Noise Perturbation in Functional Principal Component Analysis Filtering for Two-Dimensional Correlation Spectroscopy: Its Theory and Application to Infrared Spectra of a Poly(3-hydroxybutyrate) Thin Film

Yun Hu,[†] Boyan Li,[†] Harumi Sato,[†] Isao Noda,[‡] and Yukihiro Ozaki*,[†]

Department of Chemistry, School of Science and Technology, and Research Center for Environmental Friendly Polymers, Kwansei-Gakuin University, Gakuen, Sanda 669-1337, Japan, and The Procter & Gamble Company, 8611 Beckett Road, West Chester, Ohio 45069

Received: April 24, 2006; In Final Form: June 21, 2006

A method based on *noise perturbation in functional principal component analysis* (NPFPCA) is for the first time introduced to overcome the noise interference problem in two-dimensional correlation spectroscopy (2D-COS). By the systematic addition of synthetic noise to the dynamic multivariate spectral data, the functional principal component analysis (FPCA) described in this report is able to accurately determine which eigenvectors are representing significant signals instead of noise in the original data. This feature is especially useful for the data reconstruction and noise filtering. Reconstructed data resulted from the smooth eigenvectors can produce much more reliable 2D correlation spectra by removing the correlation artifacts from noise, which in turn enable more accurate interpretation of the spectral variations. The usefulness of this method is demonstrated with a theoretical framework and applications to the 2D correlation analyses of both simulated data and temperature-dependent reflection−absorption infrared spectra of a poly(3-hydroxybutyrate) (PHB) thin film.

1. Introduction

Generalized two-dimensional correlation spectroscopy (2D-COS)[1−4] is based on the correlation analysis of perturbation-induced variations of spectral intensities monitored by an electromagnetic probe, for instance, time series spectra produced by an infrared spectrometer from *n* measurements at *m* different frequencies. These spectra with an intrinsic order of sampling are not a simple disjoint set of multivariate observations commonly dealt with in the classical statistics but in reality a discrete representation of a set of continuously observed functions, formulated as a matrix **X** of size *n* by *m*. 2D-COS analysis is primarily focused on dynamic changes in spectral intensities during the measurement. The analysis is carried out by the determination of a complex cross-correlation function characterizing the relationship among the variables in spectroscopic data and their temporal physicochemical behaviors in so-called synchronous and asynchronous 2D spectral entries.

Despite its utility and popularity in recent decades, especially in vibrational spectroscopy, it has been recognized that there are certain limitations to the use of 2D-COS. Most discussions revolve around *what exactly the calculated correlation spectra represent and how exact they are*.[3−13] The problems of reliability of correlation results can be ascribed to spectral variations with respect to signal-related variations, peak shifting, baseline fluctuations, and noise.[5,13] As peak shift and distortion are intrinsic to the data, they have to be examined prior to correlation procedure and taken into account for accurate interpretation of correlation patterns.[5,6,14−16] Baseline correction methods, such as perturbation-averaged spectrum[6,13] and iterative polynomial fitting baseline correction,[17] and data pretreatments, for example, multiplicative scatter correction (MSC),[18] principal component analysis (PCA)-based strategy[19] and orthogonal signal correction (OSC)[20] have been available for removing or reducing the effects of baseline fluctuations.

Noise is often a major obstacle to the interpretation of 2D correlation patterns, because it introduces artifact peaks, and sometimes even causes peaks to enhance or attenuate. Therefore, several approaches have been suggested to handle the noise influence on the spectroscopic data prior to the correlation analysis, including Fourier filtering,[13] wavelet analysis,[21] smoothing technique[22] and eigenvector reconstruction.[23,24] Fourier filtering uses the cosine/sine function to transform a noisy spectrum forward first and then inversely back to the spectral domain. The spectrum is, thus, denoised in that when the spectrum is taken into the Fourier domain the noise by definition cannot be adequately modeled. However, the applicability of Fourier transform filtering heavily depends on the capability in specifying a level of modeling in the operation. In wavelet analysis, the elimination of small wavelet coefficients related to the noise of variance spectrum allows the data denoising through a basis function. The philosophy behind wavelet analysis is similar to that of PCA, in which the tradeoff of the irrelevant principal components associated with the small eigenvalues allows the partial elimination of the data noise. However, in wavelet analysis the definition of an optimal basis function and the determination of a proper threshold term are problem dependent. Smoothing is a kind of high-frequency filter that eliminates all signals and noise of high-frequency regardless of their amplitude or even the integrity of spectroscopic information. It is, by its nature, not a true denoising technique.

Due to its simplicity and the ease of use, eigenvector reconstruction is another preferred method to remove the noise.

* To whom all correspondence should be addressed. Fax: +81 79 565 9077. E-mail: ozaki@kwansei.ac.jp.
[†] Kwansei-Gakuin University.
[‡] The Procter & Gamble Co..

10.1021/jp062492t CCC: $33.50
Published on Web 09/14/2006

Based on the concept of PCA, it uses a few eigenvectors capturing almost all variances of the original data to reconstruct the data at an appropriate level of modeling. Noise is accordingly removed. It is crucial to find out the appropriate level of modeling, so that the eigenvector reconstruction is accomplished by the number of factors used against the amount of information captured.[21,23] This operation may influence much smaller spectral features on the correlation patterns which 2D-COS tends to highlight. The noise influence on 2D correlation analysis and the eigenvector reconstruction denoising will be further discussed in the following section.

The aim of the present study is to investigate the noise-filtering effect of eigenvector reconstruction for 2D correlation analysis. A theoretical basis is provided for the *noise perturbation in functional principal component analysis* (NPFPCA) technique. This variant form of the functional principal component analysis (FPCA) makes use of the systematic effects of noise, by actually adding synthetic noise to the spectroscopic data, rather than relying on the existing random effects in the data to identify the eigenvectors representing spectral signals. The smooth eigenvectors derived from the FPCA method are then used to model the spectroscopic data for data reconstruction and noise truncation. The resulting data can produce more reliable 2D-COS results and remove or reduce the correlation artifacts from noise, which enables the spectral variations being responsible for the 2D-COS patterns in the concerned system to be interpreted more accurately and easily. 2D correlation analyses of simulated data and reflection−absorption infrared spectra of a poly(3-hydroxybutyrate) (PHB) thin film observed during a melt process are used to demonstrate the utility of this approach.

2. Method

2.1. Two-Dimensional Correlation Spectroscopy. In 2D-COS a systemic perturbation produces complex 2D correlation spectra $\Phi + i\Psi$, which are separated into real and imaginary components.[1−4] The real component Φ gives the information about coincidental or in-phase variations caused by the perturbation, whose synchronous spectrum can be calculated as

$$\Phi(\nu_1,\nu_2) = \frac{1}{n-1}\sum_{j=1}^{n} x_j(\nu_1)\, x_j(\nu_2) \tag{1}$$

where ν_1 and ν_2 denote different frequencies, x_j corresponds to an individual dynamic spectrum, and n is the number of spectral traces. The variations that occur out-of-phase can be approximated with the imaginary component Ψ in terms of asynchronous spectrum, such that

$$\Psi(\nu_1,\nu_2) = \frac{1}{n-1}\sum_{j=1}^{n} x_j(\nu_1)\sum_{k=1}^{n} N_{jk}x_k(\nu_2) \tag{2}$$

In this equation, N_{jk} represents the jth row and kth column element of the discrete Hilbert−Noda transformation matrix $\mathbf{N}$, given by

$$N_{jk} = \begin{cases} 0; & \text{if } j = k \\ \dfrac{1}{\pi(k-j)}; & \text{otherwise} \end{cases}$$

Alternatively, eqs 1 and 2 can be written in the matrix notation,

$$\Phi = \mathbf{X}^{\mathrm{T}}\mathbf{X} \tag{3}$$

and

$$\Psi = \mathbf{X}^{\mathrm{T}}\mathbf{N}\mathbf{X} \tag{4}$$

For which the superscript T implies transpose of a matrix or vector. For simplicity, the degree-of-freedom term $1/(n-1)$ is not included here, because it contributes only as a constant multiplier.

The data matrix $\mathbf{X}$ can be decomposed into two parts, $\mathbf{X} = (\mathbf{X}_{\mathrm{signal}} + \mathbf{X}_{\mathrm{noise}})$, where $\mathbf{X}_{\mathrm{signal}}$ is a noise-free matrix that holds pure signal information of the measured system, and $\mathbf{X}_{\mathrm{noise}}$ is the matrix consisting of noise. Thus, the synchronous spectrum is given by

$$\begin{aligned}\Phi &= (\mathbf{X}_{\mathrm{signal}} + \mathbf{X}_{\mathrm{noise}})^{\mathrm{T}}(\mathbf{X}_{\mathrm{signal}} + \mathbf{X}_{\mathrm{noise}}) \\ &= \mathbf{X}_{\mathrm{signal}}{}^{\mathrm{T}}\mathbf{X}_{\mathrm{signal}} + \mathbf{X}_{\mathrm{noise}}{}^{\mathrm{T}}\mathbf{X}_{\mathrm{signal}} + \\ &\qquad \mathbf{X}_{\mathrm{signal}}{}^{\mathrm{T}}\mathbf{X}_{\mathrm{noise}} + \mathbf{X}_{\mathrm{noise}}{}^{\mathrm{T}}\mathbf{X}_{\mathrm{noise}}\end{aligned} \tag{5}$$

Similarly,

$$\begin{aligned}\Psi &= (\mathbf{X}_{\mathrm{signal}} + \mathbf{X}_{\mathrm{noise}})^{\mathrm{T}}\mathbf{N}(\mathbf{X}_{\mathrm{signal}} + \mathbf{X}_{\mathrm{noise}}) \\ &= \mathbf{X}_{\mathrm{signal}}{}^{\mathrm{T}}\mathbf{N}\mathbf{X}_{\mathrm{signal}} + \mathbf{X}_{\mathrm{noise}}{}^{\mathrm{T}}\mathbf{N}\mathbf{X}_{\mathrm{signal}} + \\ &\qquad \mathbf{X}_{\mathrm{signal}}{}^{\mathrm{T}}\mathbf{N}\mathbf{X}_{\mathrm{noise}} + \mathbf{X}_{\mathrm{noise}}{}^{\mathrm{T}}\mathbf{N}\mathbf{X}_{\mathrm{noise}}\end{aligned} \tag{6}$$

From a mathematical perspective, noise is orthogonal or nearly orthogonal to any signal except itself, and thereby cross-peaks on the synchronous 2D spectrum attributing to $\mathbf{X}_{\mathrm{signal}}{}^{\mathrm{T}}\mathbf{X}_{\mathrm{noise}}$ and $\mathbf{X}_{\mathrm{noise}}{}^{\mathrm{T}}\mathbf{X}_{\mathrm{signal}}$, do not show up because their intensity may be too weak compared with that of $\mathbf{X}_{\mathrm{signal}}{}^{\mathrm{T}}\mathbf{X}_{\mathrm{signal}}$. When the noise level (namely the trace of covariance matrix $\mathbf{X}_{\mathrm{noise}}{}^{\mathrm{T}}\mathbf{X}_{\mathrm{noise}}$) is much smaller compared to the magnitude of signals, i.e., trace$(\mathbf{X}_{\mathrm{noise}}{}^{\mathrm{T}}\mathbf{X}_{\mathrm{noise}}) \ll$ trace$(\mathbf{X}_{\mathrm{signal}}{}^{\mathrm{T}}\mathbf{X}_{\mathrm{signal}})$, the noise has little interference on the synchronous spectrum, and we have $\Phi \approx \mathbf{X}_{\mathrm{signal}}{}^{\mathrm{T}}\mathbf{X}_{\mathrm{signal}}$. On the other hand, the influence of noise on the asynchronous 2D spectrum, especially the portion represented by $\mathbf{X}_{\mathrm{signal}}{}^{\mathrm{T}}\mathbf{N}\mathbf{X}_{\mathrm{noise}}$ and $\mathbf{X}_{\mathrm{noise}}{}^{\mathrm{T}}\mathbf{N}\mathbf{X}_{\mathrm{signal}}$, may become much stronger if the noise level is very high. This is because of the following reasons: (1) on the asynchronous 2D spectrum, only the subtle spectral features that occur out-of-phase by 90° ($\mathbf{X}_{\mathrm{signal}}{}^{\mathrm{T}}\mathbf{N}\mathbf{X}_{\mathrm{signal}}$) tend to be highlighted and (2) because of the broadband nature, noise may very well contain a substantial amount of component which is 90° out-of-phase with the real signal.[4] Thus, an effective truncation of noise from the data is strongly desired for the 2D-COS analysis.

2.2. Principal Component Analysis and Functional Principal Component Analysis. In the eigenvector reconstruction of spectral data based on PCA, the data manipulation involves both the determination of significant eigenvector components related with signals in spectroscopic data and the truncation of eigenvectors merely representing noise. When it comes to the determination of the significant components, there are roughly two cardinal strategies: eigenvalue analysis and eigenvector analysis, as pointed out by Malinowski[25] and Meloun et al.[26] They stated that procedures for determining the number of components using a variety of empirical and statistical methods have a strong relationship with the consistency of the model assumption and the property of real data. In case the knowledge of the instrumental error associated with the experimental data is available, the methods based on eigenvalues such as IND (indictor function),[25] ER (eigenvalue ratio)[25] and RESO (the ratio of eigenvalues calculated by smoothed PCA and those by ordinary PCA)[27] or other indices are recommended. When the

responsive profiles are smooth, the frequency analysis,[28] the morphological approach[29] and the FPCA[27,30] based on eigenvectors should be preferred.

PCA is a popular method in applied statistical work and data analyses.[25,31−34] It can describe the data in an underlying factor analytic manner and provide information about the number of components by using a set of eigenvectors $\mathbf{r}_i$ ($i = 1, 2, ..., n$) which maximizes the objective λ_i ($i = 1, 2, ..., n$),

$$\lambda_i = \frac{\mathbf{r}_i^{\mathrm{T}}\,\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{r}_i}{\mathbf{r}_i^{\mathrm{T}}\,\mathbf{r}_i} \qquad i = 1, 2, ..., n \tag{7}$$

Here, $\mathbf{r}_i$, subject to

$$\mathbf{r}_i^{\mathrm{T}}\,\mathbf{r}_j = \begin{cases} 1; & \text{if } i = j \\ 0; & \text{otherwise} \end{cases}$$

is the ith eigenvector of the dispersion matrix $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ associated with the eigenvalue λ_i. All the eigenvectors are obtained in a numbered order according to the magnitude of eigenvalues $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_n$. This process can be realized by the singular value decomposition (SVD)[25,32] of $\mathbf{X}$.

The aim of PCA is to pick the first p principal components ($\mathbf{t}_i = \mathbf{X}\mathbf{r}_i$, $i = 1, 2, ..., p$) with $p \leq \min(n,m)$ that are relevant to the variables of interest for attacking the problems in the data, such as those of dimensionality, collinearity, baseline variations and substantial noise. The number of components p can be determined by observing the significant eigenvalues obtained. When the data are free from noise, the number of eigenvalues larger than zero is equivalent to p, providing that the spectra of components are linearly independent. That is, $\lambda_{n-p+1} = ... = \lambda_n = 0$. However, as real data usually contain experimental noise or random error, the eigenvalues and eigenvectors sought by PCA are bound to be contaminated by noise. Therefore, the number of eigenvalues different from zero is usually larger than p. The noise contained within the minor components ($\mathbf{t}_i = \mathbf{X}\mathbf{r}_i$, $i = n - p + 1, ..., n$) with the smallest eigenvalue or variance can be extracted or removed from the data by retaining only the first p components. On the other hand, the component of noise that mixes into and is carried within the first p principal component eigenvectors cannot be completely removed from the data.[25,26] The determination of p is not an easy task, although many eigenvalue-based statistical methods have been proposed with varying degrees of success to deal with the problems. If there are minor components with relatively weak signal contributions, or the signal-to-noise ratio of the data set is low, these methods may not perform well. Some methods fail in determining the real p because the eigenvalues for some true components and noise may be in the same order of magnitude.[27,35]

Besides the information of eigenvalues, the information carried out on eigenvectors is also useful for the determination of the component number of the spectroscopic data, as well as the confrontation of the aforementioned problems. The spectral observations $\mathbf{x}_i$ ($i = 1, 2, ..., n$) often appear as a smooth and continuous function of frequencies or wavenumbers (υ) as $\mathbf{x}_i^{\mathrm{T}}(\upsilon) = \mathbf{f}_i^{\mathrm{T}}(\upsilon) + \mathbf{error}_{\mathrm{noise}}{}^{\mathrm{T}}$ ($\upsilon = 1, 2, ..., m$). It is natural to assume the components $\mathbf{r}_i^f(\upsilon)$ ($i = 1, 2, ..., n$) representing spectral data in PCA should also be smooth. If such additional information is effectively utilized, a more accurate determination of the number of components may be achieved. The functional principal component analysis (FPCA),[27,30] based on functional data analysis,[36−38] was introduced to chemistry and chemometrics to deal with the eigenproblem. It is a method of finding an alternative set of eigenvectors involving the idea of smoothness. The FPCA is very similar to ordinary PCA. The main difference between the two lies in the addition of roughness penalty in FPCA. FPCA searches for a set of vector $\mathbf{r}_i^f$($i = 1, 2, ..., n$) by maximizing the objective function $F(\mathbf{r}_i^f)$,

$$F(\mathbf{r}_i^f) = \frac{\mathbf{r}_i^{f\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{r}_i^f}{\mathbf{r}_i^{f\mathrm{T}}(\mathbf{I} + \alpha\mathbf{D}^{\mathrm{T}}\mathbf{D})\mathbf{r}_i^f} \qquad i = 1, 2, ..., n \tag{8}$$

subject to

$$\mathbf{r}_i^{f\mathrm{T}}(\mathbf{I} + \alpha\mathbf{D}^{\mathrm{T}}\mathbf{D})\mathbf{r}_i^f = \begin{cases} 1; & \text{if } i = j \\ 0; & \text{otherwise} \end{cases}$$

Here $\mathbf{r}_i^f$ represents the ith smooth eigenvector, $\mathbf{I}$ denotes an identity matrix, and $\mathbf{D}$ is a second difference operator of size $(m - 2) \times m$. [27,38]

$$\mathbf{D} = \begin{bmatrix} 1 & -2 & 1 & 0 & ... & 0 & 0 & 0 & 0 \\ 0 & 1 & -2 & 1 & ... & 0 & 0 & 0 & 0 \\ ... & ... & ... & ... & ... & ... & ... & ... & ... \\ 0 & 0 & 0 & 0 & ... & 1 & -2 & 1 & 0 \\ 0 & 0 & 0 & 0 & ... & 0 & 1 & -2 & 1 \end{bmatrix}_{(m-2)\times m} \tag{9}$$

The term $\mathbf{r}_i^{f\mathrm{T}}\mathbf{D}^{\mathrm{T}}\mathbf{D}\,\mathbf{r}_i^f$ is defined as the roughness penalty of discrete function $\mathbf{r}_i^f$, and α is a penalty parameter for controlling the tradeoff between the fidelity of measured data and roughness, which can be set subjectively or by cross validation.[37,39]

Providing that the eigenvalue λ_i^f is associated with $\mathbf{r}_i^f$, eq 8 can be changed into the following form,

$$\mathbf{X}^{\mathrm{T}}\mathbf{X}\,\mathbf{r}_i^f = \lambda_i^{f\mathrm{T}}(\mathbf{I} + \alpha\mathbf{D}^{\mathrm{T}}\mathbf{D})\,\mathbf{r}_i^f \qquad i = 1, 2, ..., n \tag{10}$$

Hence, the determination of the smooth eigenvector $\mathbf{r}_i^f$ and the smooth eigenvalue λ_i^f is equivalent to solving a generalized eigenvalue problem.

Let $\mathbf{u}_i$($i = 1, 2, ..., n$) be the ith eigenvector of a symmetric matrix $\mathbf{X}(\mathbf{I} + \alpha\mathbf{D}^{\mathrm{T}}\mathbf{D})^{-1/2}(\mathbf{I} + \alpha\mathbf{D}^{\mathrm{T}}\mathbf{D})^{-1/2}\mathbf{X}^{\mathrm{T}}$, then,

$$\mathbf{r}_i^f = (\mathbf{I} + \alpha\mathbf{D}^{\mathrm{T}}\mathbf{D})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{u}_i \qquad i = 1, 2, ..., n \tag{11}$$

for which the superscript -1 implies the inverse of the symmetric matrix ($\mathbf{I} + \alpha\mathbf{D}^{\mathrm{T}}\mathbf{D}$). The vector $\mathbf{u}_i$ can be easily obtained by implementing SVD on the matrix $\mathbf{X}(\mathbf{I} + \alpha\mathbf{D}^{\mathrm{T}}\mathbf{D})^{-1/2}$. This solution has been documented in refs 27, 30, and 38 and is not repeated here.

2.3. Noise Perturbation in Functional Principal Component Analysis. As a matter of fact, the eigenproblem of eq 10 in FPCA can be regarded as a perturbation of the eigenproblem of eq 2 in ordinary PCA. Chen et al.[27] suggested that the number of components could be determined by observing the ratio change of λ_i^f/λ_i with the variation of the penalty parameter α. However, a small value of α cannot adequately distinguish noise from signals, and a large value of α will make the small eigenvalues representing signals change significantly. Therefore, how to select a good α value is not a trivial problem in that there is a tradeoff balance between curve fitting (undersmoothing) and curve distortion (oversmoothing) in any smoothing process.[39]

Motivated by the above dilemma, the concept of so-called *noise perturbation*, instead of changes in the smooth parameter α, was introduced into the FPCA method.[30] In this scheme, synthetic random noise is systematically added to the original spectroscopic data. The added noise should be small so as not

to influence the model of the original data too much, but sufficient enough to change the structure of the original noise in the data. The degree of roughness of eigenvectors $\mathbf{r}_i^f$ is utilized to determine the significant components in the noise perturbation in functional principal component analysis (NPF-PCA). An index c_i between the eigenvectors ($\mathbf{r}_i$) sought by ordinary PCA and those ($\mathbf{r}_i^f$) by FPCA before and after the noise addition is defined,

$$c_i = \mathbf{r}_i^{\mathrm{T}} \mathbf{r}_i^f \qquad i = 1, 2, ..., n \tag{12}$$

The added synthetic noise will contribute very little to the smooth eigenvectors representing signals, and the resulting c_i ($i = 1, 2, ..., p$) is close to 1. On the other hand, those representing noise will otherwise change dramatically because of the noise addition, thereby c_i ($i = n - p + 1, ..., n$) becomes much smaller than 1.

To make an objective determination of significant components, the Monte Carlo method is adopted for the synthetic noise addition. That is, the noise of the same level is generated by different random seeds and added to the data **X**. Such an experiment is repeated *num* times, and then a statistical parameter of the standard deviation of c_i is calculated as

$$d_i = \frac{1}{\mathrm{num}} \sum_{k=0}^{\mathrm{num}} \left(c_i^k - \frac{1}{\mathrm{num}+1} \sum_{k=0}^{\mathrm{num}} c_i^k \right)^2 \qquad i = 1, 2, ..., n \tag{13}$$

If c_i is close to 1 and remains rather stable in *num* noise additions, the value of the standard deviation d_i approaches zero until $i = p$. When $i = n - p + 1, ..., n$, the values of d_i become very large because c_i ($i = n - p + 1, ..., n$) varies dramatically with the noise addition. As a result, the first p eigenvectors are accurately determined as the significant components representing signals and should be retained for data reconstruction. On the other hand, all the other eigenvectors $\mathbf{r}_i$ ($i = n - p + 1, ..., n$) are cutoff for noise filtering. It is noted that the p smooth eigenvectors ($\mathbf{r}_i^{f0}$, $i = 1, 2, ..., p$) obtained by the FPCA before the synthetic noise addition rather than those ($\mathbf{r}_i$, $i = 1, 2, ..., p$) directly obtained by the ordinary PCA should be used to reconstruct the data **X**, because the latter still has some noise component, especially if the noise level is high. The resulting reconstructed data ($\mathbf{X}_{\mathrm{ev}}$) based on the smooth eigenvectors

$$\mathbf{X}_{\mathrm{ev}} = \mathbf{X}[\mathbf{r}_1^{f0}, \mathbf{r}_2^{f0}, ..., \mathbf{r}_p^{f0}][\mathbf{r}_1^{f0}, \mathbf{r}_2^{f0}, ..., \mathbf{r}_p^{f0}]^{\mathrm{T}} \tag{14}$$

are then used in eqs 3 and 4 for 2D-COS analysis, in which the noise influence on the 2D correlation patterns can be minimized.

3. Experimental and Data Analysis

The performance of the NPFPCA method is evaluated by using both simulated data and real noisy experimental spectra, which were selected to show the effectiveness of noise reduction in the eigenvector reconstruction for 2D-COS analysis. This experiment specifically covers some noise situations difficult to be handled by the ordinary PCA filtering.

MATLAB 5.3 (The MathWorks Inc., Natick, MA) was used to implement the in-house programs for calculations of simulated and real data sets, running on a personal computer with Pentium 1.7 GHz CPU and 512 MB RAM under the Microsoft Windows XP operating system. The 2D-COS software was developed on the Hilbert-Noda transformation as well as the NPFPCA routines in our laboratory.

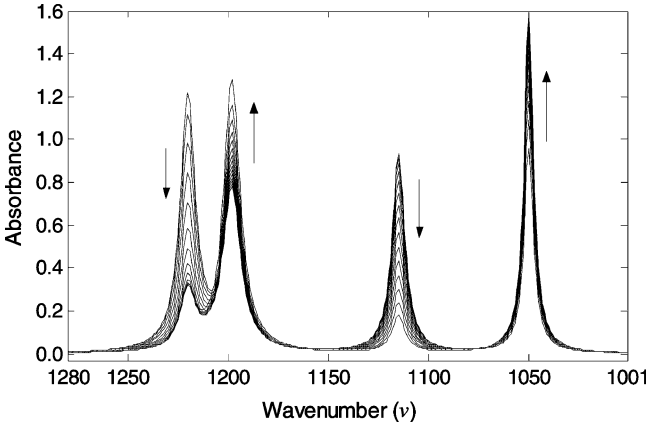

Figure 1. Original noise-free dynamic spectra that were artificially simulated with four Lorentzian bands increasing or decreasing monotonically at different rates.

3.1. Simulated Data. The data set consisted of 45 simulated dynamic spectra. Each spectrum had 280 variables ($\upsilon = 1001, ..., 1280$). The frequency axis was arbitrarily designated for the spectral variables (υ). Four Lorentzian band shapes located at 1050, 1115, 1198, and 1220 data point were created in these noise-free spectra. Figure 1 shows these series noise-free simulated spectra. The intensities of overall bands increased or decreased monotonically at different rates. Normally distributed noise of four different levels was then added to these data series to produce noisy spectra. A Gaussian noise generator was used to randomly generate the noise with zero mean and four different standard deviations of 0.001, 0.005, 0.01, and 0.05 of the maximum band absorbance. To avoid any kind of noise structure, each noise level was repeated 20 times and the averaged noise was used. Another 20-time-averaged noise with zero mean and standard deviation of 0.01 of the maximum band absorbance and a proportionality factor 2 dependent on the absorbance magnitudes were together used to make a heteroscedastic noise addition into individual data series.

3.2. Temperature-Dependent Infrared Spectra of a PHB Thin Film. Bacterially synthesized poly(3-hydroxybutyrate) (PHB) with $M_n = 2.9 \times 10^5$ and $M_w = 6.5 \times 10^5$ was obtained from the Procter & Gamble Co., Cincinnati, OH. The sample was purified by first dissolving in hot chloroform, then precipitating in methanol, and finally drying in a vacuum at 60 °C for 24 h. A thin film of PHB was prepared by the spin coating of an about 1.0 wt % PHB chloroform solution at a speed of 3000 rpm for about 40 s onto an Au-coated glass wafer. The wafer was cleaned in a fresh piranha solution (30% H_2O_2 mixed in a 1:5 ratio with concentrated H_2SO_4) prior to the spin coating. [***Caution:*** Piranha solution reacts violently with organic matter and should be handled with extreme care!] Consequently, the thin film was kept under vacuum at 60 °C for 48 h to completely remove the residual solvent.

Reflection−absorption infrared spectra were recorded by averaging 32 scans at a 2 cm^{-1} resolution with a Thermo Nicolet Magna 470 spectrometer equipped with a MCT detector. The incidence angle was 84°, and the polarization of the incoming beam was parallel to the plane of incidence (p-polarized). The infrared spectra were collected at 2 °C intervals with a heating step rate of 2 °C min^{-1} from 30 to 190 °C using a homemade variable temperature cell. Baseline correction was performed for each spectrum prior to data analysis.

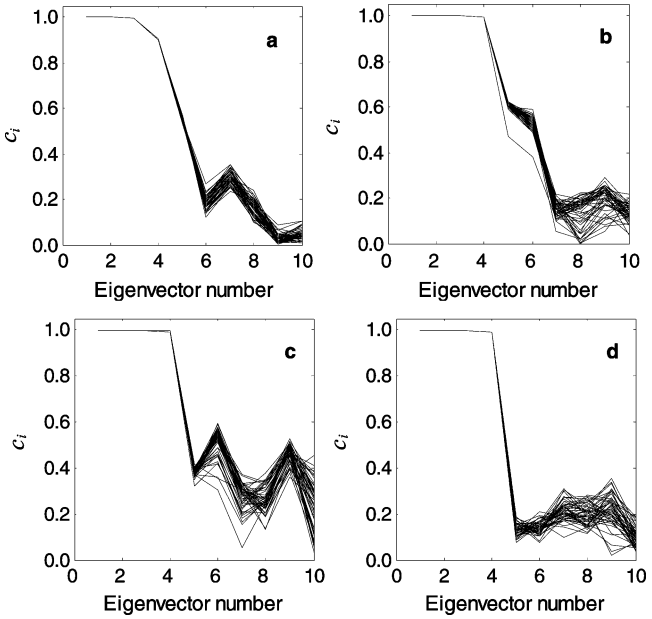

Figure 2. Eigenvector index (c_i) versus eigenvector number for noise perturbation experiments for simulated noisy spectra of four different noise addition levels: (a) 0.05 level; (b) 0.001 level; (c) 0.005 level; (d) 0.01 level.

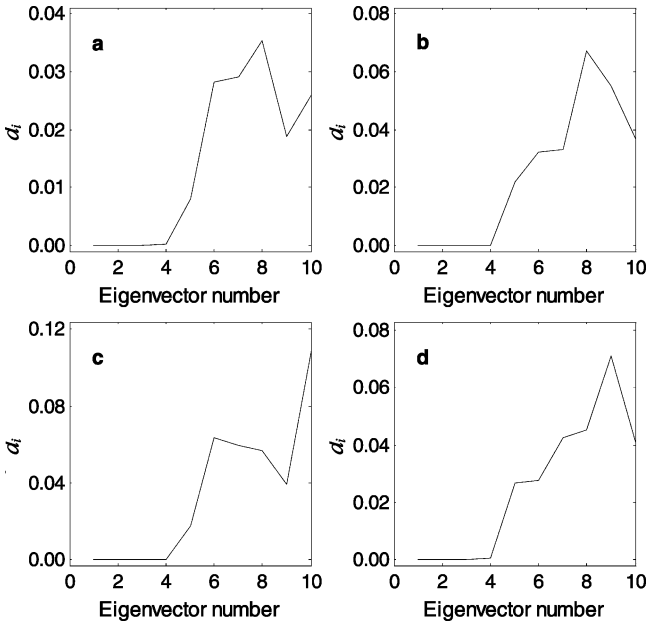

Figure 3. Standard deviation (d_i) of eigenvector index (c_i) versus eigenvector number for noise perturbation experiments for simulated noisy spectra of four different noise addition levels: (a) 0.05 level; (b) 0.001 level; (c) 0.005 level; (d) 0.01 level.

4. Results and Discussion

4.1. Two Key Parameters. Two parameters, α and the noise level for synthetic perturbation addition, are important for the NPFPCA performance and have to be ascertained in advance. Our experience points to the conclusion that the value of the penalty parameter α within the range 1−10 can lead to a satisfactory result, which well balances between curve fitting and curve distortion in the smoothing process. In this study, it was set to 2. Fortunately, the NPFPCA method is not very sensitive to the level of perturbation noise, which is an attractive feature in practice. Whatever the level of noise is added synthetically to the data, it does not influence much the result of the determination of significant eigenvectors for data reconstruction. Of course, too small noise will not provide enough perturbation. One can try different levels of noise to discern how the added noise influences the determination of significant eigenvectors.

TABLE 1: Determination Results of Significant Eigenvectors of the Simulated Spectra with Artificial Noise Addition of Four Different Levels Obtained by Five Indices

noise level	IND	IE	ER	VPVRS	NPFPCA
0.001	6	4	3	4	4
0.005	4	4	3	4	4
0.010	4	4	3	4	4
0.050	3	3	3	3	4

4.2. Simulated Data. To the simulated noise-free dynamic data shown in Figure 1 we added four different levels of noise as described in experimental section. Now, the data with a noise level of 0.05 are taken to illustrate how the NPFPCA procedure determines the significant eigenvectors.

According to eqs 7 and 11, the data matrix **X** with a noise level of 0.05 are first decomposed by ordinary PCA and FPCA using SVD. The first 10 ordinary eigenvectors $\mathbf{r}_i(i = 1, 2, ..., 10)$ obtained by PCA are selected, as well as the first 10 smooth eigenvectors $\mathbf{r}_i^{f0}(i = 1, 2, ..., 10)$ obtained by FPCA in which the penalty parameter is set to $\alpha = 2$. The index c_i between the PCA-based eigenvector $\mathbf{r}_i$ and the smooth eigenvector $\mathbf{r}_i^{f0}$ is consequently calculated with the increase in the eigenvector number from one to ten. Then, synthetic noise with zero mean and standard deviation of 0.001 of the maximum band absorbance, randomly generated by a Gaussian noise generator, is added to the data **X** as a perturbation. It should be pointed out that this synthetic noise may be of different level, and one can manipulate the noise level to achieve an optimal result. The noise-perturbed data are also decomposed by FPCA with $\alpha = 2$, and another 10 smooth eigenvectors $\mathbf{r}_i^f(i = 1, 2, ..., 10)$ are acquired. Likewise, the same levels of noise generated by different random seeds in the Monte Carlo method are added to the data **X**. There are total of 50 experiments, and synthetic noise is added to 49th one (namely, $num = 49$). For each experiment of noise perturbation, the index c_i between the ordinary eigenvectors $\mathbf{r}_i$ and the smooth eigenvectors $\mathbf{r}_i^f$ obtained by FPCA after the synthetic noise addition are calculated versus increasing eigenvector number. Finally, the standard deviation $d_i(i = 1, 2, ..., 10)$ of individual eigenvector index c_i is calculated, as referred to eq 13. We plot all the indices c_i of 50 experiments and their standard deviations d_i versus increasing eigenvector number, as displayed in Figures 2a and 3a. As seen from Figure 2a, the values of c_i are close to 1 as the eigenvector number goes from 1 to 3, whereas a sharp decline appears from 4 to 5. Between the eigenvector number of 3 and 4, the value of c_i declines a little due to heavy noise. In Figure 3a, one can observe a flat line from 1 to 4 (eigenvector number) at which the corresponding d_i remains close to 0, followed by a dramatic change from 4 to 5. This result points to the conclusion that the number of significant eigenvectors representing signals of the noisy spectra equals 4, because the smooth signals are usually unaffected by this perturbation, whereas those eigenvectors representing noise are affected.

In the same way, the numbers of significant eigenvectors representing signals of the other three noisy data sets with smaller noise of 0.001, 0.005 and 0.01 levels are all determined to be 4. The results are also shown in Figures 2 and 3. In these three cases, we kept the conditions of $\alpha = 2$ and the level of

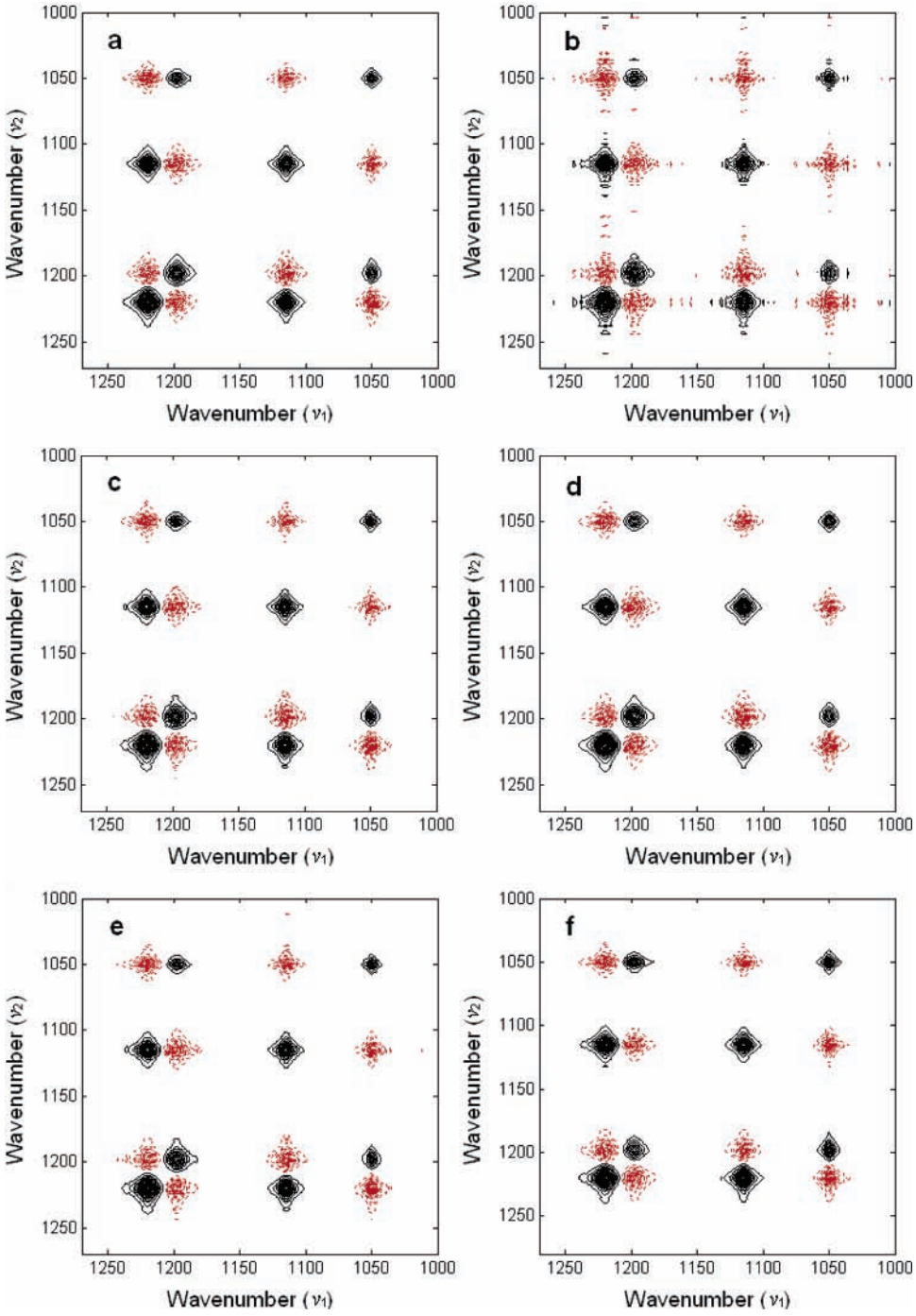

Figure 4. Synchronous 2D-COS spectra resulted from (a) noise-free simulated spectra, (b) directly simulated noisy spectra with the 0.05 level noise added, (c) reconstructed spectra with the first four ordinary eigenvectors obtained by PCA, (d) reconstructed spectra with the first four significant smooth eigenvectors obtained by NPFPCA, (e) reconstructed spectra with the first six ordinary eigenvectors, and (f) reconstructed spectra with the first two ordinary eigenvectors. Solid and dotted lines represent positive- and negative-going peaks.

synthetic noise used for perturbation experiments being 0.001 of the maximum band absorbance. These results all indicate that four significant eigenvectors should be utilized when using eigenvector reconstruction to filter the input noisy data for 2D-COS.

Comparison has been made between the NPFPCA and ordinary PCA for the determination of significant eigenvectors of noisy spectroscopic data used for the eigenvector reconstruction. Four often used indices in the ordinary PCA, i.e., IND, ER, IE (imbedded error)[25] and VPVRS (variance percentage to variance sum),[25,27] are described below to determine significant eigenvectors according to the eigenvalues λ_i ($i = 1, 2, ..., n$) of the covariance matrix $\mathbf{X}^T\mathbf{X}$ obtained by SVD. These indices are calculated as

$$\mathrm{IND}_i = \left(\frac{\sum_{j=i+1}^{n} \lambda_j}{m(n-i)} \right)^{1/2} /(n-i)^2 \qquad i = 1, 2, ..., n-1 \quad (15)$$

$$\mathrm{IE}_i = \left(\frac{i \sum_{j=i+1}^{n} \lambda_j}{nm(n-i)} \right)^{1/2} \qquad i = 1, 2, ..., n-1 \quad (16)$$

$$\mathrm{ER}_i = \frac{\lambda_i}{\lambda_{i+1}} \qquad i = 1, 2, ..., n-1 \tag{17}$$

$$\mathrm{VPVRS}_i = \frac{\lambda_i - \lambda_{i+1}}{\lambda_{i+1} - \lambda_{i+2}} \qquad i = 1, 2, ..., n-2 \tag{18}$$

Here, n and m mean the same as before.

The ability of these five methods to determine the significant eigenvectors of noisy spectra of four different noise levels has been investigated and the results are listed in Table 1. The results show that IND and ER are hardly able to determine the correct significant eigenvectors, even when the interfering noise in the data is small. IE and VPVRS are slightly inferior to NPFPCA in the four cases. However, these two methods fail to determine the significant eigenvectors when the noise level increases to a high degree. NPFPCA works best because it utilizes eigenvector information rather than eigenvalue information. A possible explanation may be that noise does affect the eigenvalues of ordinary PCA but has less influence on the significant eigenvectors. This result suggests that the eigenvalue-based methods are not suitable for determining the significant eigenvectors for the data with heavy noise.

As for 2D correlation, the mean of the series spectra is taken as a reference spectrum to create dynamic spectra. Both

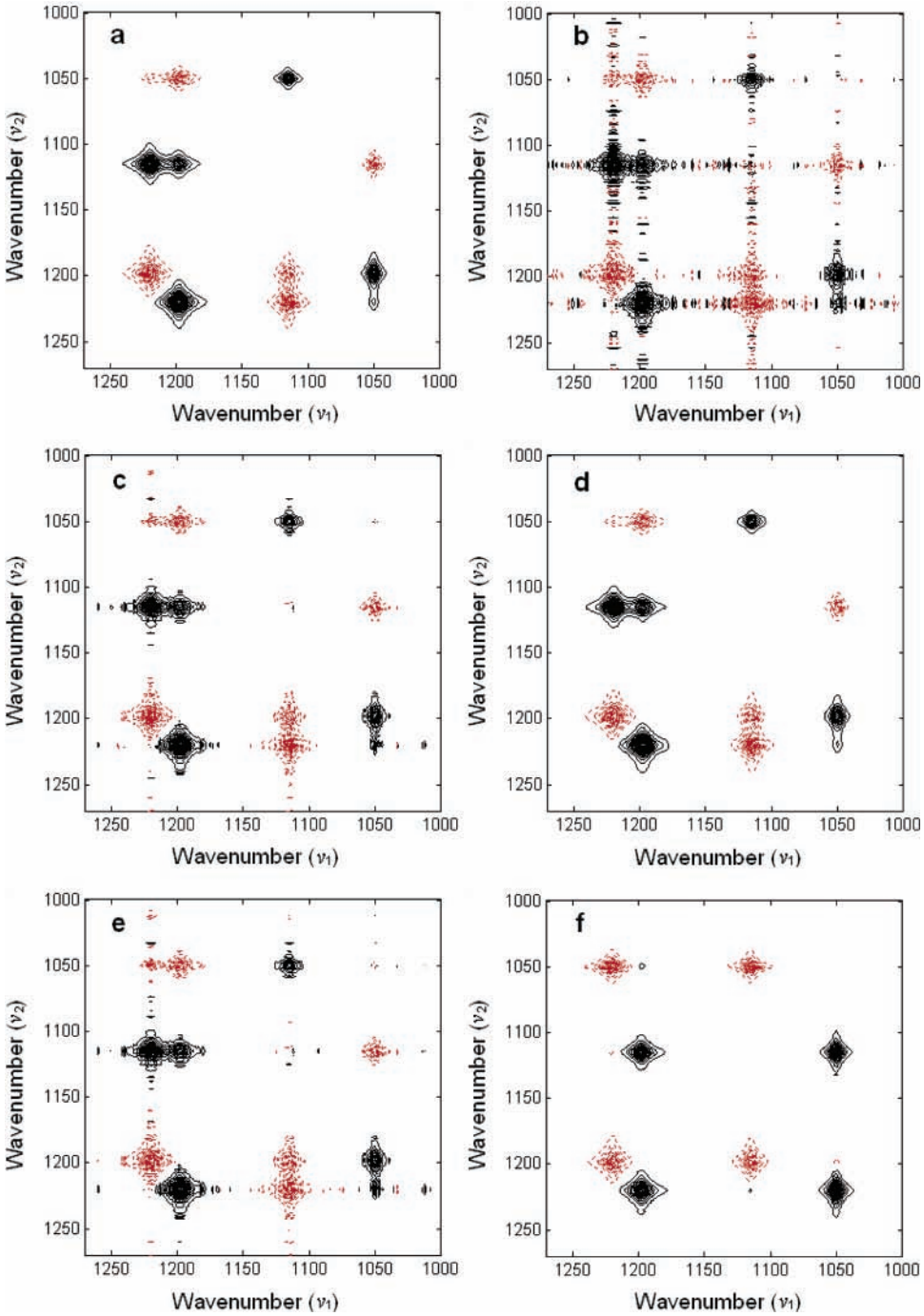

Figure 5. Asynchronous 2D-COS spectra resulted from (a) noise-free simulated spectra, (b) directly simulated noisy spectra with the 0.05 level noise added, (c) reconstructed spectra with the first four ordinary eigenvectors obtained by PCA, (d) reconstructed spectra with the first four significant smooth eigenvectors obtained by NPFPCA, (e) reconstructed spectra with the first six ordinary eigenvectors, and (f) reconstructed spectra with the first two ordinary eigenvectors. Solid and dotted lines represent positive- and negative-going peaks.

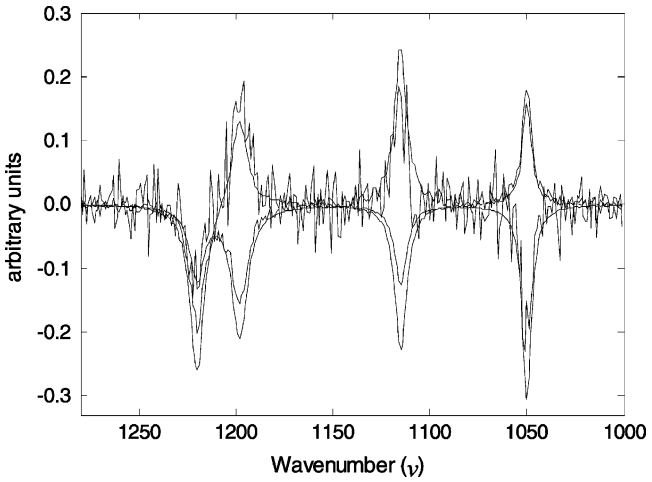

Figure 6. Plot for the first four eigenvectors obtained by ordinary PCA from simulated noisy data of the 0.05 noise level.

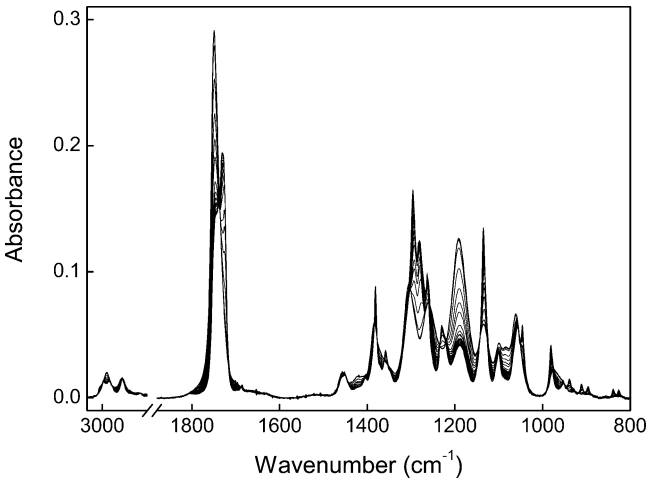

Figure 7. Reflection−absorption infrared spectra of a PHB thin film measured in the temperature range 30−190 °C, increasing at a step rate of 2 °C.

synchronous and asynchronous 2D spectra are calculated for 2D-COS analysis complying with eqs 3 and 4. Figure 4a shows a synchronous 2D correlation spectrum constructed from the original noise-free spectra. Four autopeaks (peaks at diagonal) and six cross-peaks are developed on this 2D contour plot. The positive correlation peaks are plotted solid, and the negative cross-peaks dotted. A positive synchronous cross-peak at position (υ_1, υ_2) indicates that the intensity variations at υ_1 and υ_2 proceed in the same direction, whereas a negative synchronous cross-peak reveals that the changes are in opposite directions. It can be found from this figure that $\Phi(1050, 1115) < 0$, $\Phi(1050, 1220) < 0$, $\Phi(1115, 1198) < 0$, $\Phi(1115, 1220) > 0$, $\Phi(1198, 1050) > 0$, and $\Phi(1198, 1220) < 0$. This result is in accordance with the simulation setting that two bands at 1050 and 1198 data points intensify gradually while the other two bands decrease in intensity. The corresponding asynchronous 2D spectrum is depicted in Figure 5a, where six cross-peaks are present on account of different rates of intensity variations of the four bands. From the sign of these asynchronous cross-peaks together with that of the synchronous cross-peaks, it is deduced that the temporal behaviors of their phase changes occur in the order of 1220 > 1050 > 1115 > 1198, which agrees with the simulated rates. This example reveals the fact that the asynchronous 2D spectrum can particularly reflect some subtle spectral features taking place in dynamic spectra. Hereof, the asynchronous 2D spectrum is susceptible to experimental noise.

Figures 4b and 5b show the synchronous and asynchronous 2D spectra directly derived from the noisy spectra with a noise level of 0.05. The noise introduces many serious artifact peaks and obscures the 2D correlation spectra. In particular, the asynchronous 2D correlation patterns are badly disrupted. As a consequence, it becomes very difficult to discern spectral features on the 2D-COS maps, not to mention reasonably interpreting the temporal behaviors of spectral variations. It is obviously necessary to remove noise from the input data before 2D correlation analysis so as to reduce the noise influence on the 2D correlation spectra as much as possible.

The eigenvector reconstruction is thus considered for filtering noise. Pursuant to the determination of significant eigenvectors of the noisy data $\mathbf{X}$ of 0.05 noise level exemplified previously, four ordinary eigenvectors ($\mathbf{r}_i$, $i = 1, ..., 4$) obtained by PCA and those smooth eigenvectors ($\mathbf{r}_i^{f0}$, $i = 1, ..., 4$) obtained by NPFPCA have undertaken reconstructing the data with $\mathbf{X}_{ev} = \mathbf{X}[\mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3, \mathbf{r}_4][\mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3, \mathbf{r}_4]^T$ and $\mathbf{X}_{ev} = \mathbf{X}[\mathbf{r}_1^{f0}, \mathbf{r}_2^{f0}, \mathbf{r}_3^{f0}, \mathbf{r}_4^{f0}][\mathbf{r}_1^{f0}, \mathbf{r}_2^{f0}, \mathbf{r}_3^{f0}, \mathbf{r}_4^{f0}]^T$, respectively. Afterward, these two reconstructed data are used for the 2D-COS calculation, i.e., $\Phi = \mathbf{X}_{ev}^T\mathbf{X}_{ev}$ and $\Psi = \mathbf{X}_{ev}^T\mathbf{N}\mathbf{X}_{ev}$. Figures 4c and 5c exhibit the synchronous and asynchronous 2D spectra derived from the data reconstructed by the four ordinary eigenvectors. Those from the four smooth eigenvectors are displayed in Figures 4d and 5d. The striking contrast between the 2D correlation spectra presented in Figures 4 and 5 reveals that the eigenvector reconstruction based on NPFPCA has removed the noise influence from the 2D correlation spectra much more effectively, and all the real correlation peaks are recovered clearly. This result strongly supports that the four smooth eigenvectors have rationally extracted the pertinent signal information even from the noisy spectra, which can be used to explain with a higher level of confidence the physical phenomena happening in the dynamic system. Other eigenvectors are regarded as noise representatives and hence truncated for noise filtering. In contrast, the 2D correlation spectra that are resulted from the ordinary PCA eigenvector reconstruction have not been improved so much. The asynchronous spectrum is still unacceptable, and the fringes of peaks in Figure 5c are worse. A possible reason is that the ordinary eigenvectors obtained by PCA still include some noise, which cannot be completely eliminated from the data, as delineated in Figure 6.

If eigenvectors other than significant eigenvectors are reserved for data reconstruction, the reproduced correlation peaks become worse and more spurious peaks appear. Figures 4e and 5e show the synchronous and asynchronous 2D correlation spectra derived from the reconstructed data using six ordinary PCA eigenvectors. It is noted that the information about correlation peaks of the reconstructed data has been contaminated by more noise. On the other hand, an incorrect elimination of significant eigenvectors, for example, only two ordinary PCA eigenvectors used for data reconstruction, may lead to a distinct distortion of the resulting 2D correlation spectra (see Figures 4f and 5f) because some real information about the system behavior is lost. The spectral features on such 2D-COS maps may not be used to determine correct spectral temporal behaviors.

Due to the limitation of space, further discussions of the noise-filtering effect of the NPFPCA and ordinary PCA in relation to the other noisy data of three noise levels 0.001, 0.005, 0.01 are not shown here. In general, the synchronous 2D spectrum is less sensitive to noise. The noise does not influence the asynchronous 2D spectrum much if the noise level is weak.

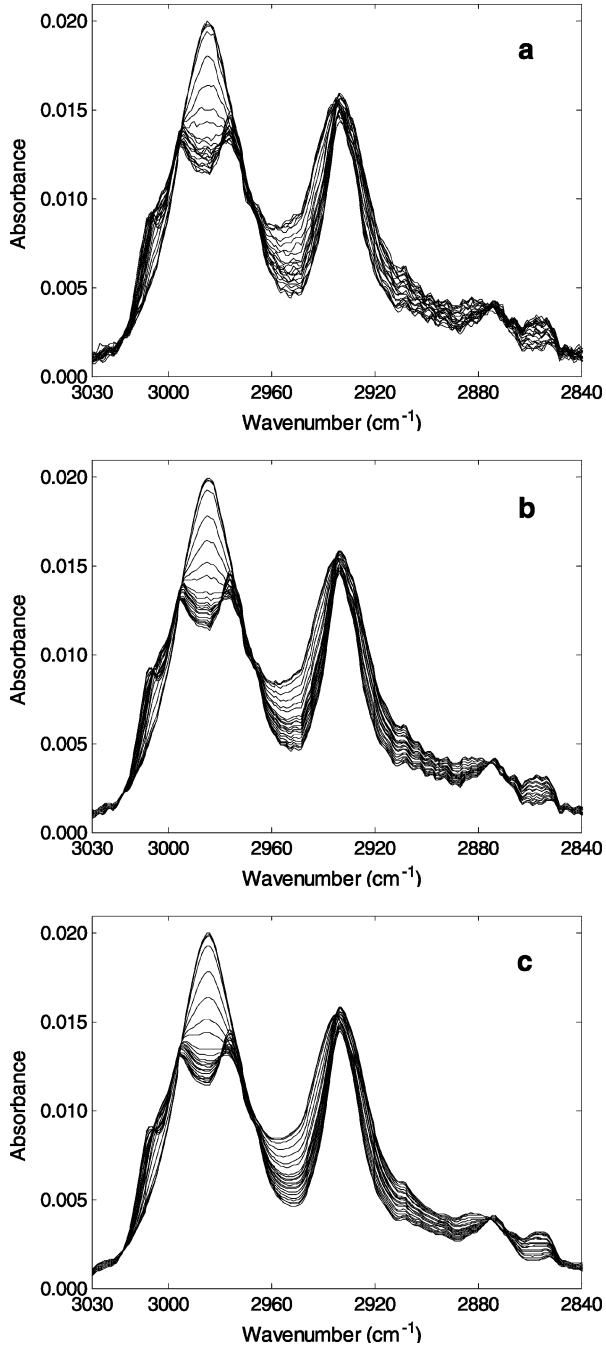

Figure 8. (a) Reflection−absorption infrared spectra in the region 3030−2840 cm^{-1} of a PHB thin film, (b) spectra reconstructed from the first three ordinary eigenvectors obtained by PCA, and (c) spectra reconstructed from the first three significant smooth eigenvectors obtained by NPFPCA

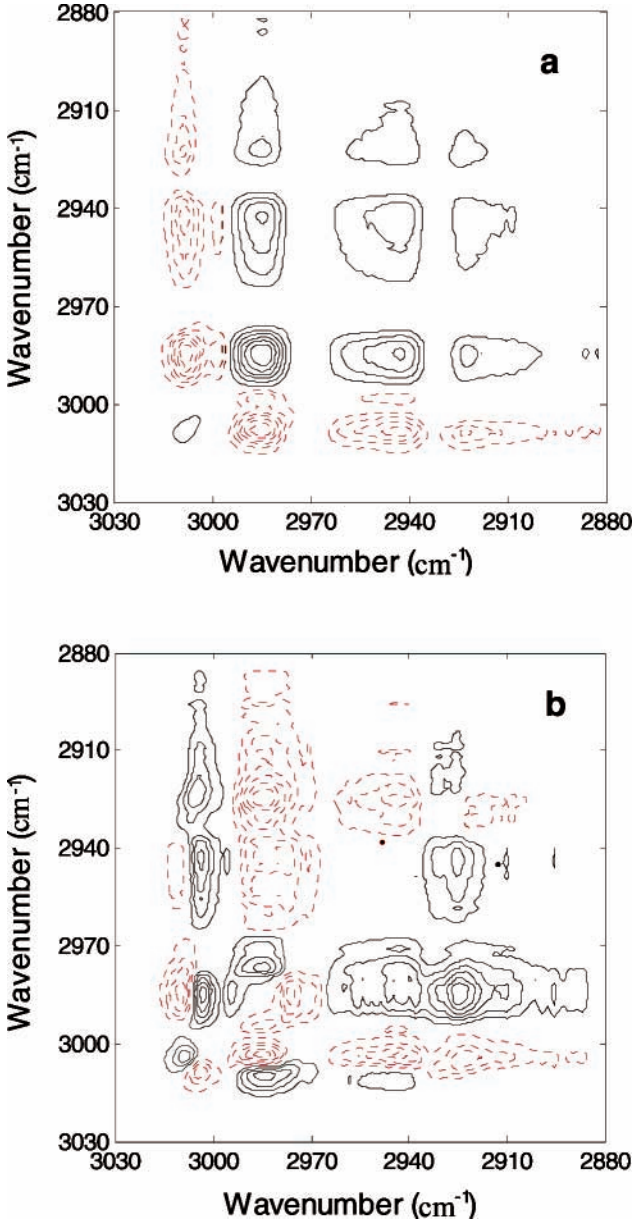

Figure 9. 2D correlation spectra calculated from original reflection−absorption infrared spectra in the region 3030−2880 cm^{-1} of a PHB thin film: (a) synchronous; (b) asynchronous. Solid and dotted lines represent positive- and negative-going peaks.

However, the influence of noise on the asynchronous 2D spectrum becomes serious in the case of heavy noise. It can introduce artifact peaks, cause correlation peaks to change, or sometimes make the asynchronous 2D spectrum useless. By using smooth eigenvectors obtained by NPFPCA to reconstruct the spectroscopic data, it is possible to reduce the serious noise interference on 2D correlation spectra much better than the ordinary PCA.

4.3. Application of Noise Perturbation in Functional Principal Component Analysis to Reflection−Absorption Infrared Spectra of a PHB Thin Film. Thin film technique is very interesting in the studies of physical behavior and mechanical properties of polymers because of its restricted geometric effects.[40,41] Fourier transform infrared spectroscopy is a very sensitive technique to probe submolecular and segmental constituents of polymeric system during the phase transition, but for thin film samples, the signal-to-noise ratio of infrared spectra is not always satisfactory. PHB is one of the thermoplastic polymers of considerable industrial interest due to its biodegradability[42,43] and has been a subject of active scientific investigations.[44−46] Herein, the reflection−absorption infrared spectra of a PHB thin film were chosen for demonstrating the noise-filtering effect of the NPFPCA method for 2D correlation analysis.

Figure 7 shows temperature-dependent reflection−absorption infrared spectra of the PHB thin film. From this figure, it can be seen that distinct spectral variations take place in several regions with temperature. Particular attention should be paid to the spectral changes in four characteristic regions, 3030−2840, 1790−1700, 1340−1150 and 1000−800 cm^{-1}, where bands are ascribed to the C−H, C═O, C−O−C, and C−C stretching

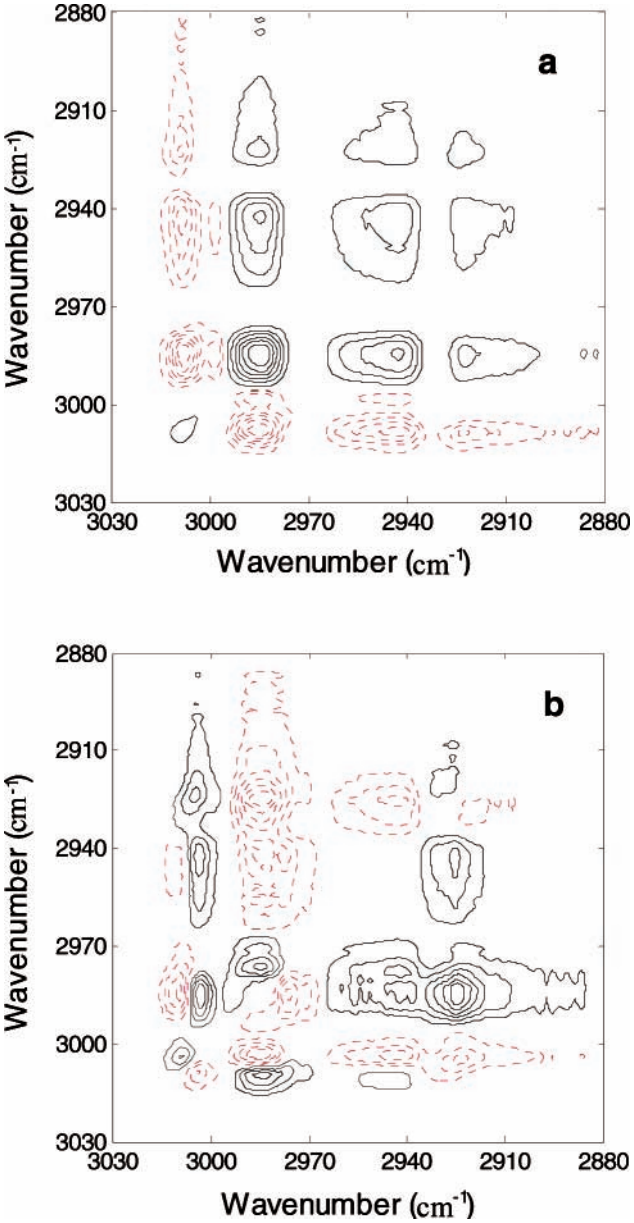

Figure 10. 2D correlation spectra calculated from reconstructed data obtained by three ordinary eigenvectors of PCA: (a) synchronous; (b) asynchronous. Solid and dotted lines represent positive- and negative-going peaks.

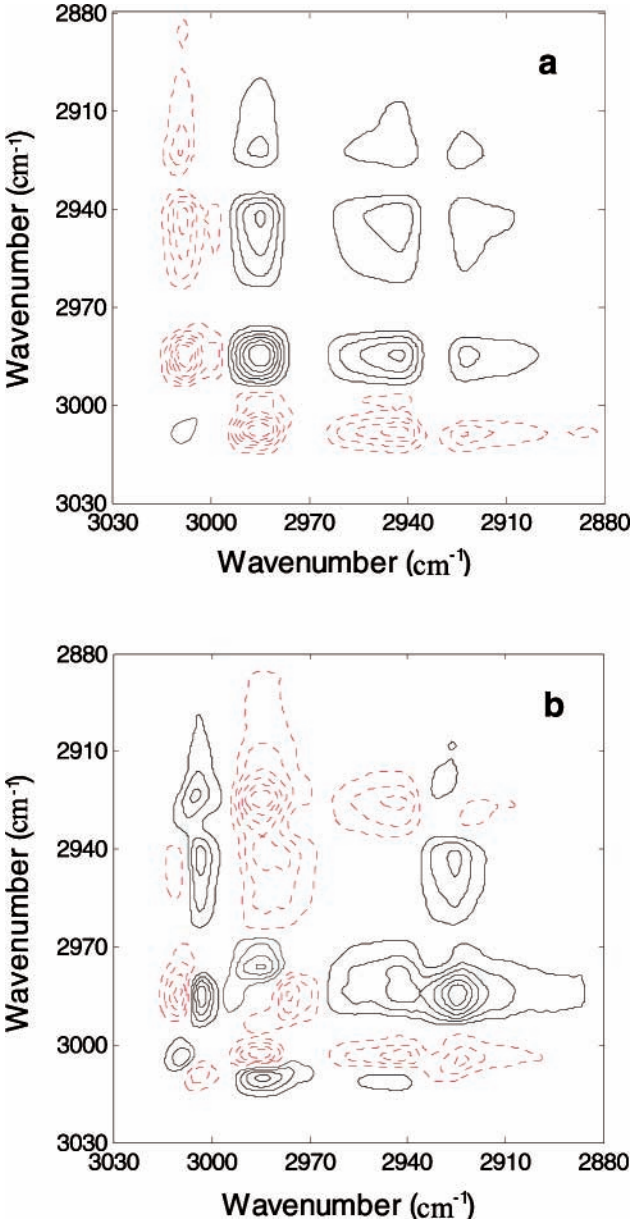

Figure 11. 2D correlation spectra calculated from reconstructed data obtained by three smooth eigenvectors of NPFPCA: (a) synchronous; (b) asynchronous. Solid and dotted lines represent positive- and negative-going peaks.

vibrations reflecting largely the melting behavior of semicrystalline PHB thin film.[44] It is relatively easy for the three regions 1790−1700, 1340−1150 and 1000−800 cm^{-1} to be directly subjected to the 2D correlation analysis, because their responsive intensities are strong enough not to be interfered with by noise. However, in the region 3030−2840 cm^{-1} the situation is different. The original reflection−absorption infrared spectra are very noisy, perhaps due to the low signals, as shown in Figure 8a. These spectra should present a significant challenge to 2D-COS. The NPFPCA method is exploited to filter the data to remove noise as much as possible. The penalty parameter value of $\alpha = 2$ and 49 synthetic perturbations of noise with zero mean and standard deviation of 0.01 of the maximum band absorbance are used for NPFPCA. Three significant smooth eigenvectors $\mathbf{r}_i^{f0}$ ($i = 1-3$) are obtained and used for the data reconstruction and noise filtering. For comparison, the same spectra are reconstructed by three ordinary PCA eigenvectors for noise filtering. Parts b and c of Figure 8 show the resulting data. The data reconstructed from the three smooth eigenvectors obtained by NPFPCA appear rather smooth, as shown in Figure 8c, and look much better than the original spectra in this region (Figure 8a) or those reconstructed from first three ordinary eigenvectors obtained by PCA (Figure 8b). Note that by using NPFPCA the primary spectral features remain intact, while noise is selectively filtered.

The synchronous 2D correlation spectrum generated from the reconstructed spectra obtained by three ordinary PCA eigenvectors is still slightly interfered with by noise. The effect of noise is much worse on the asynchronous 2D correlation spectrum, for the original infrared spectra in the 3030−2840 cm^{-1} region (Figure 9), as well as for the reconstructed data using ordinary PCA (Figure 10). In contrast, there is an obvious beneficial reduction of spectral artifacts or spurious correlation peaks on the 2D-COS maps derived from the reconstructed spectra obtained by three smooth eigenvectors of NPFPCA, as illustrated in Figure 11. Figure 12 shows the comparison of slice

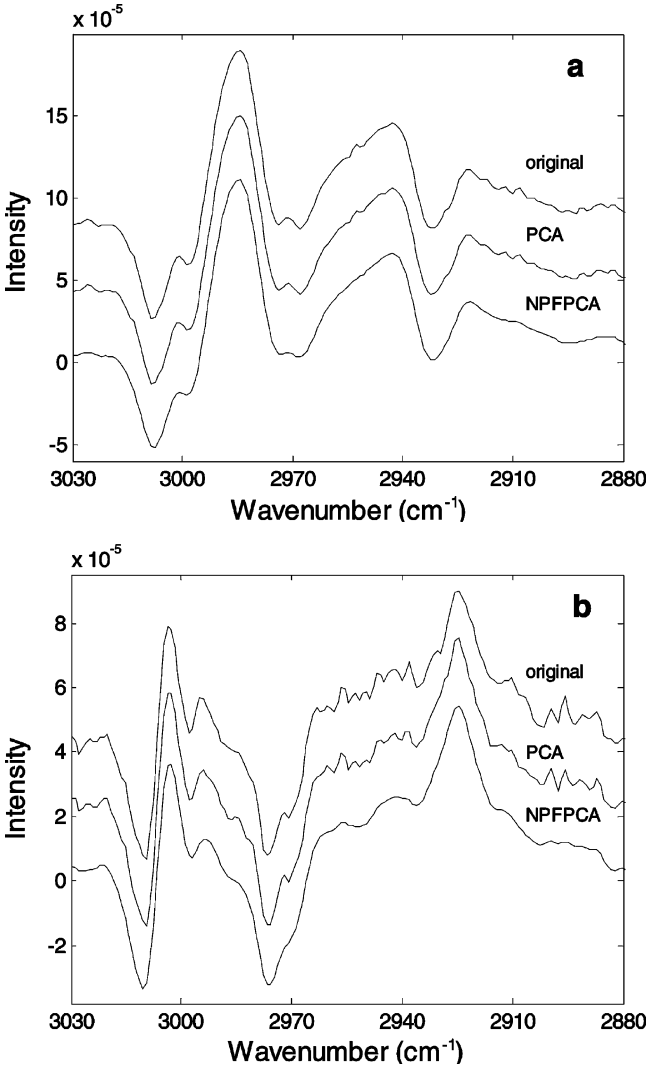

Figure 12. Slice spectra at 2985 cm^{-1} extracted from (a) synchronous and (b) asynchronous 2D correlation spectra obtained from the original spectra in the region 3030−2880 cm^{-1} and the corresponding PCA and NPFPCA eigenvector-reconstructed data, respectively.

spectra at 2985 cm^{-1} for the synchronous and asynchronous 2D correlation spectra resulted from the original spectra and reconstructed data obtained by ordinary and smooth eigenvectors. It is apparent that the slices of correlation spectra from the original spectra and ordinary PCA eigenvector-reconstructed data are almost the same. They both contain a similar level of noise interference, which results in spectral artifacts in the synchronous and asynchronous 2D correlation spectra, especially in the region 2960−2935 cm^{-1} (Figures 9 and 10). It is very hard for the PCA eigenvector reconstruction to filter the noise influence from the spectra in this situation. Not only is the noise influence reduced much from the 2D correlation spectra but also the primary informative features of the input spectra are well preserved. The quality of the correlation spectra resulted from the reconstructed data by three smooth eigenvectors of NPFPCA, in particular the asynchronous 2D spectra, is improved to a legible level. The spectral variations occurring in this region are much easier to interpret according to the correlation patterns.

Four autopeaks are clearly observed at around 3009, 2985, 2945, and 2922 cm^{-1} on the synchronous map. The peak located at 2985 cm^{-1} together with two bands at 2945 and 2922 cm^{-1} share three positive cross-peaks at (2985, 2945), (2985, 2922), and (2945, 2922) cm^{-1}. Also, they share negative cross-peaks with those located at 3009 and 2997 cm^{-1}. These observations suggest that the three bands at 2985, 2945, and 2922 cm^{-1} vary in the same direction, opposite to those at 3009 and 2997 cm^{-1}. As the temperature increases, the former are intensified. Therefore, we conclude that the bands at 2985, 2945, and 2922 cm^{-1} are all associated with the amorphous part of PHB. On the other hand, the intensities of bands at 3009 and 2997 cm^{-1} decrease, indicating they are from the crystalline state of PHB. In the asynchronous 2D spectrum, the amorphous band at 2985 cm^{-1} shares positive cross-peaks with the crystalline bands at 3009 and 2975 cm^{-1}, suggesting that the amorphous phase does not appear simultaneously with the disappearance of the crystalline state. Furthermore, there is no asynchronicity between any of crystalline bands (3009, 2997, and 2975 cm^{-1}) in this spectral region. As a result, it is revealed that the intensity changes of the crystalline bands in this region likely occur simultaneously.

The bands in the 3030−2840 cm^{-1} region are mainly due to the CH_3 stretching, CH_2 stretching and CH stretching modes. The 2985 cm^{-1} band is ascribed to the amorphous part and those at 3009 and 2975 cm^{-1} are from the PHB crystal structure. For PHB, the CH_3 group exists only on the side chain and the CH_2 group is located on the skeletal chain. The unusual appearance of the CH_3 asymmetric stretching band at 3009 cm^{-1} is attributed to the existence of a C−H···O═C hydrogen bonding in this system.[44,45]

5. Conclusion

The present study discusses how the presence of heavy noise influences 2D-COS results. It was demonstrated that the NPFPCA method effectively filters out a significant amount of the noise from the spectral data for more reliable 2D correlation analysis. This method is able to accurately determine the number of significant eigenvectors representing signals. The technique was especially effective for the reflection−absorption infrared spectroscopic data of a very thin polymer film interfered with heavy noise, by providing a set of smooth eigenvectors to reconstruct the data. The resulting reconstructed data free from the noise effect can lead to more reliable 2D correlation spectra, enabling much easier and accurate interpretation.

Acknowledgment. We thank Dr. Chenjian Xu of Central South University, P. R. China, for the valuable discussion on the NPFPCA method. This study was supported by "Open Research Center" Project for Private Universities: matching fund subsidy from MEXT (Ministry of Education, Culture, Sports, Science and Technology), 2001-2005, and by Kwansei Gakuin University Special Research Project "Environment Friendly Polymers", 2004−2008.

References and Notes

(1) Noda, I. *J. Am. Chem. Soc.* **1989**, *111*, 8116.
(2) Noda, I. *Appl. Spectrosc.* **1993**, *47*, 1329.
(3) Ozaki Y.; Noda, I. *Two-Dimensional Correlation Spectroscopy*; American Institute of Physics: New York, 2000.
(4) Noda I.; Ozaki, Y. *Two-Dimensional Correlation Spectroscopy: Applications in Vibrational and Optical Spectroscopy*; John Wiley & Sons: Chichester, U.K., 2004.
(5) Czarnecki, M. A. *Appl. Spectrosc.* **1998**, *52*, 1583.
(6) Harrington, P. de B.; Urbas, A.; Tandler, P. J. *Chemom. Intell. Lab. Syst.* **2000**, *50*, 149.
(7) Meier, R. J.; Steeman, P. A. M. *Appl. Spectrosc.* **2002**, *56*, 401.
(8) Noda, I.; Ozaki, Y. *Appl. Spectrosc.* **2003**, *57*, 110.
(9) Huang, H.; Malkov, S.; Coleman, M. M.; Painter, P. C. *Macromolecules* **2003**, *36*, 8148.
(10) Czarnecki, M. A. *Appl. Spectrosc.* **2003**, *57*, 107.
(11) Czarnecki, M. A. *Appl. Spectrosc.* **2003**, *57*, 991.

(12) Haider, P.; Haller, G. L.; Pfeferle, L.; Ciuparu, D. *Appl. Spectrosc.* **2005**, *59*, 1060.
(13) Tandler, P. J.; Harrington, P. de B.; Richardson, H. *Anal. Chim. Acta* **1998**, *368*, 45.
(14) Huang, H.; Malkov, S.; Coleman, M. M.; Painter, P. C. *Macromolecules* **2003**, *36*, 8156.
(15) Huang, H.; Malkov, S.; Coleman, M. M.; Painter, P. C. *J. Phys. Chem. A* **2003**, *107*, 7697.
(16) Morita, S.; Miura, Y. F.; Sugi, M.; Ozaki, Y. *Chem. Phys. Lett.* **2005**, *402*, 251.
(17) Gan, F.; Ruan, G. H.; Mo, J. Y. *Chemom. Intell. Lab. Syst.* **2006**, *82*, 59.
(18) Czarnik-Matusewicz, B.; Murayama, K.; Tsenkova, R.; Ozaki, Y. *Appl. Spectrosc.* **1999**, *53*, 1582.
(19) Sun, J. *J. Chemom.* **1997**, *11*, 525.
(20) Wold, S.; Antii, H.; Lindgren, F.; Öhman, J. *Chemom. Intell. Lab. Syst.* **1998**, *44*, 175.
(21) Berry, R. J.; Ozaki, Y. *Appl. Spectrosc.* **2002**, *56*, 1462.
(22) Buchet, R.; Wu, Y.; Lachenal, G.; Raimbault, C.; Ozaki, Y. *Appl. Spectrosc.* **2001**, *55*, 155.
(23) Kramer, R. *Chemometric Techniques for Quantitative Analysis*; Marcel Dekker: New York, 1998.
(24) Jung, Y. M.; Shin, H. S.; Kim, S. B.; Noda, I. *Appl. Spectrosc.* **2002**, *56*, 1562.
(25) Malinowski, E. R. *Factor Analysis in Chemistry*, 3rd ed.; Wiley: New York, 2002.
(26) Meloun, M.; Čapek, J.; Mikšík, P.; Brereton, R. G. *Anal. Chim. Acta* **2000**, *423*, 51.
(27) Chen, Z. P.; Liang, Y. Z.; Jiang, J. H.; Li, Y.; Qian, J. Y.; Yu, R. Q. *J. Chemom.* **1999**, *13*, 15.
(28) Rossi, T. M.; Warner, I. M. *Anal. Chem.* **1986**, *58*, 810.
(29) Wang, J. H.; Jiang, J. H.; Xiong, J. F.; Li, Y.; Liang, Y. Z.; Yu, R. Q. *J. Chemom.* **1998**, *12*, 95.
(30) Xu, C. J.; Liang, Y. Z.; Li, Y.; Du, Y. P. *Analyst*, **2003**,*128*, 75.
(31) Anderson, T. W. *An Introduction to Multivariate Statistical Analysis*, 2nd ed.; Wiley: New York, 1984.
(32) Joliffe, I. T. *Principal Component Analysis*; Springer: New York, 1986.
(33) Golub, G.; Van Loan, C. *Matrix Computations*, 2nd ed.; Johns Hopkins University Press: Baltimore, MD, 1989.
(34) Press: W.; Teukolsky, S.; Vetterling, W.; Flannery, B. *Numerical Recipes in C*, 2nd ed.; Cambridge University Press: New York, 1992.
(35) Elbergali, A. K.; Nygren, J.; Kubista, M. *Anal. Chim. Acta* **1999**, *379*, 143.
(36) Ramsay, J. O.; Dalzell, C. J. *J. R. Statist. Soc. B* **1991**, *53*, 539.
(37) Green, P. J.; Silverman, B. W. *Nonparametric Regression and Generalized linear models: A Roughness Penalty Approach*; Chapman and Hall: London, 1994.
(38) Silverman, B. W. *Ann. Statist.* **1996**, *24*, 1.
(39) Liang, Y. Z.; Leung, A. K. M.; Chau, F. T *J. Chemom.* **1999**, *13*, 511.
(40) Shin, H. S.; Jung, Y. M.; Oh, T. Y.; Chang, T.; Kim, S. B.; Lee, D. H.; Noda, I. *Langmuir* **2002**, *18*, 5953.
(41) Zhang, Y.; Zhang, J. M.; Lu, Y. L.; Duan, Y. X.; Yan, S. K.; Shen, D. Y. *Macromolecules* **2004**, *37*, 2532.
(42) Dawes, E. A. *Novel Biodegradable Microbial Polymers*; Kluwer Academic: Dordrecht, The Netherlands, 1990.
(43) Doi, Y.; Fukuda, K. *Biodegradable Plastics and Oilymers*; Elsevier: Amsterdam, 1994.
(44) Sato, H.; Murakami, R.; Padermshoke, A.; Hirose, F.; Senda, K.; Noda, I.; Ozaki, Y. *Macromolecules* **2004**, *37*, 7203.
(45) Sato, H.; Mori, K.; Murakami, R.; Ando, Y.; Tagahashi, I.; Zhang, J. M.; Terauchi, H.; Hirose, F.; Senda, K.; Tashiro, K.; Noda, I.; Ozaki, Y. *Macromolecules* **2006**, *39*, 1525.
(46) Zhang, J. M.; Sato, H.; Noda, I.; Ozaki, Y. *Macromolecules* **2005**, *38*, 4274.