New Approach to Generalized Two-Dimensional Correlation Spectroscopy. 1: Combination of Principal Component Analysis and Two-Dimensional Correlation Spectroscopy

YOUNG MEE JUNG,* HYEON SUK SHIN, SEUNG BIN KIM*, and ISAO NODA*

Department of Chemistry, Pohang University of Science and Technology, San 31, Hyojadong, Pohang 790-784, Korea (Y.M.J., H.S.S., S.B.K.); and The Procter and Gamble Company, 8611 Becket Road, West Chester, Ohio 45069 (I.N.)

The direct combination of chemometrics and two-dimensional (2D) correlation spectroscopy is considered. The use of a reconstructed data matrix based on the significant scores and loading vectors obtained from the principal component analysis (PCA) of raw spectral data is proposed as a method to improve the data quality for 2D correlation analysis. The synthetic noisy spectra were analyzed to explore the novel possibility of the use of PCA-reconstructed spectra, which are highly noise suppressed. 2D correlation analysis of this reconstructed data matrix, instead of the raw data matrix, can significantly reduce the contribution of the noise component to the resulting 2D correlation spectra.

Index Headings: Two-dimensional (2D) correlation spectroscopy; Principal component analysis; PCA; PCA 2D correlation spectroscopy.

INTRODUCTION

In recent years generalized two-dimensional (2D) correlation spectroscopy has been applied to infrared (IR), near-infrared (NIR), Raman, ultraviolet-visible (UV-Vis), fluorescence, and circular dichroism spectroscopy for a wide range of applications, including analytical chemistry, complex reaction kinetics, electrochemistry, and photochemistry.1-6 It is safe to state that 2D correlation spectroscopy has established itself as one of the techniques of choice for the characterization of complex systems, including polymers, liquid crystals, proteins, etc. In general, 2D correlation analyzes systematic patterns of subtle spectral changes induced by an external perturbation that are not readily noticeable in conventional 1D spectra. Bands are spread over a second spectral dimension, thereby enhancing spectral resolution, simplifying complex spectra consisting of many overlapped peaks, establishing unambiguous assignments of bands selectively coupled by various interaction mechanisms, and probing the specific order of the spectral intensity changes by an external perturbation. Furthermore, the use of hetero-spectral correlation as a part of 2D analysis provides a convenient means to investigate the relationship between dynamic variations observed by two different spectroscopic techniques.7-10

There have recently been some new and exciting developments in the mathematical treatments concerning the field of 2D correlation analysis, especially those related to the multivariate statistical theory.11-15 For example, the sample–sample correlation spectroscopy,13-15 which has two independent sample axes defining the 2D correlation coordinates instead of the more traditional wavenumber axes, provides information about the perturbation-dependent dynamics of individual species sharing similar spectral signatures. The conventional 2D correlation (i.e., wavenumber–wavenumber correlation) and sample–sample correlation analyses provide complementary information for the study of complex chemical systems.

Chemometrics is already very popular in spectral analysis in many fields of spectroscopic study.16-19 It is generally defined as the chemical discipline that uses mathematical and statistical methods to design or select optimal measurement procedures and experiments and to provide maximum chemical information by analyzing chemical data.16-19 It is interesting to compare the similarities and differences between techniques used in chemometrics and 2D correlation spectroscopy. While they both share similar mathematical operations based on linear algebraic manipulation of data matrices, each focuses on somewhat different aspects of the spectral data structure. For example, chemometrics techniques usually treat each spectrum as a whole entity defined within some chosen spectral region and often represent it as a linear combination of a set of representative loading vectors. In contrast, 2D correlation tends to primarily focus on the dynamics of local features of individual spectra, such as individual peaks and spectral bands, more so than the whole spectral representation in a vector space.

Several papers have recently appeared to explore the possible combination of 2D correlation spectroscopy and chemometrics to provide potentially promising methods applicable in many areas of spectroscopic study.20-25 A noticeable trend among these papers is that instead of forcing 2D correlation to become a fully quantitative technique like a traditional chemometrics approach, they tend to use 2D correlation analysis in conjunction with other well-established techniques to extract the desired information. In general, most workers up to now have focused their attention on the parallel or sequential use of 2D correlation analysis and chemometrics as complementary but essentially independent data analysis techniques. Thus, spectral data are analyzed by a technique such as principal component analysis (PCA) and then a separate analysis is carried out using 2D correlation spectroscopy. The results from different methods are then compared to each other. Similar conclusions are often drawn from different approaches to the spectral data analysis.20-25

Some direct comparisons among outcome features of
chemometrics and 2D correlation spectroscopy have also been made.\textsuperscript{20,25} For example, it has been pointed out that the autopower spectrum located at the diagonal position of a synchronous 2D correlation spectrum resembles the first PCA loading vector of mean-centered data.\textsuperscript{20,21} Likewise, some similarity was noted between certain slices of asynchronous 2D correlation spectra and the second loading vector of mean-centered data.\textsuperscript{20,21} There are also some more detailed discussions about the similarity and difference between the two approaches, but the extent of discussion has been limited to the analysis of synchronous spectra.\textsuperscript{20,21,24}

It should be pointed out that true comparison between 2D correlation spectroscopy and chemometrics must eventually delve into the concept of asynchronicity, which is so central to 2D correlation analysis. Among all the chemometrics techniques, evolving factor analysis (EFA)\textsuperscript{26–28} may carry the closest conceptual commonality with 2D correlation analysis. Both techniques place key emphasis on the order and sequence of spectral data collection. However, there still exist significant differences in the fundamental approach to exploiting the sequentially ordered spectral data set. Future development of chemometrics techniques capable of sorting out the sequential order of spectral signal changes would certainly bring the two fields conceptually much closer.

In the classical chemometrics approaches, such as PCA and factor analysis, one decomposes the original spectral data matrix into a product of score and loading vector matrices.\textsuperscript{16–19} The loading vectors are usually separated into two categories: significant loading vectors representing the linear combinations of spectral contributions of actual components and the noise or insignificant vectors. The fundamental hypothesis of factor analysis is that the improved proxy of the original data matrix can be reconstructed from a limited number of significant loading vectors and scores. There are three direct practical outcomes of factor analysis in the classical chemometrics: (1) the determination of the number of components (i.e., number of significant factors) involved in the description of the data matrix; (2) rejection of noise by discarding the insignificant factors presumably consisting exclusively of noise; and (3) reduction of the information into a compact set of factors.

In this contribution, we will introduce our first attempt at developing the fully integrated true combination of the two powerful spectral data analysis techniques. In this approach, PCA treatment of the spectral data set now becomes an essential and integral part of the subsequent 2D correlation analysis. We will exploit the most obvious advantage of the simple PCA-based abstract factor analysis, i.e., effective identification and rejection of noise components from a raw spectral data set, for subsequent 2D correlation analysis. This feature is most interesting for asynchronous 2D correlation analysis. An asynchronous 2D correlation spectrum is often contaminated by artifactual peaks attributed to the fortuitous correlation of noise. If instead of the raw data matrix, the reconstructed data matrix obtained only from the significant factors for 2D correlation analysis is used, we should be able to essentially eliminate this problem. Although conceptually straightforward, such an attempt to carry out the 2D correlation analysis by using the PCA-reconstructed data matrix (the product of significant or selected scores and loading vectors) has never been reported.

**THEORY**

Let us set the raw data matrix $A$, made up of the original set of perturbation-dependent spectra, to be an $n \times m$ matrix with $n$ spectra and $m$ wavenumber points. The loading vector matrix $V$ is an $m \times r$ matrix, where each column is the loading (i.e., the eigenvector of the dispersion matrix $A^T A$) obtained by PCA. Here, $A^*$ stands for the transpose of $A$. The total number of loading vectors $r$ selected for the analysis must be less than or equal to $n$. It is customary to normalize each column of $V$ (i.e., each loading), such that the product $V^T V$ is an identity matrix. Associated with the PCA loading vectors are the scores (sometime called latent variables). Score matrix $W$ is a relatively small $n \times r$ matrix.

The fundamental idea of factor analysis based on PCA is that the significant part of the data matrix can be expressed as the product of score and loading vectors matrices:

$$A = WV^* + E = A^* + E$$ (1)

where $E$ is the residual matrix often associated with pure noise. Thus, the matrix product $A^*$ can be regarded as the noise-free reconstructed data matrix of the raw data matrix $A$.

$$A^* = WV^*$$ (2)

This reconstructed data set represented in the new matrix $A^*$ may be used instead of the raw data $A$ for the 2D correlation analysis to minimize the effect of noise. A somewhat similar concept of noise reduction by using PCA as a simple filter to reconstruct less noisy data was mentioned by Gillette and Koenig.\textsuperscript{29}

If the total number of significant factors is one (i.e., rank one data matrix), we are dealing with a trivial case of a collection of spectral data sets essentially comprising the same spectrum component with some noise. Such a data matrix will yield a strictly synchronous response if 2D correlation analysis is applied. All the spectral features will vary in a well-coordinated fashion such that there will be no discernible difference in the sequential order of changes (i.e., asynchronicity) among detectable spectral intensities. Typically, however, the number of factors needed to describe a spectral data set will be more than one. Most interestingly, any spectral data set having the matrix rank of three or more, and thus requiring more than two loading vectors for description, will automatically have an asynchronous 2D correlation spectrum with significant features. Any strictly synchronous system, on the other hand, can be adequately and completely described by the linear combination of only two independent spectra. In the case of mean-centered data constructed in conjunction with an additional reference (i.e., average) spectrum, the system can be described by one less number of eigenvectors.\textsuperscript{30} In short, the matrix rank of synchronously varying data simply reduces to one for mean-centered data. Thus, the number of significant eigenvalues (or number of principal components) deduced from classical chemometrics analysis will immediately tell us about the presence of asynchronicity and thus the
potential benefit of constructing the asynchronous 2D correlation spectrum.

At the present moment, classical chemometrics analysis cannot tell us much useful information derived from the sequential order of events described individually by different spectral bands, as in the case of 2D correlation analysis. As noted earlier, however, a very important benefit derived from chemometrical analysis combined with 2D correlation is the ability of factor analysis to rationally reject the noise contribution, which often obscures the truly asynchronous spectrum responses of the system. Thus, the quality of asynchronous correlation spectra will be substantially improved by introducing the noise-elimination advantage of PCA to 2D analysis. In other words, the 2D correlation analysis of the PCA-reconstructed data matrix should selectively accentuate the most important features of synchronicity and asynchronicity without being hampered by noise, or sometimes even by the presence of minor and insignificant components.

EXPERIMENTAL

The set of spectra to which we applied the combination of 2D correlation analysis and PCA were synthetic spectra with random noise. The spectra consist of three Lorentzian peaks located at 1070, 1080, and 1100 cm$^{-1}$, respectively, with gradual increase or decrease with three different rates in each step. There are a total of ten such steps. The choice of the wavenumber axis is quite arbitrary, and thus, it can be settled for any other spectral variable.

Prior to PCA calculation, the mean-centering operation was applied to the raw data matrix. To preserve the amplitude information of the variation of spectral intensities, which becomes important later for 2D correlation analysis, other steps commonly used in PCA, such as normalization scaling of data according to the standard deviation, were not carried out. PCA analysis was performed in the Pirouette software (Infometrix Inc.).

Synchronous and asynchronous 2D correlation spectra were calculated using the algorithm based on the numerical method developed by Noda.³ A subroutine named KG2D,† composed with the Array Basic language (GRAMS/386; Galactic Inc., Salem, NH), was employed for the 2D correlation analysis.

All the reconstructed data matrices used in this study were calculated in MATLAB software (Version 6, The Math Works Inc., Natick, MA).

RESULTS AND DISCUSSION

Figure 1 shows the raw synthetic noisy spectra used in the present study. Two bands at 1070 and 1080 cm$^{-1}$ are gradually increasing while a band at 1100 cm$^{-1}$ is gradually decreasing through ten steps. Conventional synchronous and asynchronous 2D correlation spectra constructed from the noisy spectra in Fig. 1 are displayed in Figs. 2a and 2b, respectively. In this study, we will focus our attention on the potential of PCA 2D correlation analysis for efficient noise filtering as applicable to 2D cor-

† The program can be downloaded from the homepage of Prof. Yukihiro Ozaki of Kwansei Gakuin University, Japan (http://science.kwansei.ac.jp/~ozaki/).

³ A subroutine named

![Fig. 1. Synthetic noisy spectra having three Lorentzian peaks with gradual increase or decrease with three different rates in each step. There are a total of 10 such steps.](image1)

![Fig. 2. (a) Conventional synchronous and (b) asynchronous 2D correlation spectra obtained from noisy synthetic spectra. Solid and dashed lines represent positive and negative cross peaks, respectively.](image2)

Volume 56, Number 12, 2002
relation spectroscopy, so no further interpretation of synchronous and asynchronous 2D correlation spectra will be given here.

The raw spectral data set in Fig. 1 was decomposed into the scores and loading vectors by standard PCA analysis. The reconstructed data matrix $A^*$ obtained by Eq. 2 was used instead of the raw spectral data matrix $A$ for the 2D correlation analysis. Figure 3a depicts spectra of the reconstructed data represented in the matrix $A^*$ from loading vectors and scores of PC1, PC2, and PC3. The PCA-reconstructed spectra with the average spectrum added back, $A^* + \bar{A}$, are shown in Fig. 3b, where $\bar{A}$ is the reference spectrum obtained by averaging the raw spectral data over the entire temperature range. A comparison between Figs. 3a and 1 clearly shows the effect of noise rejection by PCA.

Figures 4a and 4b display synchronous and asynchronous 2D correlation spectra, respectively, constructed from the spectra of the reconstructed data matrix $A^*$ in Fig. 3a. The main feature of the 2D correlation spectra is very much like those directly obtained from the original data but with many fewer effects from noise. It is noted that we can successfully obtain noise-free 2D correlation spectra by reconstructing the data matrix from loading vectors and scores.

For the reconstructed data matrix $A^*$, we used only three principal components in Fig. 4. Of course, we can use more PCs, even though the noise level of the additional loading vectors may increase. As the data matrix can be reconstructed from different numbers of loading and score vectors, one may wonder if 2D correlation spectra generated from the data matrix reconstructed from the first three principal components may become substantially different from those reconstructed from four or more PCs. If the physical phenomena studied here can be explained by only three species, the additional principal components will contain only noise contributions. As one increases the number of PCs, the asynchronous 2D correlation spectrum can become noisy. The main feature of the 2D spectra, other than noise effects obtained from the reconstructed data with four or more principal components (not shown), is virtually indistinguishable from those obtained from the reconstructed data matrix.
A* with the first three principal components in Fig. 4. Thus, the spectral data can be faithfully reconstructed with only three principal components. The rest of the PCs are therefore safely truncated without any noticeable loss of information as noise contribution. This feature should be especially useful if the signal-to-noise ratio of the raw spectral data set is poor.

On the other hand, if we decrease the number of PCs below three, more dramatic changes are observed in the resulting 2D correlation spectra. Eventually, the 2D correlation spectra display the feature as if the system were behaving more or less synchronously. The effect of reducing the number of PCs to only two is quite interesting. Figure 5 depicts the 2D correlation spectra constructed from only two principal components, PC1 and PC2. We can clearly create less noisy 2D spectra by simply using a smaller number of PCs. Simultaneously, however, since we are discarding some real information from the raw data, we may experience some loss of sequential information from the asynchronous spectrum. It may actually make it easier to interpret such 2D spectra if one reduces the number of PCs to the minimum. Furthermore, we have proven unambiguously that the asynchronicity exists even for a system made up of only two principal components. This result is consistent with the expectation from the 2D correlation theory, which requires only the existence of nominal nonlinearity in the sample dimension to exhibit detectable asynchronicity.

The most extreme case of PCA-based 2D correlation analysis is demonstrated when the reconstructed data matrix $A^*$ has the rank of only one, corresponding to the reconstruction of the data matrix using only one principal component. Figure 6 displays the 2D correlation spectra constructed from only one principal component, PC1. The asynchronous 2D correlation spectrum (Fig. 6b) with PC1 shows no significant feature at all. It is thus confirmed that at least two orthogonal spectra (loading vectors) are needed to construct a meaningful set of synchronous and asynchronous 2D correlation spectra. This conclusion should not be surprising since the requirement...
is the same as the well-known need for two independent spectra (i.e., in-phase spectrum and quadrature spectrum) in the case of sinusoidally perturbed 2D studies.\textsuperscript{31}

We would like to briefly discuss here the special case of 2D correlation spectra obtained from the reconstructed data matrix with several major PCs intentionally missing. If one creates a data matrix reconstructed from the second and third principal components, for example, the resulting matrix should contain much more detailed information about subtle dynamic sample-to-sample variations, which are no longer obscured by the dominant effect of the very big contribution of the first principal component. The data matrix reconstructed in this way is devoid of the dominant contribution by the first principal component, which is closely related to the variance, i.e., the autospectrum found in the ordinary synchronous 2D correlation map. As expected, we should be able to accentuate a lot more asynchronous features in this way by eliminating the dominant variance contribution from the data matrix. In the following paper,\textsuperscript{32} we will explore in more detail the above cases.

**CONCLUSION**

We have introduced the new and potentially important idea of using a reconstructed data matrix from scores and loading vectors obtained by applying PCA to raw data before 2D correlation analysis. The 2D correlation analysis of such a reconstructed data matrix should more effectively accentuate the most important features of synchronicity and asynchronicity without being hampered by noise or insignificant minor components. For the demonstration of the power of PCA-based 2D (PCA 2D) correlation analysis, we used noisy synthetic data. As expected, the resulting 2D correlation spectra of noisy data, especially the asynchronous 2D correlation spectrum, are full of noise-induced artifacts.

The PCA reconstruction of the data matrix from the significant scores and loading vectors of PCA could successfully remove the noise component. When such a reconstructed data matrix was used instead of a noisy raw data matrix for 2D correlation analysis, we could successfully obtain improved 2D correlation spectra, which are highly noise suppressed but faithfully retain fine spectral features. Thus, PCA 2D correlation analysis enables us to effectively sort out the similarities and differences in the perturbation-induced variations of the spectral intensities, even for noisy data. There is no doubt that we can faithfully reconstruct the key features of 2D correlation spectra, which are made to be highly noise-suppressed, from PC loading vectors and scores. The application and potential of PCA 2D correlation analysis will be discussed in the following paper.\textsuperscript{32}

**ACKNOWLEDGMENTS**

This work was supported in part by the Korea Research Foundation (BK 21 Project). Y. M. Jung also acknowledges support from the Korea Research Foundation (2001-003-D00059).