Principal Component Analysis Based Two-Dimensional (PCA-2D) Correlation Spectroscopy: PCA Denoising for 2D Correlation Spectroscopy

Young Mee Jung

Department of Chemistry, Pohang University of Science and Technology, San 31, Hyojadong, Pohang 790-784, Korea
Received August 7, 2003

Principal component analysis based two-dimensional (PCA-2D) correlation analysis is applied to FTIR spectra of polystyrene/methyl ethyl ketone/toluene solution mixture during the solvent evaporation. Substantial amount of artificial noise were added to the experimental data to demonstrate the practical noise-suppressing benefit of PCA-2D technique. 2D correlation analysis of the reconstructed data matrix from PCA loading vectors and scores successfully extracted only the most important features of synchronicity and asynchronicity without interference from noise or insignificant minor components. 2D correlation spectra constructed with only one principal component yield strictly synchronous response with no discernible a asynchronous features, while those involving at least two or more principal components generated meaningful asynchronous 2D correlation spectra. Deliberate manipulation of the rank of the reconstructed data matrix, by choosing the appropriate number and type of PCs, yields potentially more refined 2D correlation spectra.

Key Words : Two-dimensional (2D) correlation spectroscopy, Principal component analysis (PCA), PCA-2D correlation spectroscopy, Noise filtering effect

Introduction

The promising possibility of the direct combination of two-dimensional (2D) correlation spectroscopy and principal component analysis (PCA) using simulated spectral data has been recently reported by Jung *et al.*[1] I now apply this powerful technique to experimental data to further examine the benefit of this approach. Both generalized 2D correlation spectroscopy[2-4] and PCA[5-7] are very popular in spectral analyses in many fields of study. Generalized 2D correlation spectroscopy has been certainly one of the most sensitive techniques for interpreting perturbation-dependent spectra. However, it has been well recognized in the field that the presence of a high level of noise can complicate the analysis of 2D correlation spectra.

Jung *et al.*[1] have formulated our PCA-based analysis of spectral data by treating the data matrix **A**, *i.e.*, the original set of perturbation-dependent spectra with each row of the matrix corresponding to a spectral trace, as a somewhat arbitrarily determined sum of an informationally significant portion **A*** and a residual portion **E** comprising predominantly noise contributions.

$$\mathbf{A} = \mathbf{W}\,\mathbf{V}' + \mathbf{E} = \mathbf{A}^* + \mathbf{E} \quad (1)$$

The significant part of the data matrix **A*** was expressed as a product of score matrix **W** and loading vector matrix **V**, and thus **A*** could be regarded as the highly noise-suppressed reconstructed data matrix of the original data **A**. Here **V'** stands for the transpose of **V**.

$$\mathbf{A}^* = \mathbf{W}\,\mathbf{V}' \quad (2)$$

For this reason, it has been named **A*** the *reconstructed* data matrix for 2D correlation analysis. In this new way of 2D correlation analysis, the reconstructed data matrix **A*** is utilized for the calculation of 2D correlation spectra instead of the original data matrix. The analysis of this new data matrix, reconstructed strictly from a few selected significant scores and loading vectors of PCA, has been proposed as a useful method to substantially improve the data quality for 2D correlation analysis by focusing the attention more efficiently to the dominant feature of spectral intensity variations.

Furthermore, another very important benefit derived from this generalized PCA-based two-dimensional (PCA-2D) correlation analysis is the ability to rationally and systematically reject the noise distributed within the original data. It is generally recognized that the asynchronous 2D correlation spectrum is often contaminated by artifactual peaks attributed to the fortuitous correlation of noise. The contribution of noise can be suppressed most effectively as the quality and quantity of data increases, *i.e.*, low noise level of the measurement and more spectral traces per experiment. Unfortunately, for very noisy spectra, it is hard to separate the real signals and noise completely in 2D correlation spectra. It is sometimes possible to filter out the noise from 2D correlation maps. There are some reports for schemes capable of removing or reducing the noise-based peaks from asynchronous 2D correlation spectrum.[8-11] However, none of these techniques has provided truly satisfactory results.

Jung *et al.*[1] have proposed a somewhat different approach to the noise reduction in 2D correlation spectra. Instead of using the original data **A** with noise, if a PCA-reconstructed data set comprising only selected few principal component

*Tel: +82-54-279-2776; Fax: +82-54-279-3399; E-mail: ymjung@postech.ac.kr

(PC) factors are represented in the new matrix **A***, as reported previously[1] this noise problem can be essentially eliminated in 2D correlation analysis. The 2D correlation analysis of such a reconstructed data matrix accentuates only the most important features of synchronicity and asynchronicity without being hampered by the noise, or even by minor signal components, if the number of principal components is restricted.

In this report, I will demonstrate the application of PCA-2D correlation analysis to experimental spectra with a finite but unknown level of noise contributions. I will also examine the practical implication of the efficient noise filtering effect applicable to 2D correlation spectroscopy in the interpretation of fine spectral features, as this was not so obvious in our previous simulation study.

Experimental Section

The PCA-2D correlation spectroscopy was applied to the time-dependent FTIR spectra of a mixture of methyl ethyl ketone, deuterated toluene, and polystyrene during the evaporation, which were injected substantial amount of artificial noise to emphasize the practical benefit of this technique not so obvious in the previous simulation study. The detailed system studied and the methods used to collect FTIR spectra have already been described,[3] so only a brief explanation of the experimental conditions is provided here. A three-component solution mixture of PS dissolved in a 50 : 50 blend of MEK and perdeuterated toluene was examined. The initial concentration of the PS was 1.0% (wt/wt). Once the solution mixture was exposed to air, the solvents started to evaporate, and the PS concentration increased. Eventually, both MEK and toluene are removed from the system, so that only PS remains behind. Transient IR data were collected with a Bio-Rad Model 165 FTIR spectrometer. The sample solution mixture was analyzed by depositing it on a horizontal ZnSe attenuated total reflectance (ATR) plate (Bio-Rad HATR accessory). Sets of IR spectra were collected at intervals of 9 s, with each set consisting of eight coadded scans at 4 cm^{-1} resolution. A total of 25 sets of spectra were collected in this manner. After the data collection, a total of 12 consecutive scan sets were chosen to represent the system during solvent evaporation.

Prior to PCA calculation, the mean centering operation was applied to the 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 using the Pirouette software (Infometrix Inc.).

Synchronous and asynchronous 2D IR correlation spectra were calculated using an algorithm based on a numerical method developed by Noda.[4] A subroutine named KG2D[12] composed by using Array Basic language (GRAMS/386; Galactic Inc., Salem, NH) was employed for the 2D correlation analysis. All the reconstructed data matrixes used in this study were calculated using MATLAB software (Version 6, The Math Works Inc.).

Results and Discussion

Figure 1 shows the model spectra, where substantial amount of artificial noise is injected to the raw transient IR spectra of a PS/MEK/toluene solution mixture during the solvent evaporation process, measured every minute over a time range of 0 to 12 min. The spectral region between 1520 and 1320 cm^{-1}, which covers the individual dynamics of three components well, is examined in the present study. A detailed information of these spectra including bands assignment was already given in the previous paper.[3]

Conventional synchronous and asynchronous 2D correlation spectra constructed from these raw spectra in Figure 1 are displayed in Figure 2(a) and (b), respectively. The basic properties and interpretational procedure of synchronous and asynchronous 2D correlation spectra have already been described in more detail previously,[2] so no further in-depth discussion will be given here.

A common problem encountered in interpreting real-world 2D correlation spectra has been the treatment of relatively weak correlation intensities and poorly resolved peak shoulders. It is often difficult to distinguish the true feature of 2D correlation spectra from artifactual correlation intensities generated by the fortuitous matching of noise components sharing somewhat similar time-dependent behavior. Two specific examples of such potentially ambiguous correlation features found in Figure 2 are: (1) a set of very small synchronous and asynchronous peaks appearing

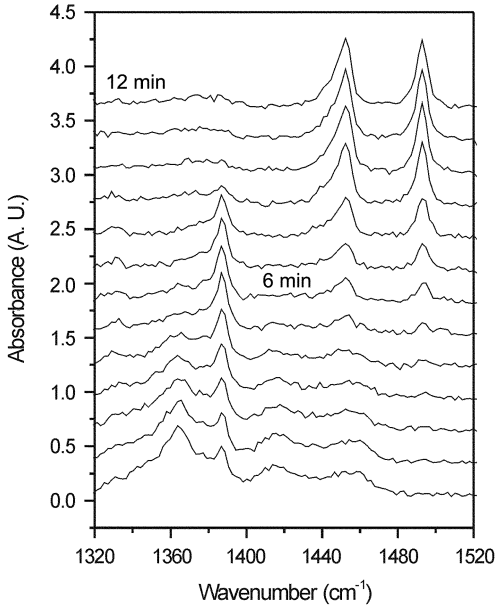

Figure 1. Synthetic noisy spectra which substantial amount of artificial noise is injected to the time-dependent FTIR spectra of a mixture of a PS/MEK/toluene solution mixture during solvent evaporation, measured every minute over a time range of 0 to 12 min, in the region of 1320-1520 cm^{-1}.

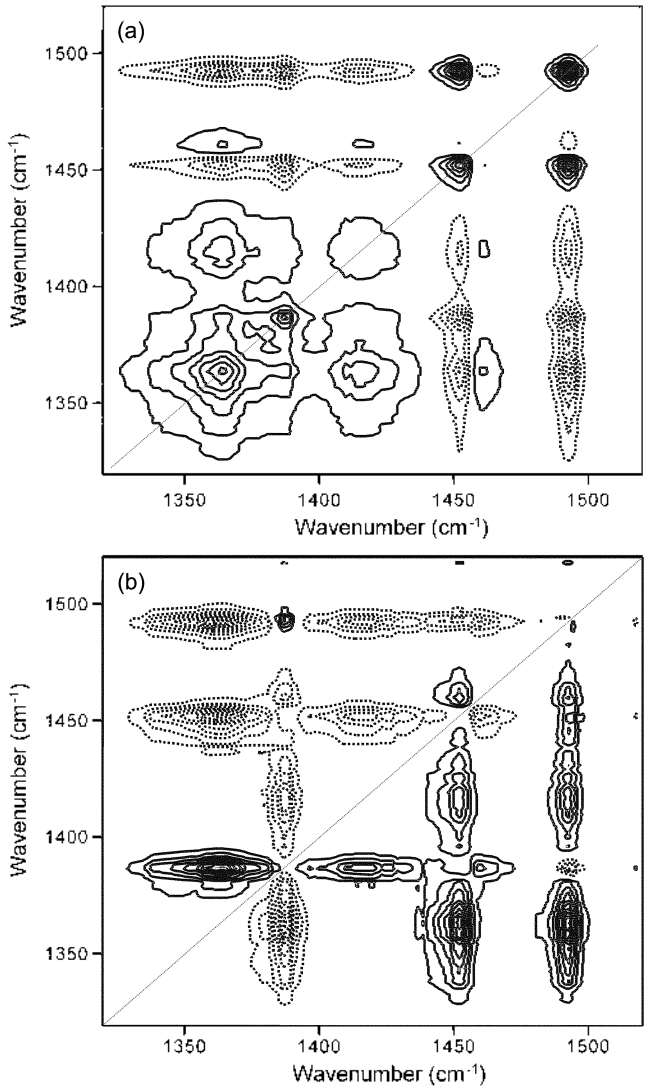

Figure 2. Conventional synchronous (a) and asynchronous (b) 2D correlation spectra obtained from the raw spectra in Figure 1. Solid and dashed lines represent positive and negative cross peaks, respectively.

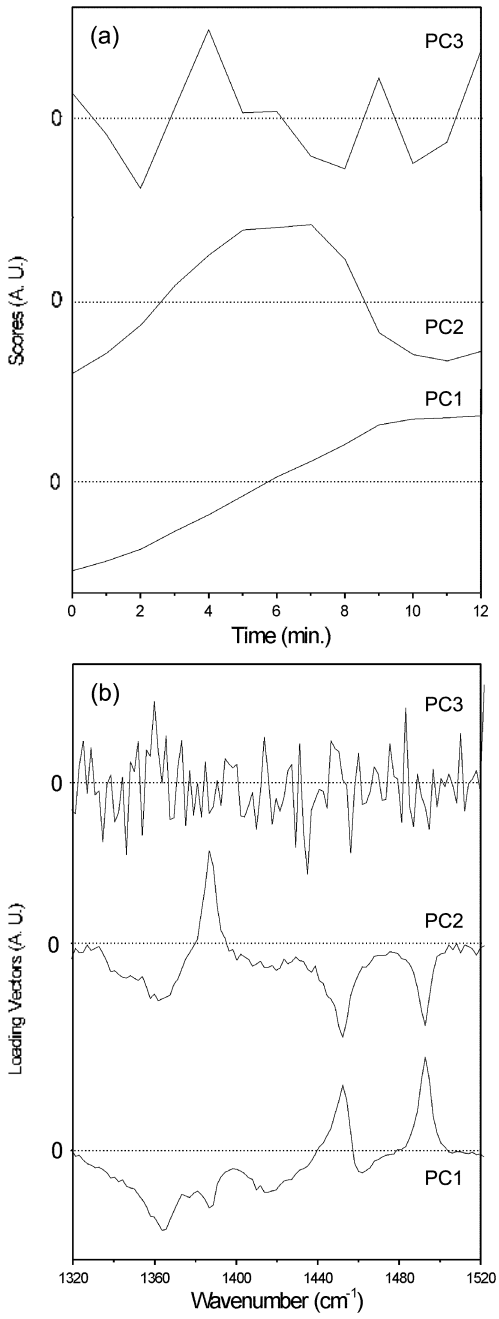

Figure 3. Score plots (a) of the raw spectra in Figure 1 and plots (b) of PC loading vectors for the appropriate scores plotted.

around 1459 cm^{-1} and (2) subtle shoulders observed for some of the synchronous cross peaks and autopeak around 1379 cm^{-1}. Are they real correlation peaks or artifacts generated by the noise contributions? In the past, such questions required subjective judgment of the spectroscopist. The introduction of PCA-2D analysis should provide a more precise answer to such ambiquities, and with a higher level of confidence.

At the outset, the original spectral data set in Figure 1 was decomposed into the scores and loading vectors by standard PCA analysis. The plots of scores and loading vectors of PCs are shown in Figure 3(a) and (b), respectively. PCA factor 1 (PC1), factor 2 (PC2), and factor 3 (PC3) accounts for 83.0%, 16.1%, and 0.2%, respectively, of the total variance of spectral intensities along the time axis. The reconstructed data matrix **A*** obtained by eq. (2) from the three principal components was used instead of the original raw spectral data matrix **A** for the subsequent 2D correlation analysis. Figure 4(a) depicts a set of spectra of the reconstructed data represented in the matrix **A*** from loading vectors and scores of PC1, PC2, and PC3. Figure 4(b) shows the PCA-reconstructed spectra with average spectrum added back allow a meaningful comparison to be made with Figure 1. Even a cursory observation clearly indicates that the reconstructed spectra [Figure 4(b)] are virtually indistinguishable from the original spectra (Figure 1), suggesting that most of the pertinent information is retained in the PCA-reconstructed data.

Figure 5(a) and (b) display a new set of synchronous and asynchronous 2D correlation spectra, generated from the PCA-2D reconstructed spectral data matrix **A*** in Figure 4.

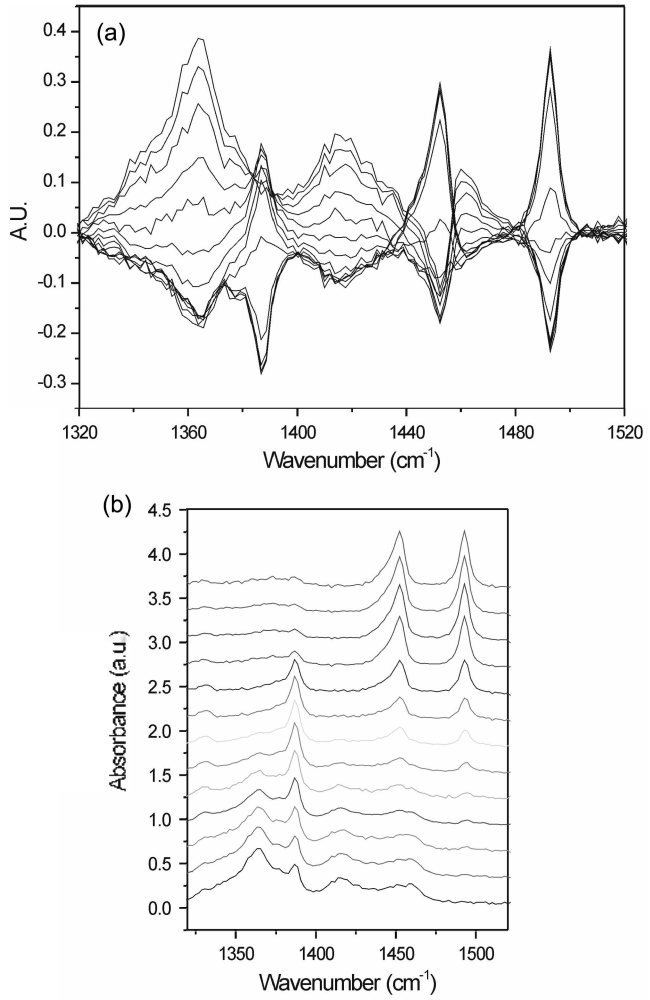

Figure 4. Mean-centered (a) and base-centered (*i.e.*, average added back) (b) spectra of the reconstructed data from loading vectors and scores of PC1, PC2 and PC3.

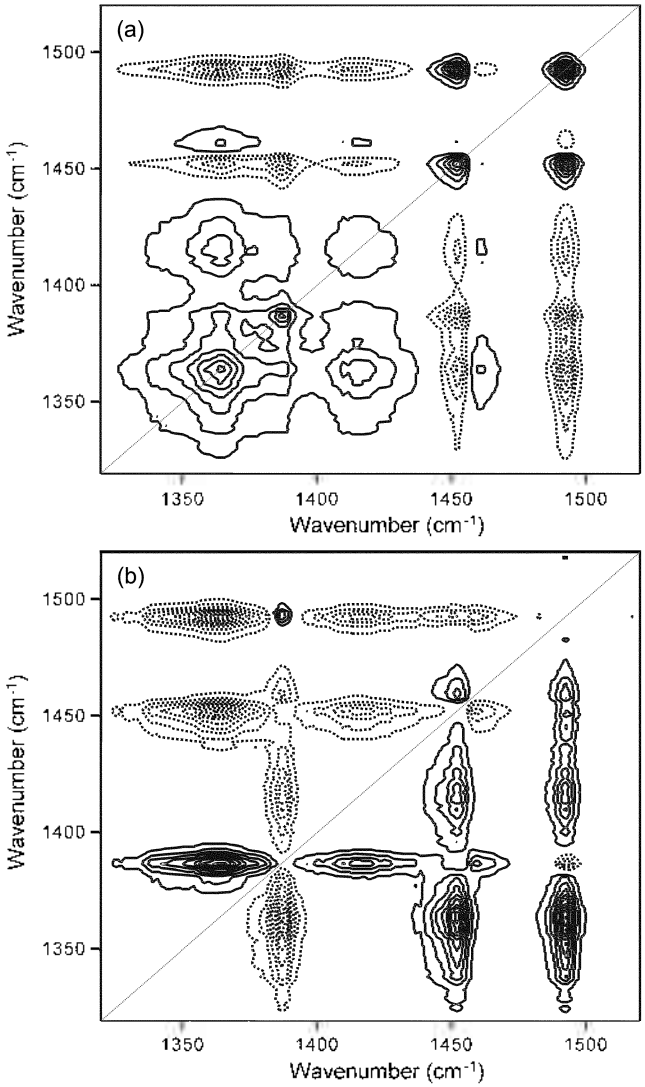

Figure 5. Synchronous (a) and asynchronous (b) 2D correlation spectra obtained using reconstructed data from loading vectors and scores of PC1, PC2 and PC3. Solid and dashed lines represent positive and negative cross peaks, respectively.

These 2D correlation spectra in Figure 5 are remarkably similar to the original 2D correlation spectra in Figure 2, although the details are slightly different. The result indicates that we can successfully truncate the noise component **E** from **A** to calculate **A*** by applying classical PCA treatment to the original data without using a mathematical smoothing filter. Such filtering may distort the spectral features. This result confirms again that the key features of 2D correlation spectra can faithfully reconstructed from PC loading vectors and scores.

Most importantly, finer features appearing in the asynchronous 2D correlation spectrum now can be taken as real manifestations with a much higher level of confidence, since the possibility of noise-induced correlation intensity effect is now substantially eliminated. For example, even small shoulders on 2D peaks around 1383 cm^{-1} can be treated as real spectral features with certain level of confidence. Using the 2D correlation spectra generated from the highly noise suppressed PCA-reconstructed data, the small feature around 1459 cm^{-1} can be unambiguously assigned to MEK band related to other MEK contributions at 1366 and 1417 cm^{-1} and asynchronous correlation with a toluene band at 1389 cm^{-1} and PS bands at 1454 and 1494 cm^{-1}. Likewise, the subtle shoulders appearing around 1379 cm^{-1} are now assigned with reasonable certainty to the MEK contribution, instead of possibly being a noise-induced artifact.

The 2D correlation spectra from the reconstructed data matrix with a smaller number of PCs (*e.g.*, two and one) are displayed in Figures 6 and 7, respectively. As predicted in the previous study,[1] less noisy and somewhat simplified 2D correlation spectra can be created using a smaller number of PCs. Simultaneously, however, since some real information is discarded from the raw data, one may lose some loss of informational content, such as a loss of sequential information from the asynchronous 2D correlation spectrum. For the specific system studied here, most of the key features of the 2D correlation spectra, including the very subtle correlations observed around 1459 and 1379 cm^{-1}, are fully preserved if not enhanced in PCA-2D spectra.

Most importantly, much finer features appearing in

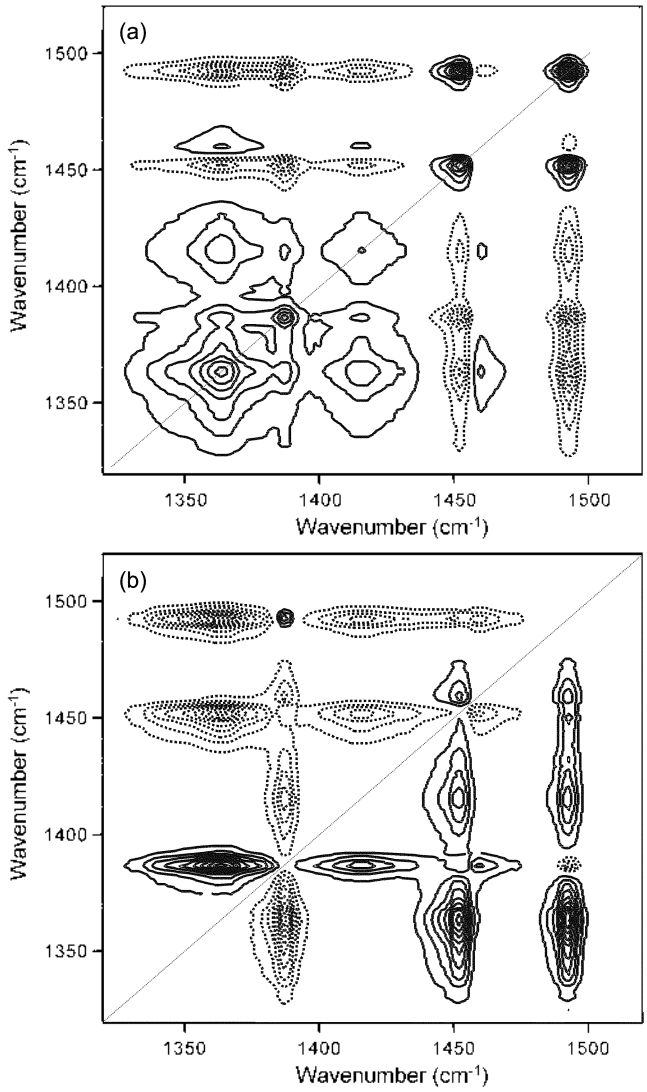

Figure 6. Synchronous (a) and asynchronous (b) 2D correlation spectra obtained using reconstructed data from loading vectors and scores of PC1 and PC2. Solid and dashed lines represent positive and negative cross peaks, respectively.

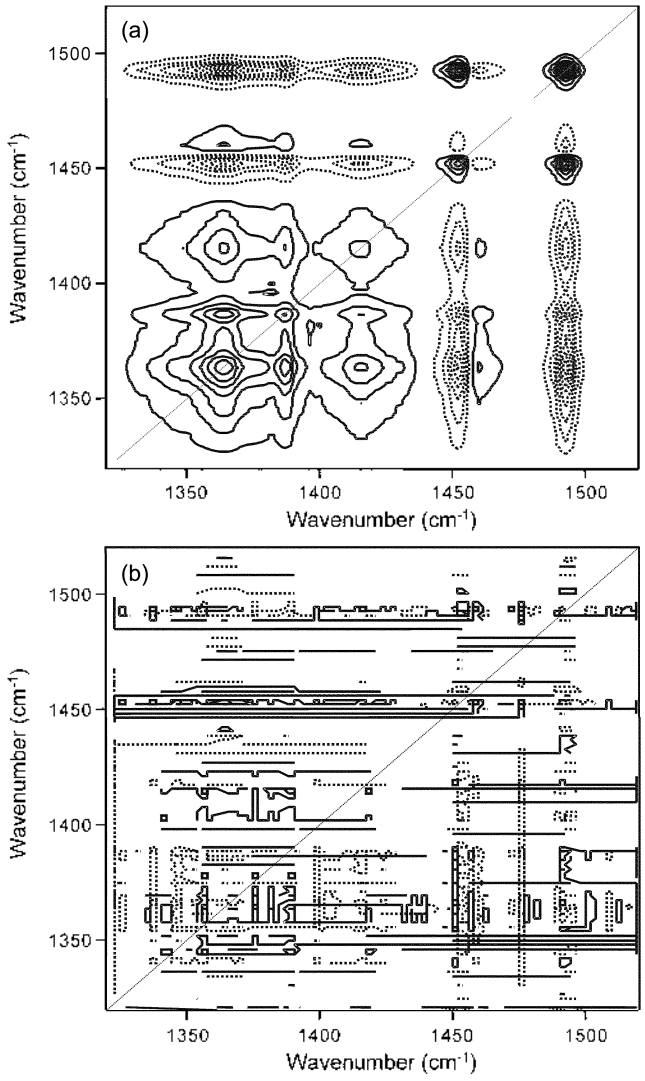

Figure 7. Synchronous (a) and asynchronous (b) 2D correlation spectra obtained using reconstructed data from loading vector and score of only PC1. Solid and dashed lines represent positive and negative cross peaks, respectively.

asynchronous 2D correlation spectra can now be treated as real signals with a reasonable level of confidence, as the possibility of them being noise-induced artifactual correlation effects is now substantially reduced. For example, even small shoulders of 2D peaks around 1387 cm^{-1} can now be analyzed. It turned out that the asynchronous 2D correlation spectrum obtained from the reconstructed data even with only two principal components, PC1 and PC2, [Figure 6(b)] is still very rich in the useful information and shows development of numerous asynchronous peaks. It is still possible unambiguously to distinguish the time-dependent behavior of three separate chemical components from these spectra. For the specific system studied here, most of the key features of 2D correlation spectra, including the very subtle correlation observed around 1459 and 1399 cm^{-1}, are fully preserved if not slightly enhanced. It is important to note that the number of PC factors used in the analysis is not usually the same as the total number of chemical species identified by the specific behavior of band intensity changes. A vary large number of chemical species can be readily differentiated by 2D correlation analysis based on a limited number of principal components, as few as two loadings. This fact has been well demonstrated in the earlier days of 2D IR correlation spectroscopy, where only two orthogonal spectra were used to generate 2D correlation spectra.[13]

On the other hand, the asynchronous 2D correlation spectrum [Figure 7(b)] based on only one factor (PC1) shows only extremely small noise speckles, arising from machine-computational truncation error and random bit noise. As expected, PC1 alone cannot capture the asynchronicity at all. Interestingly, however, the synchronous 2D correlation spectrum [Fig. 7(a)] still retains the main features of the original 2D spectrum [Fig. 2(a)] although fine details are clearly lost.

Conclusion

This study demonstrates the successful application of PCA-2D correlation spectroscopy to real experimental spectral data. It has been demonstrated that the key features of 2D correlation spectra can be faithfully reconstructed from a limited number of PC loading vectors and scores. Such spectra are also highly nose suppressed. Thus, PCA-2D correlation analysis enables me effectively to differentiate the similarities and differences in the perturbation-induced variations in spectral intensities, even for noisy data.

This attribute of PCA-2D correlation analysis is especially useful in interpreting complex real-world 2D spectra, where potential presence of artifactual correlation intensities from noise contributions sharing similar time-dependent behavior can limit the extraction of useful information from weak correlation peaks and small shoulders. This approach to noise rejection without using conventional smoothing techniques inspires greater confidence in interpreting subtle features of complex 2D correlation spectra.

These results demonstrate that the synchronous 2D correlation spectrum is relatively insensitive to the use of reconstructed data with a varying number of principal components. However, the asynchronous 2D correlation spectrum is more sensitive to this effect. The total number of PC factors used in the PCA-2D analysis will not limit the number of chemical components distinguished by the 2D correlation analysis, as long as at least two PCs are involved. If the 2D correlation spectra are constructed with only one factor, the asynchronous 2D correlation spectrum vanishes. In other words, at least two orthogonal spectra (loadings) are needed to construct a meaningful set of 2D correlation spectra.

Acknowledgment. Author thanks Dr. Isao Noda for valuable discussions. This work was supported by the Korea Research Foundation (2001-015-CP0164).

References

1. Jung, Y. M.; Shin, S. H.; Kim, S. B.; Noda, I. *Appl. Spectrosc.* **2002**, *56*, 1562.
2. Noda, I. *Appl. Spectrosc.* **1993**, *47*, 1329.
3. Noda, I.; Dowrey, A. E.; Marcott, C.; Story, G. M.; Ozaki, Y. *Appl. Spectrosc.* **2000**, *54*, 236A.
4. Noda, I. *Appl. Spectrosc.* **2000**, *54*, 994.
5. Malinowski, E. R. *Factor Analysis in Chemistry*; 2nd Ed.; Wiley: New York, 1991.
6. Martens, H.; Næs, T. *Multivariate Calibration*; John Wiley & Sons: New York, 1991.
7. Vandeginste, B. G. M.; Massart, D. L.; Buydens, L. M. C.; De Jong, S.; Lewi, P. J.; Smeyers-Verbeke, J. *Handbook of Chemometrics and Qualimetrics. Part B*; Elsevier Science, B. V.: Amsterdam, The Netherlands, 1998; pp 88-104.
8. Buchet, R.; Wu, Y.; Lachenal, G.; Raimbault, C.; Ozaki, Y. *Appl. Spectrosc.* **2001**, *55*, 155.
9. Tandler, P. J.; Harrington, P. B.; Richardson, H. *Anal. Chim. Acta* **1998**, *368*, 45.
10. Czarnecki, M. A. *Appl. Spectrosc.* **1998**, *52*, 1583.
11. Müller, M.; Buchet, R.; Fringeli, U. P. *J. Phys. Chem.* **1996**, *100*, 10810.
12. The program can be downloaded from the homepage of Prof. Yukihiro Ozaki of Kwansei Gakuin University, Japan. (*http://science.kwansei.ac.jp/~ozaki/*).
13. Noda, I. *Appl. Spectrosc.* **1990**, *44*, 550.