.9237	97.2946	57.8228	39.0923	39.3496	57.9449	56.2540	38.2816
		96.9151		97.2945			

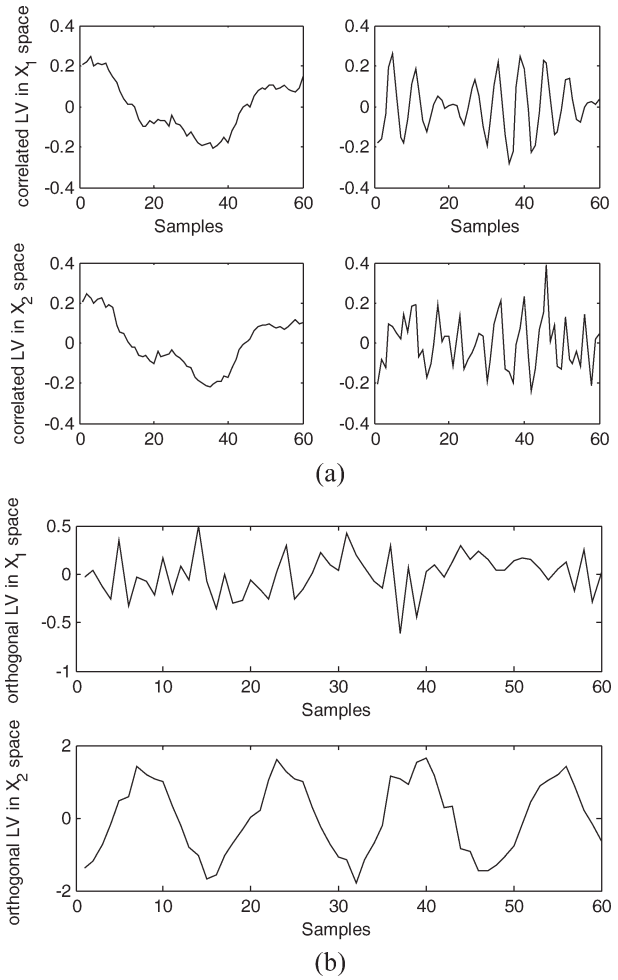

Figure 5. Postprocessing result for PLS LVs (a) correlated components (b) orthogonal.

strategy shown in Appendix, orthogonal MsubLVs can be sorted automatically by their relationships with supLVs. The first 2 MsubLVs can collect the major systematic variation, and meanwhile show larger correlations with one $\mathbf{d}_i$, whereas the third MsubLV only denotes noise information. Combining the indications of variance and correlation, in each data space, only two MsubLVs should be retained. Moreover, according to the profiles of $\mathbf{d}_i$ shown in Figure 2, the MsubLVs in $\mathbf{X}_1$-space actually recover $\mathbf{d}_2$ and $\mathbf{d}_1$ and the MsubLVs in $\mathbf{X}_2$-space actually recover $\mathbf{d}_2$ and $\mathbf{d}_3$, which well agrees with the real situation. As shown in Table 4, correlation analysis is performed for original subLV pairs and MsubLV pairs, respectively. As the bold values indicate, only the first pair of MsubLVs is correlated with each other, whereas the first 2 original subLVs in $\mathbf{X}_1$ are both correlated with the first subLV in $\mathbf{X}_2$ space. Then postprocessing is performed to separate the between-set closely related part from the irrelevant part. Only one canonical variable should be kept according to the correlation coefficients of the two pairs of CCA components, which are 0.9863 and 0.0410, respectively. One OSC component is retained to extract the between-set irrelevant information. As shown in Figure 4, the extracted CCA components and OSC components are comparatively plotted in both data spaces. The CCA component is similar to $\mathbf{d}_2$, the common signal in both data spaces as shown in Eq. 26. The OSC component in $\mathbf{X}_1$-space is similar to $\mathbf{d}_1$ and that in $\mathbf{X}_2$-space is similar to $\mathbf{d}_3$, both being unique direction in each data space. Their associated variations are also calculated as shown in Table 5, which quantitatively demonstrate that different types of systematic information can be separated corresponding to different between-set relationships. Moreover, it can be clearly seen that $R^2\mathbf{X}_{1c}$ and $R^2\mathbf{X}_{2c}$ are different from $R^2\hat{\mathbf{X}}_1$ and $R^2\hat{\mathbf{X}}_2$, respectively, as mentioned after Eq. 18.

Comparatively, PLS is used to prepare LVs for the following postprocessing procedure. As analyzed before, because of the deflation by $\mathbf{X}_1$-subLVs ($\mathbf{T}$) in $\mathbf{X}_2$-space, the $\mathbf{X}_1$-information is introduced to $\mathbf{X}_2$-space and mixed in $\mathbf{X}_2$-LVs $\mathbf{U}$. To demonstrate this point, the PLS LVs ($\mathbf{T}$ and $\mathbf{U}$) are postprocessed by CCA, where the number of PLS LVs are kept to be up to 3 so that the influence of information mixing problem can be more obvious. For the first 2 canonical components, their profiles are shown in Figure 5. Clearly, the first CCA component looks like $\mathbf{d}_2$ which agrees well with the real situation. The second component is like $\mathbf{d}_1$, which, however, conflicts with the real case because it is well known that $\mathbf{d}_1$ only exists in $\mathbf{X}_1$-space. Resulting from the information mixing problem, the between-set similarity is enhanced wrongly as indicated by correlations between CCA component pairs, which are 0.9881 and 0.6331, respectively, telling both CCA components should be retained. As more information ($\mathbf{d}_1$ and $\mathbf{d}_2$) are regarded as between-set related part, no OSC component in $\mathbf{X}_1$-space can be extracted as shown in Figure 5b where the forcibly extracted orthogonal component in $\mathbf{X}_1$ is more like to be noise signal rather than systematic information. The above analysis reveals the influence of ''information mixing'' problem in

Table 7. Process Disturbances for Tennessee Eastman Process (TEP)

Fault No.	Process Variable	Fault Type
1	A/C feed ratio, B composition constant	Step
2	Reactor cooling water valve	Sticking
3	Reaction kinetics	Slow drift
4	C feed temperature	Random variation

Table 6. Bi-JPLV Modeling and Statistical Analysis Comparison for Case Study 1: PLS-Based Joint Modeling and Quantitative Analysis Result

PLS Modeling Result		Post-Processing Result				Prediction	
$R^2\tilde{\mathbf{X}}_1$	$R^2\tilde{\mathbf{X}}_2$	$R^2\mathbf{X}_{1c}$	$R^2\mathbf{X}_{1o}$	$R^2\mathbf{X}_{2c}$	$R^2\mathbf{X}_{2o}$	$R^2\hat{\mathbf{X}}_1$	$R^2\hat{\mathbf{X}}_2$
97.9280	98.0230	96.9643	0.9637	40.2229	56.5210	72.0826	41.8197
		96.9151		97.2945			

Table 8. Bi-JPLV Modeling Result for Tennessee Eastman Process (TEP)

Data Space	$\mathbf{X}_1$-Space				$\mathbf{X}_2$-Space			
Model order	Bi-LV	Post-processing			Bi-LV	Post-processing		
		CCA	OSC	PCA		CCA	OSC	PCA
	23	7	16	0	11	7	4	0
Modeling variations (%)	94.1037	40.8800	53.2273	0	90.7612	62.1522	28.6091	0

PLS on the between-set relationship interpretation. Especially, it can be readily deduced that when more PLS LVs are kept and more deflation calculations are performed, the influence will be more serious. The variations counted by different parts are also calculated and shown in Table 6 in comparison with the result shown in Table 5, which quantitatively demonstrates the above analysis. The PLS $\mathbf{X}_2$-LVs have been contaminated by information from $\mathbf{X}_1$-space and will seriously influence the postprocessing result. For example, the correlated components in $\mathbf{X}_1$-space wrongly count much more variations ($R^2\mathbf{X}_{1c}$) than real case where the common direction $\mathbf{d}_2$ actually only accounts for ~60% in $\mathbf{X}_1$-space. Moreover, it should be noted that although the $\mathbf{X}_2$-CCA components have been contaminated, they approximately count the same amount of variations in $\mathbf{X}_2$ (40%) as what $\mathbf{d}_2$ does because the introduced $\mathbf{d}_1$ information actually does not exist in $\mathbf{X}_2$-space. Moreover, because $\mathbf{d}_1$ is wrongly taken as the common direction, there should be no orthogonal systematic information in $\mathbf{X}_1$-space. This is well demonstrated by $R^2\mathbf{X}_{1o}$, where the value is very less (approaching zero), means that it may only reveal noises.

In summary, the simple numerical example illustrates the objective function shown in Eq. 1, e.g., how the Bi-LVs are extracted from both ways and how they are prepared for the second-step postprocessing. By the proposed between-set statistical analysis strategy, the underlying information in each data space is explored in detail under the supervision of each other. Expressed by different types of LVs, the related part, revealing between-set similarity, are separated from the orthogonal part, revealing between-set dissimilarity. The associated different types of variations are quantitatively explored in each data space. Comparatively, the disadvantage of PLS with respect to its unidirectional fashion is revealed and its influence on between-set relationship analysis is also probed.

Case Study 2

In this case study, the proposed monitoring strategy is tested by the well-known Tennessee Eastman (TE) benchmark chemical process. TE process has been widely used for verifying various process monitoring methods[30–33] as it was first introduced by Downs and Vogel.[30] The process is

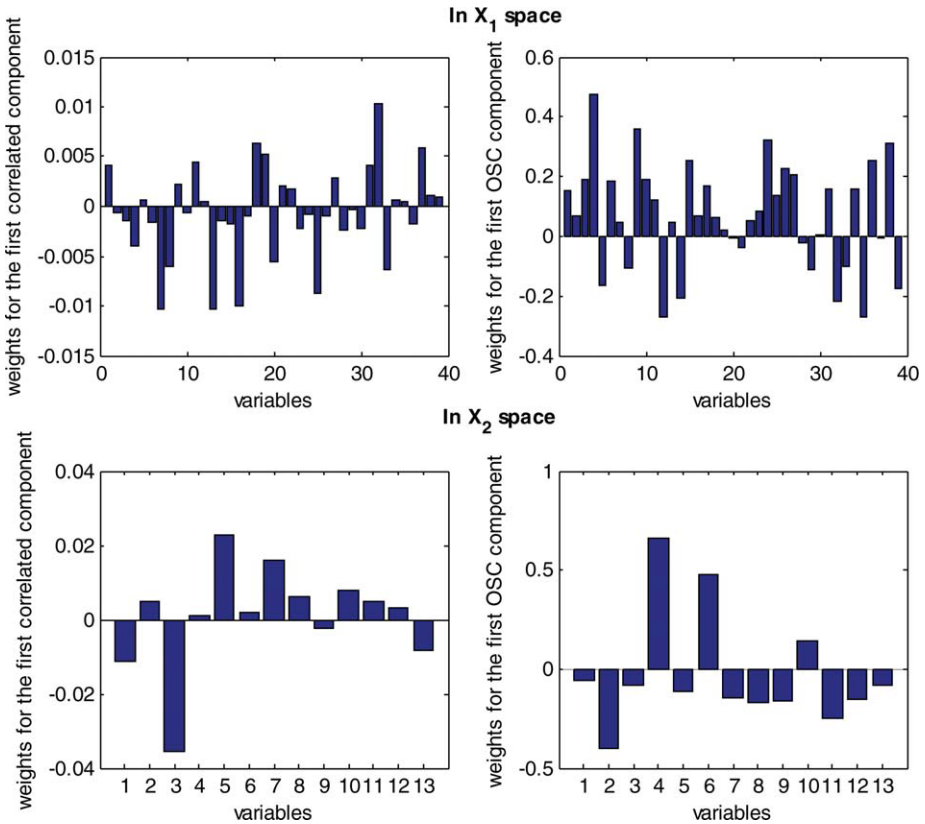

Figure 6. Weight coefficients for systematic components in each data space for TEP.

[Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

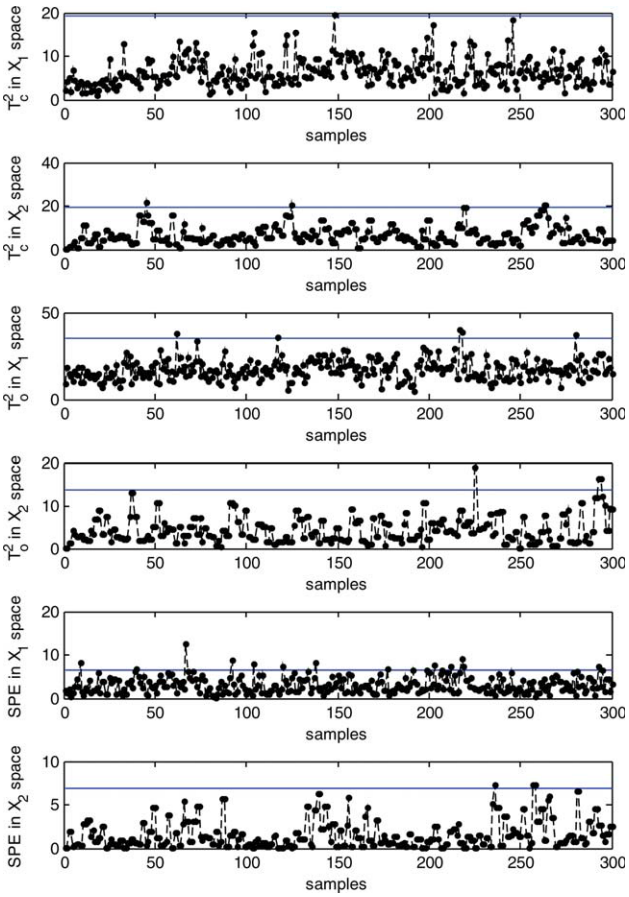

Figure 7. Online monitoring result for the normal case (solid line, 99% control limit; dot line, on-line monitoring statistics).

[Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

constructed by five major operation units: a reactor, a product condenser, a vapor–liquid separator, a recycle compressor, and a product stripper. The operation flow is described as follow.[30] The gaseous reactants are fed to the reactor where they react to form liquid products. The gas phase reactions are catalyzed by a nonvolatile catalyst dissolved in the liquid phase. The reactor has an internal cooling bundle for removing the heat of reaction. The products leave the reactor as vapors along with the unreacted feeds. The catalyst remains in the reactor. The reactor product stream passes through a cooler for condensing the products and from them to a vapor–liquid separator. Noncondensed components recycle back through a centrifugal compressor to the reactor feed. Condensed components move to a product stripping column to remove remaining reactants. In the process, two products are created from four reactants by means of two irreversible and exothermic reactions. Also present are an inert and a byproduct making a total of eight components. More details of this industrial process model have been described by Downs and Vogel.[30]

In this study, the simulation data is downloaded from the website http://brahms.scs.uiuc.edu. It includes 52 input variables. Based on the reaction characteristics, streams exiting the stripper base and the vapor–liquid separator, which cover the concerned product measurement, are used as one data set, including 13 variables, whereas the other data set involves the left 39 variables. In this way, the product variability can be related with the operation conditions by performing between-set statistical analysis, in which correlated and orthogonal variations are separated and distinguished. Therefore, faults can be explored more specifically by checking their impacts on different types of variations. Three hundred normal data samples are used to train the monitoring model and another 300 samples to verify the model. In addition, more process data are used for testing, including one normal case and 12 fault cases. The programmed faults are listed in Table 7, covering four different types of disturbances, step, random variation, slow drift, and sticking.

First, the modeling result is summarized in Table 8, telling the number of different systematic components retained to describe different types of variations. Moreover, based on the rate of variations, PCA is not needed because CCA plus OSC has counted most systematic information in each data space (>90%) and the PCA scores count little variations. Moreover, in $\mathbf{X}_2$-space, obviously, the correlated variations are more than orthogonal ones, whereas in $\mathbf{X}_1$-space, the two variations share comparative amounts. The weight

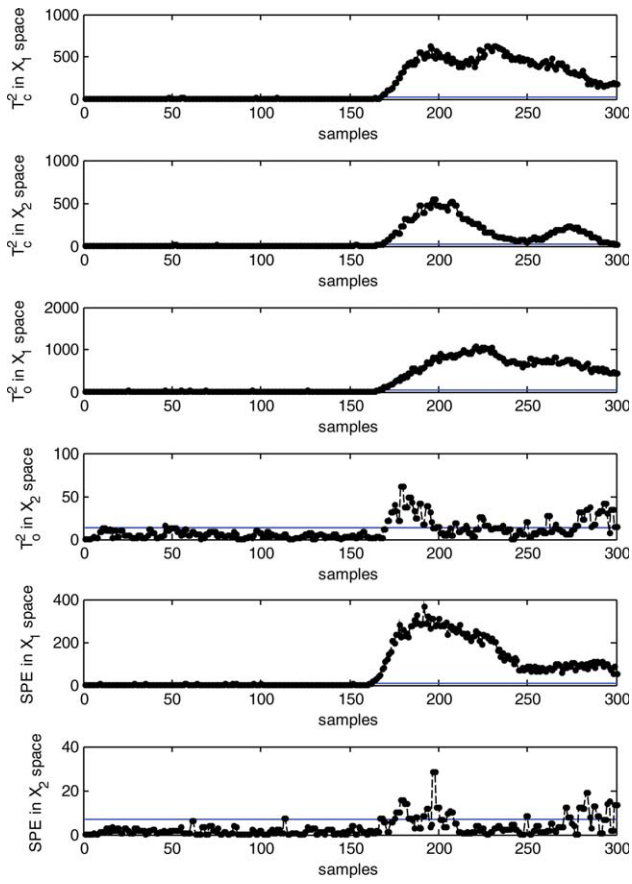

Figure 8. Online monitoring result for fault 1 (solid line, 99% control limit; dot line, on-line monitoring statistics).

[Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

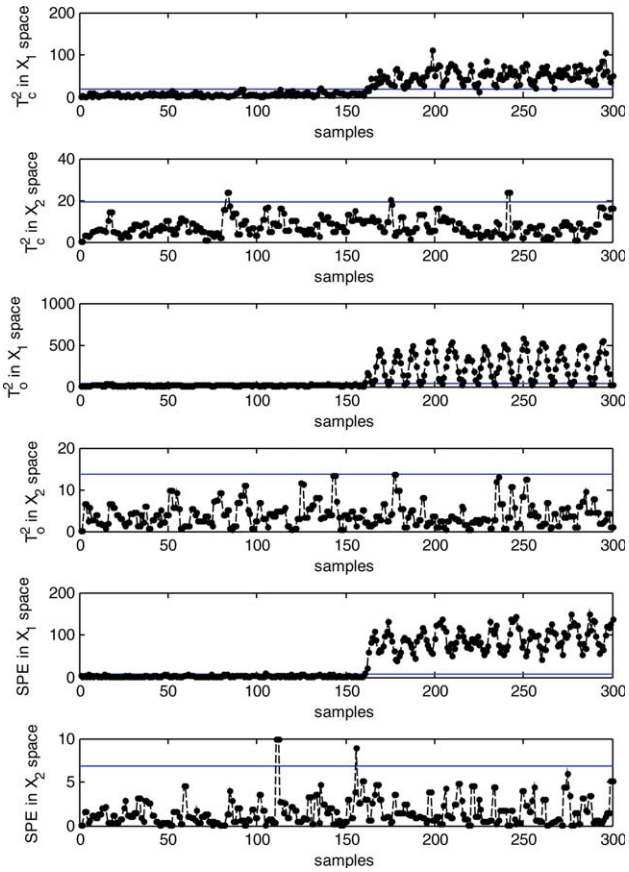

Figure 9. Online monitoring result for fault 2 (solid line, 99% control limit; dot line, on-line monitoring statistics).

[Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

coefficients to derive the first CCA and OSC scores are plotted in Figure 6 for each data space. Based on the weights, it is easy to check what variables dominate in explaining the between-set similarity and what are important in revealing between-set dissimilarity. For example, in $\mathbf{X}_2$-space, the 3rd, 5th, and 7th variables are important when explaining the correlated part, whereas the 2nd, 4th, and 6th variables are of concern to capture the orthogonal part. Therefore, before online monitoring, preanalysis can be conducted and a preliminary understanding of the underlying information is obtained.

Then the monitoring results for the normal case and four fault cases are illustratively shown in Figures 7–11 using the proposed algorithm. For the normal case, the monitoring chart of Bi-JPLV shows that the process is tracking the desired operation trajectory because all monitoring statistics stay well below the control limits. For abnormal cases, generally, the monitoring charts make a quick response to the disturbances with the statistic values delivering significant deviation from the normal region. In detail, for fault 1, where *A*/*C* feed ratio and *B* composition constant have step changes, the disturbance actually roots in $\mathbf{X}_1$-space. From the monitoring plots, it is seen that all monitoring indices in $\mathbf{X}_1$-space clearly show out-of-control indications. Resulting from the between-set correlation, the influence of abnormal behaviors is introduced to $\mathbf{X}_2$-space, which is more serious in the $\mathbf{T}_{2c}^2$ monitoring chart since $\mathbf{T}_{2c}$ are closely correlated with $\mathbf{T}_{1c}$. Moreover, although $\mathbf{T}_{2o}$ is orthogonal with $\mathbf{T}_{1c}$, it has some relationship with $\mathbf{T}_{1o}$ more or less. Therefore, $\mathbf{T}_{2o}^2$ also escapes from normal operation region but with smaller extent. For fault 2, where reactor cooling water valve is stuck, the monitoring illustrations tell one that the abnormal behaviors mainly influence the $\mathbf{X}_1$-space with all monitoring statistics beyond the normal region. Moreover, the influence of this fault is not introduced into $\mathbf{X}_2$-space, where all three monitoring indices follow the desired operation trajectory well. For fault 3, where slow drifting happens in reaction kinetics, the underlying variations in both spaces are seriously disrupted as shown in Figure 10 where continuous alarms are clearly indicated after a period of time. For fault 4, random variations occur in *C* feed temperature. From the monitoring graphs, its influence on the process behaviors mainly lies in $\mathbf{X}_1$-space although the alarms are not so continuous compared with those shown in faults 1–3. Moreover, taking example for the fourth fault, the monitoring results using PCA and PLS algorithms, respectively, are shown in Figure 12. By PLS, to make use of the $\mathbf{X}_2$-information, SPE statistic index is also calculated using the residual in $\mathbf{X}_2$ after prediction by $\mathbf{X}_1$-LVs. By PCA and PLS, one can only know

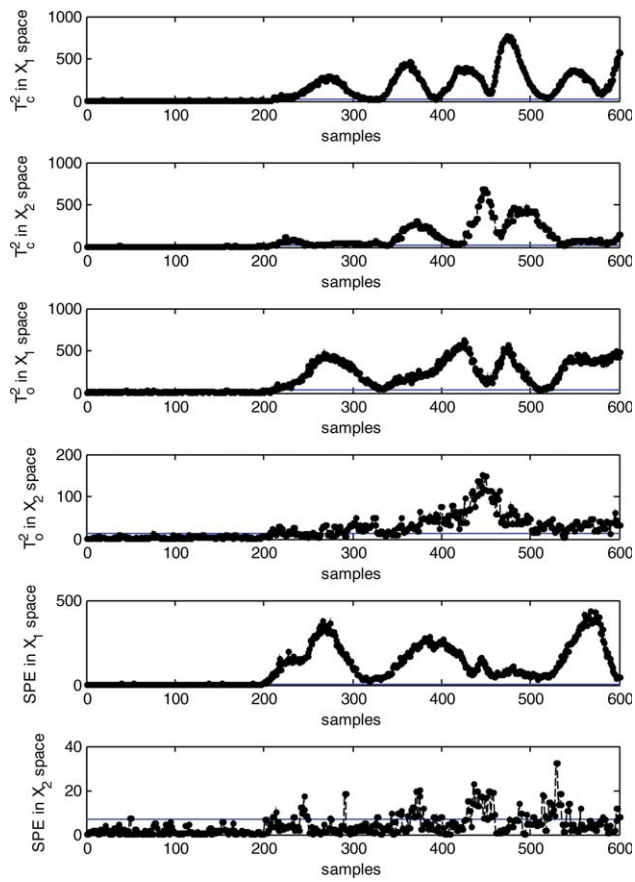

Figure 10. Online monitoring result for fault 3 (solid line, 99% control limit; dot line, on-line monitoring statistics).

[Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

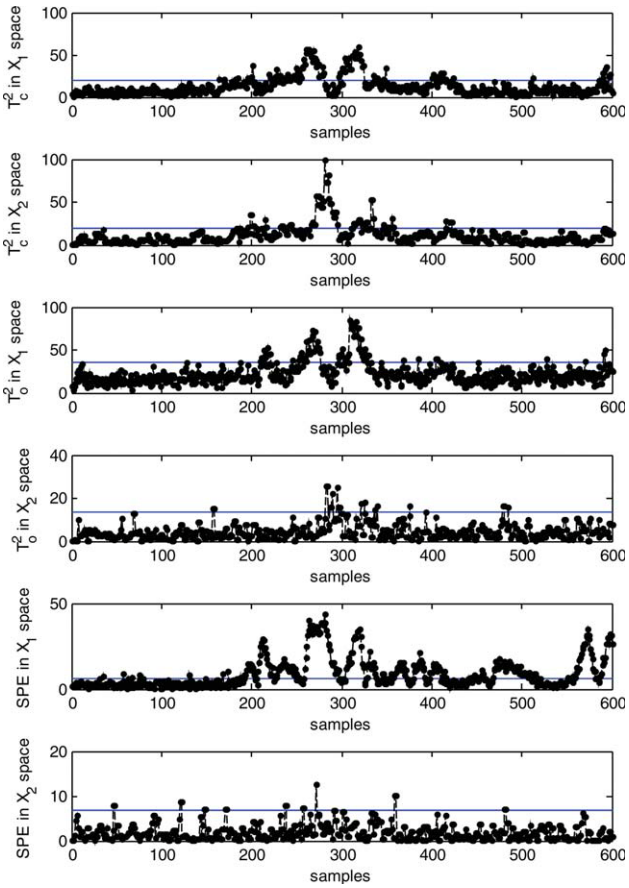

Figure 11. Online monitoring result for fault 4 (solid line, 99% control limit; dot line, on-line monitoring statistics).

[Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

whether changes of the variable correlation structure occur or whether the disturbances in $\mathbf{X}_1$-space will influence the prediction of $\mathbf{X}_2$. However, we cannot know how the underlying information changes in detail under the influence of disturbances as what has been shown in Figure 11 by the proposed method. Bidirectional between-set analysis enables the separation of different variations and a better representation of underlying information so that abnormality in each data space could be noticed better. Hence, an experienced operator would quickly check and locate the problem within the operation process.

Summary and discussion

This report has illustrated how the proposed modeling and analysis strategy performs by both simulated and real data sets. There may be still many modeling issues to be investigated in future, but the results of this study can provide the basis for further work and improvement, which might profitably take the following directions.

The Variation Balance Between X_1 and X_2. From the objective function shown in Eq. 1, there is no guarantee that each data table contributes the same, revealing the common bias optimization issue in multioptimization problem, which thus result in the following considerations. On one hand, in this work, both data sets are just simply normalized to be zero mean and unit variance. Other scaling methods can also be used as alternatives when data sets have obviously different numbers of variables. For example, each data set can be scaled to equal the square root of the number of variables in that space or the sum of squares. This may partially facilitate the modeling performance. On the other hand, by imposing constraints, the variations explained in each data space can be controlled to a certain extent. For example, tuning parameters that range between 0 and 1 can be introduced so that

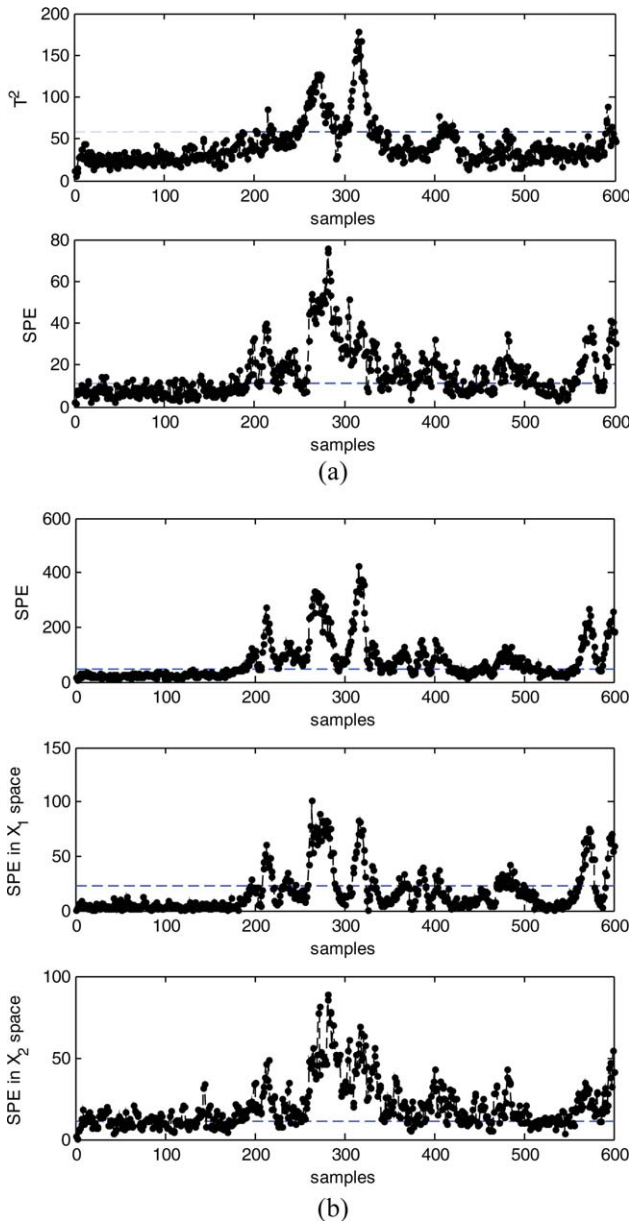

Figure 12. Monitoring result for fault 4 using (a) PCA and (b) PLS (solid line, 99% control limit; dot line, on-line monitoring statistics).

[Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

the objective function can be expressed as $\alpha\left(\mathbf{t}_g^T\mathbf{X}_1\mathbf{a}_1\right)^2+\beta\left(\mathbf{t}_g^T\mathbf{X}_2\mathbf{a}_2\right)^2$ subject to $\alpha+\beta=1$. Considering the extreme case, $\alpha=1$, $\beta=0$, it converges to simple PCA in $\mathbf{X}_1$-space without considering the influence of $\mathbf{X}_2$-space, and vice versa. With larger α value and less β value, the systematic variations in $\mathbf{X}_1$-space are more focused on. On the contrary, a larger β value will pay more attention to the underlying variations in the $\mathbf{X}_2$-space. The effects of tuning parameters on LV extraction and between-set analysis have not yet been reduced to an explicit mathematical formula. The proper setting of parameter values may have to be determined by cross validation. Considering that the parameter values actually determine the variation balance between the two different data spaces, is it possible to work out some simple rule to follow? Answers to the question may not easily be given. As a meaningful issue, this will be the subject of follow-up work, which may be especially fruitful and deserve further investigation.

Nonlinear Problem. For reasons of computational and conceptual simplicity, the nature of the current modeling method is linear. For real industry batch processes, it is not uncommon that the underlying characteristics are nonlinear to some extent, which introduces extra complications. Linear analysis method may not function well to exploit the nonlinear data structure. In such a case, it is desirable to use nonlinear techniques to handle the problem aroused by nonlinear behaviors.

Dynamics Consideration. When process dynamics cannot be ignored, the between-set relationship cannot be explored sufficiently enough by using static statistical analysis strategy. In such a case, dynamic model structures should be built. The simplest way of building a dynamic system is to use past measurements combined with the current ones as inputs because one data space at time k also has relationships with the other data space in the past.

Conclusions

In this work, a Bi-LV modeling version is designed and combined with a Bi-JPLV for between-set statistical analysis. In the general formulation, the proposed method differs from conventional regression methods in the way it distinguishes the underlying between-set related and unique sources in both data spaces from bidirectional viewpoint. It improves the interpretational functionalities to give estimate of (a) the $\mathbf{X}_1$–$\mathbf{X}_2$ correlated information, (b) $\mathbf{X}_1$-irrelated variation (i.e., unique variation) in $\mathbf{X}_2$-space, and (c) the $\mathbf{X}_2$-unique variation in $\mathbf{X}_1$-space. All systematic variation information is explained in different model parameters. This allows for the interpretation of each data space to be more comprehensive and thus better process understanding. It should be generally applicable to a broad range of practical cases. The simulations designed in this study only illustrate its feasibility for online monitoring due to the lack of space. It would be interesting to consider its use in other analysis fields. For example, for calibration analysis, when the same set of samples are subject to different instrumental analysis methods, e.g., IR and mass spectrometer, respectively, their underlying characteristics may be different but also share similar ones to some extent. It is interesting to perform between-set statistical analysis and explore their underlying information under the supervision of each other, which can give insight into what information about the samples is uniquely present in either the IR or the mass spectra, and what kind of information commonly exists in both spectra. Considerable further research is thus recommended.

Acknowledgments

The work is supported in part by China National 973 program (2009CB320603), Hong Kong Research Grant Council (613107), and National Natural Science Foundation of China (No. 60774068).

Literature Cited

1. Martens H, Naes T. *Multivariate Calibration*, 2nd ed. Chichester: Wiley, 1994.
2. Burnham AJ, Viveros R, MacGregor JF. Frameworks for latent variable multivariate regression. *J Chemom.* 1996;10:31–45.
3. Doyal BS, MacGregor JF. Improved PLS algorithms. *J Chemom.* 1997;11:73–85.
4. Brereton RG. Introduction to multivariate calibration in analytical chemistry. *Analyst.* 2000;125:2125–2154.
5. Kleinbaum DG, Kupper LL, Muller KE, Nizam A. *Applied Regression Analysis and Other Multivariable Methods*, 3rd ed. California: Wadsworth Publishing Co. Inc., 2003.
6. Gustafsson MG. Independent component analysis yields chemically interpretable latent variables in multivariate regression. *J Chem Inf Model.* 2005;45:1244–1255.
7. Ergon R. Reduced PCR/PLSR models by subspace projections. *Chemom Intell Lab Syst.* 2006;81:68–73.
8. Zhao CH, Gao FR, Wang FL. An improved independent component regression modeling and quantitative calibration procedure. *AIChE J.* 2010;56:1519–1535.
9. Cserhati T, Kosa A, Balogh S. Comparison of partial least-square method and canonical correlation analysis in a quantitative structure-retention relationship study. *J Biochem Biophys Methods.* 1998;36:131–141.
10. Anderson TW. Canonical correlation analysis and reduced rank regression in autoregressive models. *Ann Stat.* 2002;30:1134–1154.
11. Hardoon DR, Szedmak S, Taylor JS. Canonical correlation analysis: an overview with application to learning methods. *Neural Comput.* 2004;16:2639–2664.
12. Yamamoto H, Yamaji H, Fukusaki E, Ohno H, Fukuda H. Canonical correlation analysis for multivariate regression and its application to metabolic fingerprinting. *Biochem Eng J.* 2008;40:199–204.
13. Yu HL, MacGregor JF. Post processing methods (PLS-CCA): simple alternatives to preprocessing methods (OSC-PLS). *Chemom Intell Lab Syst.* 2004;73:199–205.
14. Wold S, Antti H, Lindgren F, Öhman J. Orthogonal signal correction of near-infrared spectra. *Chemom Intell Lab Syst.* 1998;44:175–185.
15. Fearn T. On orthogonal signal correction. *Chemom Intell Lab Syst.* 2000;50:47–52.
16. Westerhuis JA, De Jong S, Smilde AK. Direct orthogonal signal correction. *Chemom Intell Lab Syst.* 2001;56:13–25.
17. Trygg J, Wold S. Orthogonal projections to latent structures (O-PLS). *J Chemom.* 2002;16:119–128.
18. Zhou DH, Li G, Qin SJ. Total projection to latent structures for process monitoring. *AIChE J.* 2010;56:168–179.
19. Trygg J. O2-PLS for qualitative and quantitative analysis in multivariate calibration. *J Chemom.* 2002;16:283–293.
20. Trygg J, Wold S. O2-PLS, a two-block (X-Y) latent variable regression (LVR) method with an integral OSC filter. *J Chemom.* 2003;17:53–64.
21. Lindgren F, Geladi P, Wold S. The kernel algorithm for PLS. *J Chemom.* 1993;7:45–59.
22. Nomikos P, MacGregor JF. Multi-way partial least squares in monitoring batch processes. *Chemom Intell Lab Syst.* 1995;30:97–108.
23. Duchesne C, MacGregor CD. Multivariate analysis and optimization of process variable trajectories for batch processes. *Chemom Intell Lab Syst.* 2000;51:125–137.

24. Chu YH, Lee YH, Han C. Improved quality estimation and knowledge extraction in a batch process by bootstrapping-based generalized variable selection. *Ind Eng Chem Res*. 2004;43:2680–2690.
25. Zhao CH, Wang FL, Mao ZZ, Lu NY, Jia MX. Improved batch process monitoring and quality prediction based on multiphase statistical analysis. *Ind Eng Chem Res*. 2008;47:835–849.
26. Sharma SK, Kruger U, Irwin GW. Deflation based nonlinear canonical correlation analysis. *Chemom Intell Lab Syst*. 2006;83:34–43.
27. Lu NY, Gao FR, Wang FL. Sub-PCA modeling and on-line monitoring strategy for batch processes. *AIChE J*. 2004;50:255–259.
28. Ryan TP. *Statistical Methods for Quality Improvement*. New York: Wiley, 1989.
29. Nomikos P, MacGregor JF. Multivariate SPC charts for monitoring batch processes. *Technometrics*. 1995;37:41–59.
30. Downs JJ, Vogel EF. A plant-wide industrial process control problem. *Comput Chem Eng*. 1993;17:245–255.
31. Kano M, Ohno H, Hasebe S, Hashimoto I. Statistical process monitoring based on dissimilarity of process data. *AIChE J*. 2002;48:1231–1240.
32. Wilson DJH, Irwin GW. PLS modeling and fault detection on the Tennessee Eastman benchmark. *Int J Syst Sci*. 2000;31:1449–1457.
33. Lee G, Han CH, Yoon ES. Multiple-fault diagnosis of the Tennessee Eastman process based on system decomposition and dynamic PLS. *Ind Eng Chem Res*. 2004;43:8037–8048.

Appendix: Bi-LV Extraction Algorithm

Input data

Two different data tables $\mathbf{X}_1$ and $\mathbf{X}_2$.

Step 1: By Eq. 10, the systematic variations underlying the joint data space $[\mathbf{X}_1, \mathbf{X}_2]$ are extracted and collected in R supLVs $\mathbf{T}_g = [\mathbf{T}_{g,1}, \mathbf{T}_{g,2},\ldots,\mathbf{T}_{g,R}]$. The original subLVs ($\overline{\mathbf{T}}_1$ and $\overline{\mathbf{T}}_2$) are also calculated in each specific data space based on Eq. 13.

Step 2: Calculate the correlation coefficients between each pair of subLVs. Find the maximum correlation value as well as the corresponding supLV vector $\mathbf{t}_g^*$ and subLVs pairs ($\mathbf{t}_1^*$ and $\mathbf{t}_2^*$). Then deflate the associated variation from their respective data space:

$$\begin{aligned} \mathbf{p}_1^{\mathrm{T}} &= \left(\mathbf{t}_1^{*\mathrm{T}}\mathbf{t}_1^*\right)^{-1}\mathbf{t}_1^{*\mathrm{T}}\mathbf{X}_1 \\ \mathbf{E}_1 &= \mathbf{X}_1 - \mathbf{t}_1^*\mathbf{p}_1^{\mathrm{T}} \end{aligned} \tag{A1}$$

$$\begin{aligned} \mathbf{p}_2^{\mathrm{T}} &= \left(\mathbf{t}_2^{*\mathrm{T}}\mathbf{t}_2^*\right)^{-1}\mathbf{t}_2^{*\mathrm{T}}\mathbf{X}_2 \\ \mathbf{E}_2 &= \mathbf{X}_2 - \mathbf{t}_2^*\mathbf{p}_2^{\mathrm{T}} \end{aligned} \tag{A2}$$

Step 3: Exclude $\mathbf{t}_g^*$ from supLVs and get the new version, $\mathbf{T}_g^\bullet$. Then update the original subLV pairs ($\mathbf{T}_1^\bullet$ and $\mathbf{T}_2^\bullet$) based on new $\mathbf{T}_g^\bullet$ and data spaces ($\mathbf{E}_1$ and $\mathbf{E}_2$) using Eq. 13.

For subsequent MsubLVs, repeat Steps 2 and 3.

Output results

In two different data spaces ($\mathbf{X}_1$ and $\mathbf{X}_2$), respectively, MsubLVs $\mathbf{T}_1$ and $\mathbf{T}_2$, weights $\mathbf{W}_1$ and $\mathbf{W}_2$, and loadings $\mathbf{P}_1$ and $\mathbf{P}_2$.

The MsubLVs can be also directly computed from the original matrices by the equation $\mathbf{T}_1 = \mathbf{X}_1\mathbf{R}_1\mathbf{T}_2 = \mathbf{X}_2\mathbf{R}_2$, where $\mathbf{R}_1 = \mathbf{W}_1(\mathbf{P}_1^{\mathrm{T}}\ \mathbf{W}_1)^{-1}$, $\mathbf{R}_2 = \mathbf{W}_2(\mathbf{P}_2^{\mathrm{T}}\ \mathbf{W}_2)^{-1}$.

Manuscript received Dec. 10, 2009, and revision received May 4, 2010.