Multiple Band-Pass Filtering Method for Improvement on Prediction Accuracy of Linear Multivariate Analysis

JIANAN Y. QU* and LAN SHAO

Department of Electrical and Electronic Engineering, Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong, P. R. China

An approach coupling signal processing and partial least-squares regression analysis (PLS) is described in which raw spectral data are processed with a multiple band-pass filter and the filtered spectra are used in a PLS to build a calibration model for the analyte of interest. The multiple band-pass filter is specifically designed for a desired analyte based on the Fourier frequency characteristics of the pure spectrum of the desired analyte and the spectra of the interference background. It maximizes the ratio of signal to background. This combined multiple band-pass filtering and PLS method (MFPLS) was evaluated by determining clinically relevant levels of glucose, urea, ethanol, and acetaminophen in simulated human sera, in which triglyceride was simulated with triacetin; bovine serum albumin and globulin were used to model protein molecules in the serum. The results demonstrate that MFPLS produces better accuracy of prediction than PLS in all instances.

Index Headings: **Multivariate calibration; Near-infrared Raman spectroscopy; Digital filter.**

INTRODUCTION

Linear multivariate calibration is frequently used to relate the concentration of a desired analyte to a measured response such as absorption and Raman spectra of multi-component mixtures or complex biological samples.[1,2] The optimal choice of a multivariate algorithm depends on the available information on the measurements. In many cases, only the concentrations of one or a few analytes of interest in a set of multi-component mixtures are known. The corresponding spectra of the mixtures, called the calibration set, can be used with linear multivariate calibration methods such as a partial least-squares (PLS) and principal component regression (PCR) to build a calibration model for the analytes of interest. Investigations have demonstrated that PLS is an effective method for quantification of a desired analyte in the multi-component mixtures because of the quality of the calibration models produced.[2,3]

In a standard PLS regression procedure, the latent variables are calculated simultaneously along with the development of the calibration model. The concentration information concerning the desired analyte is used to determine the latent variables and ensure the maximum correlation of the variables with the response of the desired analyte in the calibration set. Hence, PLS has better capability than other methods to reject interference and build a reliable calibration model. However, even with such a sophisticated regression procedure, confounding of the desired signal by interferants such as irrelevant components, noise, and baseline variation will affect the accuracy of PLS prediction. The calibration model can be severely corrupted when the contribution of the desired signal to the signal variance of a multi-component mixture is much smaller than that of the interferants. In a previous study of near-IR Raman tests on glucose in the human sera, we found that the signal of glucose at the physiological level is hundreds of times lower than that of the proteins.[4] The PLS calibration model built on the Raman spectral data of human serum samples produced poor accuracy of prediction of glucose concentration. To demonstrate that the large molecules are the major interferants causing the corruption of the PLS model, we removed them by ultrafiltration. The prediction accuracy was improved tremendously in the PLS calibration model built on the spectra recorded from the ultrafiltered serum samples.[4] Though the ultrafiltration can ''physically'' eliminate major interference, the procedure has many limitations. In particular, it becomes invalid when the molecular weights of desired analytes are not significantly different from those of the major interferants. An approach that does not require complicated preprocessing of samples and can improve the multivariate calibration is then more desirable.

Various approaches have been explored in an attempt to minimize the prediction error in PLS caused by uninformative signals, which confound the calibration model. It was demonstrated that an optimally designed digital filter for spectral preprocessing could improve the results of PLS by removing spectral artifacts prior to the building of the calibration model.[5,6] A method using wavelet analysis to extract the relevant component for multivariate calibration was introduced.[7] The improved PLS results show that this approach could successfully remove noise and irrelevant information from the spectra for multivariate calibration. In the past ten years, considerable effort has been made to develop the variable selection method for identifying a subset of spectral data that produces the best accuracy in PLS calibration.[8–13] Furthermore, it has been mathematically proved that variable selection can remove the useless data that confounds PLS models and improve the calibration.[14]

It should be noted that the standard PLS procedure does not assume the pure spectrum of the desired analyte as known *a priori.* In PLS, the spectral information of a single analyte is hidden in the latent variables. For the methods summarized above, the pure spectrum of the desired analyte is also not used in the procedure of spectral preprocessing to improve the multivariate calibration. Recently, a new linear multivariate calibration method, called hybrid linear analysis (HLA), was reported.[15] It incorporates the spectrum of the analyte of interest into

Received 5 February 2001; accepted 14 May 2001.
* Author to whom correspondence should be sent.

0003-7028/01/5510-1414$2.00/0

the calibration and achieves better prediction accuracy than PLS. Work in our laboratory is focused on the study of biological samples by spectroscopy. The pure spectrum of various analytes of interest can easily be measured. In view of the poor prediction accuracy of PLS under circumstances where the desired signal to interfering background ratio (SBR) is small in the spectral data of the biological analyte mixture, such as serum,[4] we propose a digital filtering method, which enhances the SBR, before applying the PLS calibration. Here, the filter is designed by taking advantage of knowing both the pure spectrum of the desired analyte and the spectra of the biological analyte mixtures. This study was motivated by the work reported in Refs. 5, 6, and 15.

Mathematically, an optimization problem can be formulated to maximize the SBR after the filtering is applied to the raw spectral data:

$$\min_{f(k)} \frac{\|f(k)\cdot s(k)\|}{\|f(k)\cdot p(k)\|} \tag{1}$$

subject to $s(k) \gg p(k)$, where k is the variable, $f(k)$ is the impulse response of the optimal filter, $s(k)$ is a raw spectrum of the sample, and $p(k)$ is the pure spectrum of the desired analyte in the sample. The variable $s(k)$ is a good approximation to the interfering background when $p(k)$ is much smaller than $s(k)$. To design a filter for PLS calibration, the optimization must be applied to all the spectra in the calibration set as follows:

$$\min_{f(k)} \sum_{i=1}^{N} \frac{\|f(k)\cdot \hat{s}_i(k)\|}{\|f(k)\cdot \hat{p}(k)\|} \quad \text{or} \tag{2}$$

$$\min_{f(k)} \frac{\|f(k)\cdot \overline{\hat{s}(k)}\|}{\|f(k)\cdot \hat{p}(k)\|} \tag{3}$$

where $\hat{s}_i(k) = s_i(k)/\|s_i(k)\|$ and $\hat{p}(k) = p(k)/\|p(k)\|$ are the normalized spectrum of the ith sample and normalized pure spectrum, respectively. $\overline{\hat{s}(k)}$ is the mean spectrum of all the normalized spectra in the calibration set. N is the number of samples. Here, each spectrum in the calibration set was equally weighted in the optimization by using the normalized spectra instead of the raw ones. In the frequency domain, the optimization problem becomes

$$\min_{f(e^{j\omega})} \frac{\|f(e^{j\omega})\cdot \overline{\hat{s}(e^{j\omega})}\|}{\|f(e^{j\omega})\cdot \hat{p}(e^{j\omega})\|} \tag{4}$$

In principle, if the Fourier frequency distributions of the desired signal and the interfering background are different, the SBR must have local maxima at a few certain frequency bands. Therefore, an optimal filter of multiple frequency bands, at which SBR reaches a maximum, can be built to preprocess the raw spectra and enhance the SBR for PLS calibration. The calibration results based on the filtered spectral data are expected to be improved.

EXPERIMENTAL

Instrumentation. The Near-IR Raman spectra were recorded with a single-stage holographic grating imaging spectrograph (model No. HoloSpec f/1.8i, Kaiser Optical Systems, Inc.), equipped with a liquid nitrogen cooled CCD with 400 × 1340 pixels (Model LN/CCD-1340/400-EB/1, RS Roper Scientific) and a diode laser (Model PI-ECL-745-300, Process Instruments, Inc.) of wavelength 745 nm and output of 150 mW as an excitation source. The Raman cell was a home-made disposable quartz capillary of inner diameter ranging from 250 to 300 μm and length of 20 mm. The laser was conducted into the capillary with a 100 μm diameter optical fiber. The excitation light was totally guided by the capillary when it was filled with the aqueous sample. The Raman signal was collected with two multiple optical fiber arrays at right angles to the capillary. The fiber arrays were placed along the capillary. The distance from the distal tips of the fibers to the capillary was about 200 μm. The other ends of the fibers were combined and lined up at the entrance of the spectrograph to form an entrance slit. The Raman signals from the fibers were dispersed by the spectrograph with a wavenumber resolution of 20 cm^{-1}. Raman spectra were then formed by binning over 400 pixels of the CCD vertically.

Reagents and Procedures. The three groups of samples used in this work were: (1) four aqueous stock solutions of pure glucose, urea, ethanol, and acetaminophen; (2) a set of mixtures of the four analytes in phosphate-buffered saline (PBS); and (3) a set of mixtures of the four analytes in the aqueous phosphate buffer matrix containing triactin, bovine serum albumin, and globulin for the simulation of human serum. Glucose and urea are important metabolites. Ethanol and acetaminophen are potentially toxic substances when their concentrations in blood are high. The triactin, bovine serum albumin, and globulin were used to model triglycerides and total proteins in the blood, respectively.[6] The metabolites in the second and third group of samples were present in levels spanning the human physiological range. The concentration range of ethanol and acetaminophen covered their toxic levels in human blood. All the analytes used in preparation of the samples were reagent grade and purchased from Sigma Chemical Co., St. Louis, MO.

The four stock solutions in the first group of samples were used to measure the pure Raman spectra of glucose, urea, ethanol, and acetaminophen. The mean physiological levels of glucose and urea in normal adult serum are 4.6 mM and 4.4 mM, respectively.[16] The toxic levels of ethanol and acetaminophen in human serum are 21.7 mM and 1.3 mM, respectively.[17] To ensure the extraction of the pure spectrum with high signal-to-noise ratio (SNR) for the development of multiple band-pass filters, concentrations of the four analytes were higher than their mean physiological levels and toxic levels. Specifically, the concentrations of glucose, urea, ethanol, and acetaminophen were 30 mM, 30 mM, 100 mM, and 30 mM, respectively. The second group of samples was thirty mixtures of glucose, urea, ethanol, and acetaminophen in PBS. Glucose, urea, ethanol, and acetaminophen in the stock solutions were prepared at thirty levels spanning the range 1–11 mM, 2–10 mM, 0–46 mM, and 0–10 mM, respectively. The concentrations of the four individual analytes were randomized in each sample to eliminate possible correlation in the samples. The Raman spectra recorded from the second group of samples were used in the PLS calibration for glucose, urea, ethanol, and acetaminophen. The calibration results should then set the accuracy limits of Raman tests on the four analytes in the mixture without interference from proteins and triglyceride. The third group of samples was used for the

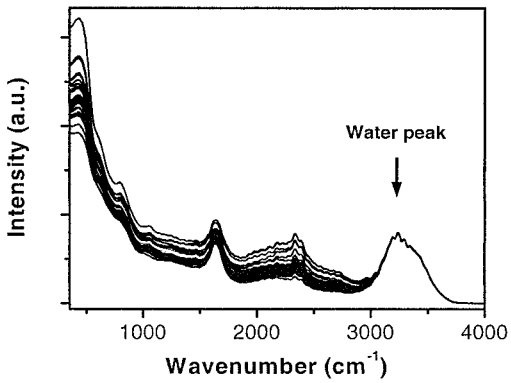

FIG. 1. Raman spectra of thirty mixtures of glucose, urea, ethanol, and acetaminophen in PBS.

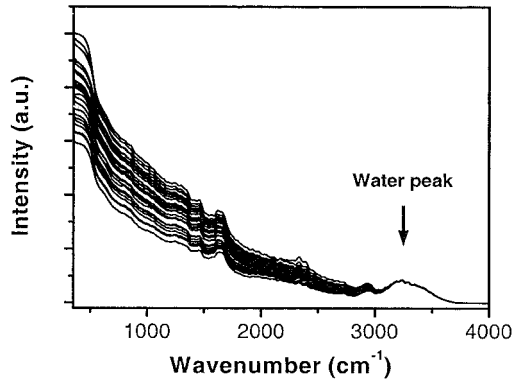

FIG. 2. Raman spectra of simulated serum samples.

multivariate calibration of glucose, urea, ethanol, and acetaminophen in simulated human serum. Thirty stock solutions were prepared by mixing glucose, urea, ethanol, acetaminophen, triactin, albumin, and globulin in PBS. Concentrations of glucose, urea, ethanol, and acetaminophen were in the same range as the second group of samples. The concentrations of albumin, globulin, and triactin were in the ranges 27–64 g/L, 20–40 g/L, and 0.5–5 g/L, respectively. Again, the concentrations of all the analytes in each sample were randomly set.

Experimental Spectra. The near-IR Raman spectra of the stock solutions in the second and third groups of samples are shown in Figs. 1 and 2. For each sample, a Raman spectrum was acquired in 10 s. The wavelength range of a full spectrum was from 350 to 4000 cm^{-1}. The recorded spectral data were processed with a 5-point adjacent averaging filter to remove the high frequency noise and preserve the spectral features. All spectra used in the calibration were normalized to the strong water peak at 3350 cm^{-1} to eliminate errors caused by the variation of sample volume and excitation/collection conditions from measurement to measurement. The fluctuations of the spectra in the range from 350 cm^{-1} to 3000 cm^{-1} were mainly caused by the thickness variation of the walls of the Raman cells because the cells were homemade in the department glassblowing facility and the capillary size could not be well controlled.

The pure spectra of glucose, urea, ethanol, and acetaminophen were obtained by subtracting the PBS spectrum from the spectra collected from the first group of stock solutions. To ensure the accurate extraction of the pure spectrum, the same Raman cell was used to measure the spectra of PBS and the four stock solutions. The spectra were acquired over 60 s. Figures 3A–3D show the pure spectra of glucose, urea, ethanol, and acetaminophen.

RESULTS AND DISCUSSION

PLS calibration and prediction were performed separately on the spectra collected from the second and third groups of stock solutions. Thirty spectra in each group were used for the calibration. A cross-validation based on the leave-one-out procedure was employed in which 29 spectra were used to develop the calibration model to predict the concentrations of the desired analytes in the omitted sample. Each spectrum was rotated out in turn. The prediction accuracy was evaluated by computing the root mean squared error of prediction (RMSEP) for the 30 rounds of cross-validations. This gave an estimate of the average prediction ability for the PLS models built on 29 samples.

The predictions of concentrations of glucose, urea, ethanol, and acetaminophen in the two groups of samples are shown in Figs. 4A–4D. Here, we observe that the

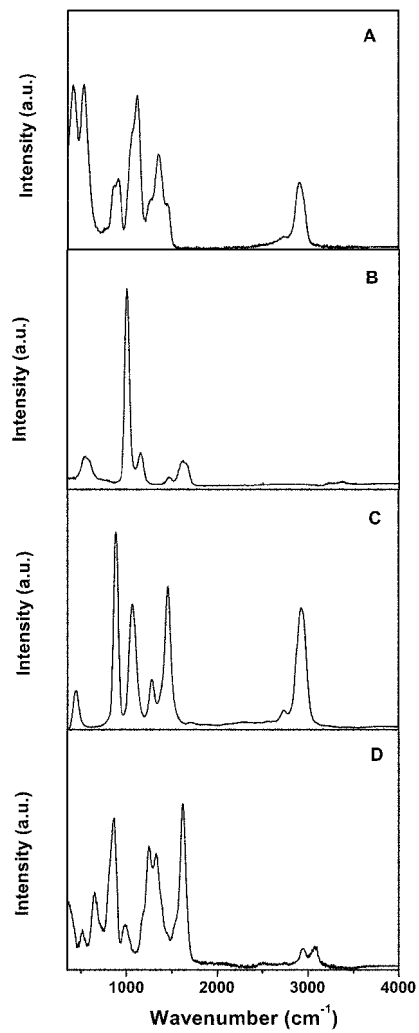

FIG. 3. Pure Raman spectra of (**A**) glucose, (**B**) urea, (**C**) ethanol, and (**D**) acetaminophen.

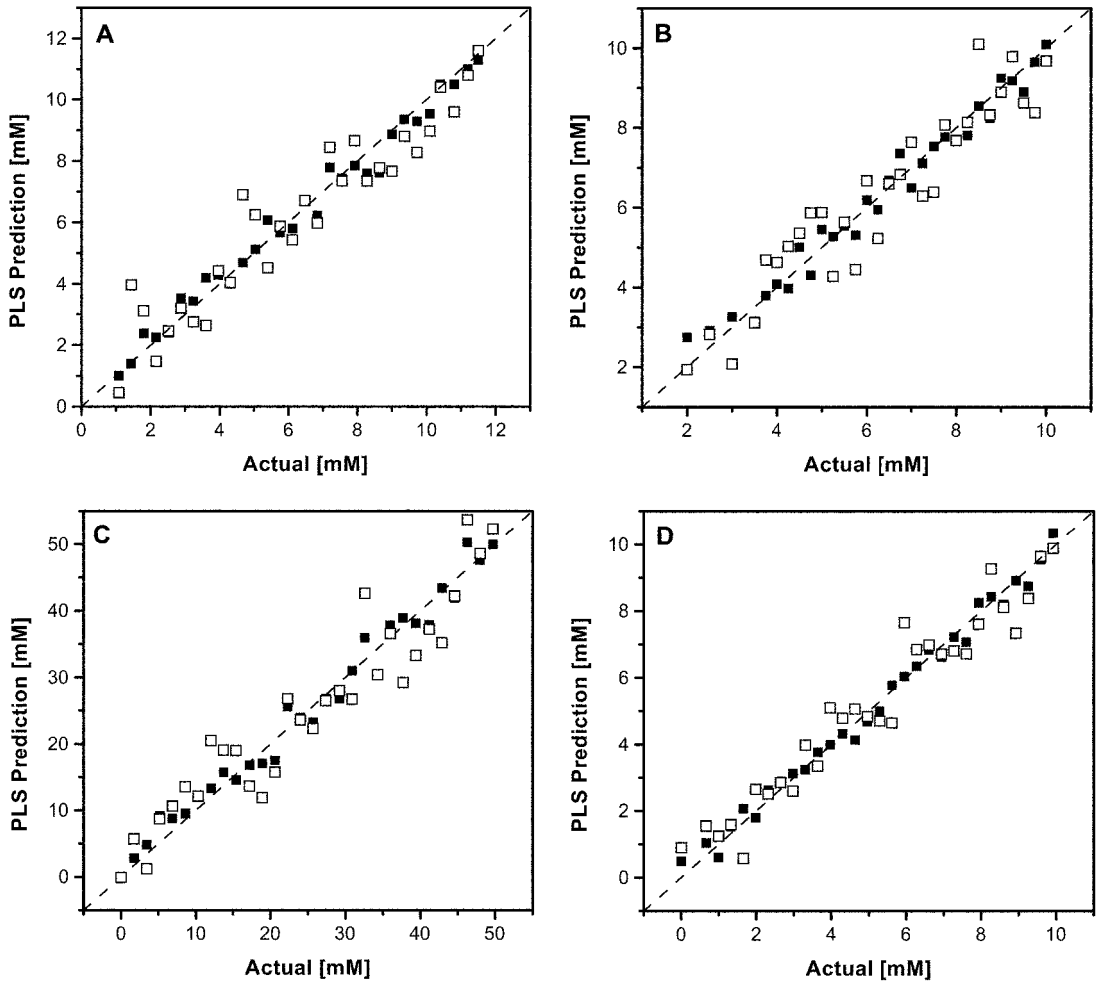

FIG. 4. PLS prediction plots for (**A**) glucose, (**B**) urea, (**C**) ethanol, and (**D**) acetaminophen. The solid squares represent the predicted concentration vs. actual concentration in the mixtures of the four analytes. The open squares represent the predictions in simulated sera.

prediction errors for the four analytes obtained in the second group of samples are significantly lower than that in the third group of samples. This demonstrates once again that the presence of proteins and triactin contributes to the errors of PLS calibration for glucose, urea, ethanol, and acetaminophen as reported in Ref. 4.

To design an optimal filter for the enhancement of SBR and improvement of PLS calibration, we analyzed the frequency characteristics of the pure spectra and the interfering background. Figure 5 shows the normalized Fourier transforms of the pure spectra and mean spectrum of the calibration set. Their differential spectra are also shown in the figure. It is obvious that the frequency distributions of the desired analytes are different from that of the interfering background. The major frequency components for pure signals and background are distributed in the range of the normalized frequency, from 0 to 0.15 Hz. Intuitively, the SBR at the frequency bands, where $\hat{p}(e^{j\omega})$ is greater than $\overline{\hat{s}(e^{j\omega})}$, should be higher than that of the raw signal. On the contrary, the SBR is poorer than the raw signal in the frequency range where $\hat{p}(e^{j\omega})$ is lower than $\overline{\hat{s}(e^{j\omega})}$. The optimization problem then becomes to build a filter to select the frequency components with high SBR and reject the components with low SBR.

An exhaustive search method was used to find the optimal filter. The initial multiple band-pass filter for the search was defined by the function $\Gamma(\omega) = \{\mathrm{Sgn}[\hat{p}(e^{j\omega}) - \hat{s}(e^{j\omega})] + 1\}/2$, which selects all the frequency bands making SBR higher than that of the mean spectrum of the calibration set. The components beyond the normalized frequency of 0.15 Hz were cut off because their contributions to the total signal are negligible. For this research, the finite impulse response (FIR) filter, which is a commonly used digital filter with stable performance,[18] was used to form the multiple band-pass filter. One FIR filter was employed in each single frequency band. The Γ-function here determined the total number of FIR band-pass filters used and the frequency range for each FIR band-pass filter. During the optimization, the bandwidth and central frequency of each FIR filter was allowed to change. In detail, the bandwidth and central frequency are defined by a pair of uniformly distributed random numbers with the criteria that the FIR filter must be in its corresponding frequency range determined by the Γ-function. The optimal filter was found under the condition that the PLS model based on the filtered calibration set produced the lowest RMSEP of cross-validation. The exhaustive search needs a computation time of up to a few days with a PC depending on the total number of frequency bands for the FIR filter set.

The optimal filters for glucose, urea, ethanol, and acetaminophen are shown in the Fig. 6. The differential spec-

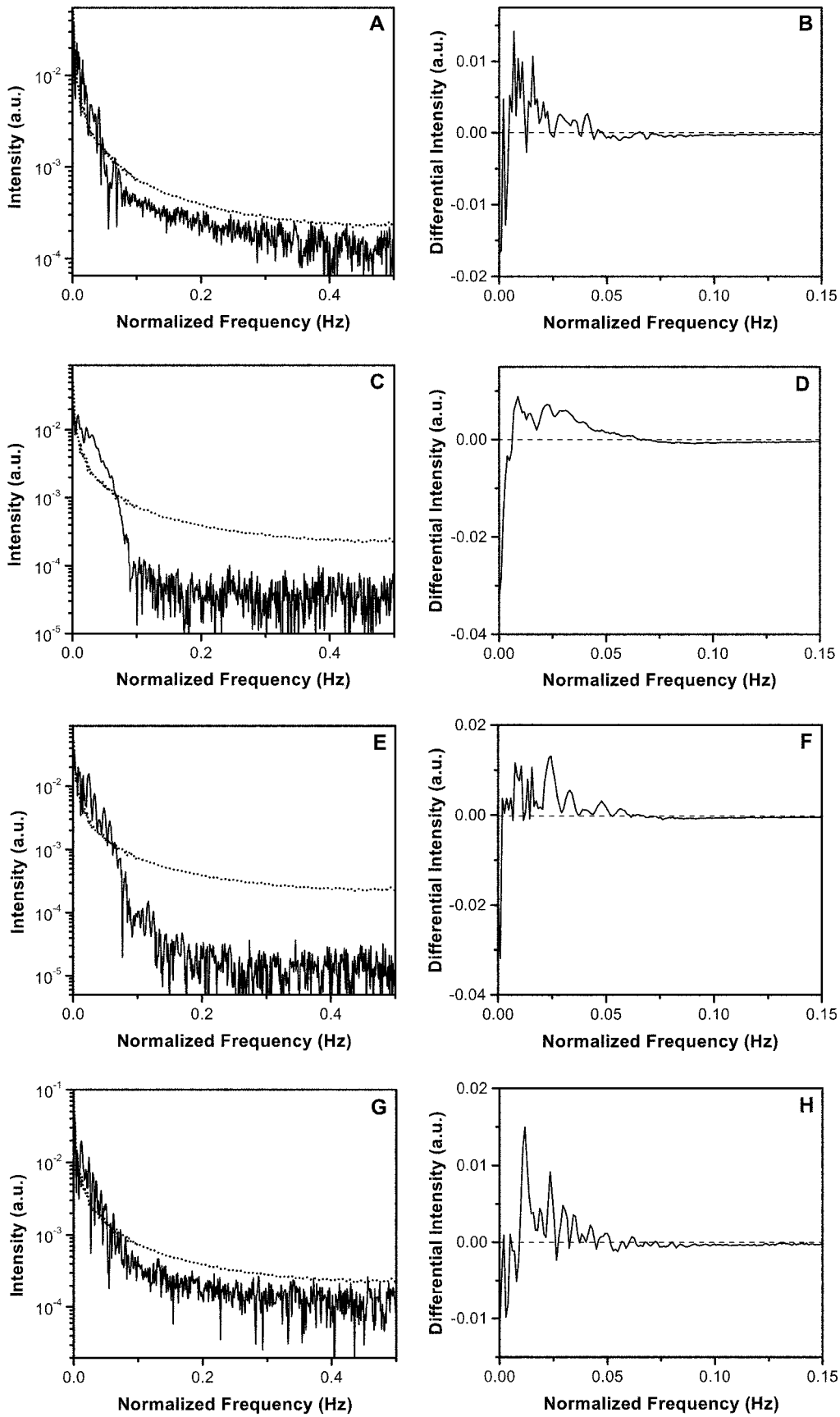

FIG. 5. Fourier transforms of pure spectrum, $\hat{p}(e^{j\omega})$, mean spectrum of calibration set, $\overline{\hat{s}(e^{j\omega})}$, and differential spectrum, $\hat{p}(e^{j\omega}) - \overline{\hat{s}(e^{j\omega})}$. (**A, B**) $\hat{p}(e^{j\omega})$ and $\hat{p}(e^{j\omega}) - \overline{\hat{s}(e^{j\omega})}$ for glucose. (**C, D**) $\hat{p}(e^{j\omega})$ and $\hat{p}(e^{j\omega}) - \overline{\hat{s}(e^{j\omega})}$ for urea. (**E, F**) $\hat{p}(e^{j\omega})$ and $\hat{p}(e^{j\omega}) - \overline{\hat{s}(e^{j\omega})}$ for ethanol. (**G, H**) $\hat{p}(e^{j\omega})$ and $\hat{p}(e^{j\omega}) - \hat{s}(e^{j\omega})$ for acetaminophen. The dotted lines represent the Fourier transform of background $\overline{\hat{s}(e^{j\omega})}$.

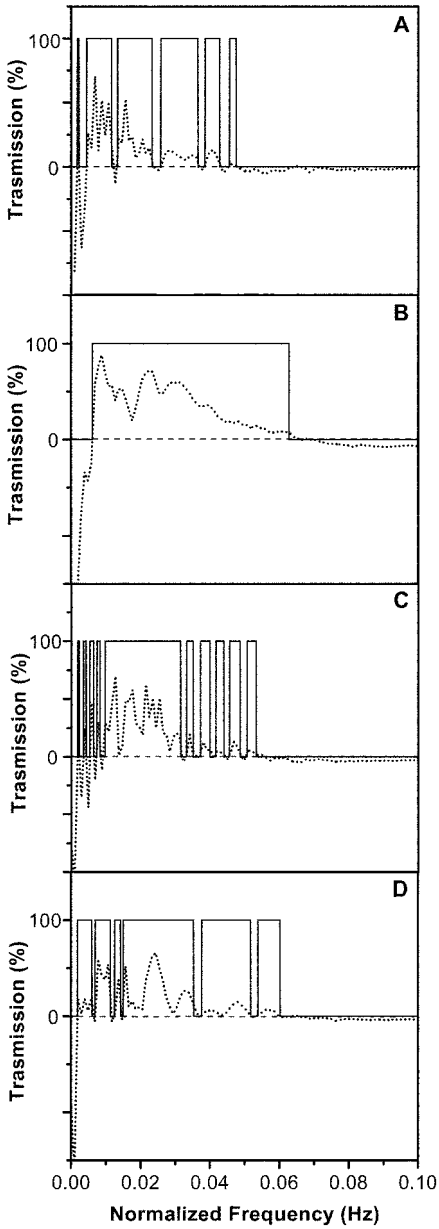

FIG. 6. Optimal multiple band-pass filter (solid lines) for (**A**) glucose, (**B**) urea, (**C**) ethanol, and (**D**) acetaminophen. The differential spectra (dotted lines) are scaled up for better illustration of the high SBR bands and low SBR bands.

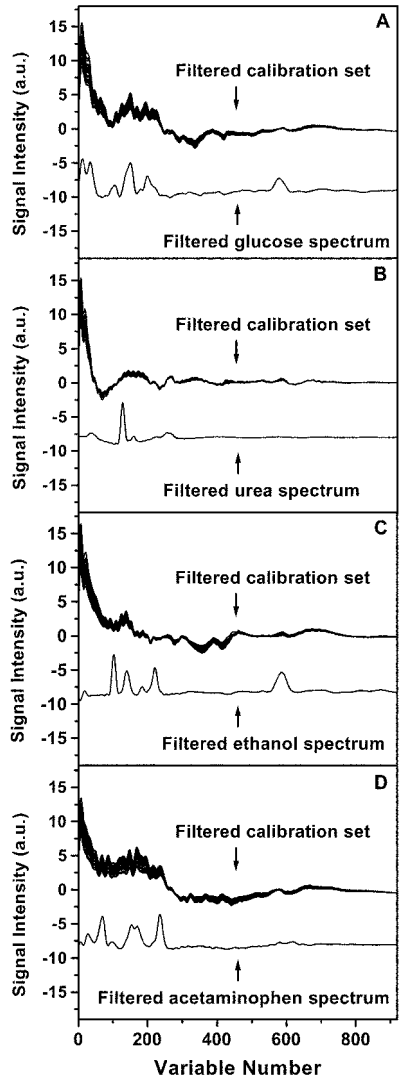

FIG. 7. Optimally filtered calibration set and pure spectra. (**A**) Calibration set and pure glucose spectrum processed by the optimal filter for glucose. (**B**) Calibration set and pure urea spectrum processed by the filter for urea. (**C**) Calibration set and pure ethanol spectrum processed by the filter for ethanol. (**D**) Calibration set and pure acetaminophen spectrum processed by the filter for acetaminophen. The filtered pure spectra are offset and scaled up for clarity.

tra between the Fourier transforms of pure spectra and the background are displayed as dotted lines, as the references. It is not surprising that the frequency bands covered by the optimal multiple band-pass filters are almost the same to those defined by the Γ-function. This means that the optimal filter preserves most of the frequency components with the SBR higher than the raw data and rejects the components with the SBR lower than the raw data. The filtered spectra of the calibration set and the pure spectra are shown in Fig. 7. We found that most of the features of the pure spectra remain the same after filtering. However, the variance of the filtered spectra has been significantly reduced compared with the raw spectra shown in Fig. 2.

A side by side comparison of PLS and MFPLS predictions for glucose, urea, ethanol, and acetaminophen in the simulated sera are displayed in Fig. 8. Here, we observe that the filtering method substantially improved the prediction accuracy of the PLS calibration for all four analytes. Table I summarizes the RMSEP and r^2 values for glucose, urea, ethanol, and acetaminophen generated by PLS and MFPLS in the simulated sera. The RMSEP and r^2 values produced by PLS in the mixtures of the four analytes are presented as the limits of prediction accuracy. As can be seen, the MFPLS gives lower RMSEP values in the prediction of each of the four analytes. The RMSEP values obtained by MFPLS are comparable with those produced by PLS in the mixtures of the four analytes. The results are consistent with our expectation that optimal filtering improves PLS calibration by enhancing the SBR and suppressing the confounding of desired signals by proteins, the major interferants.

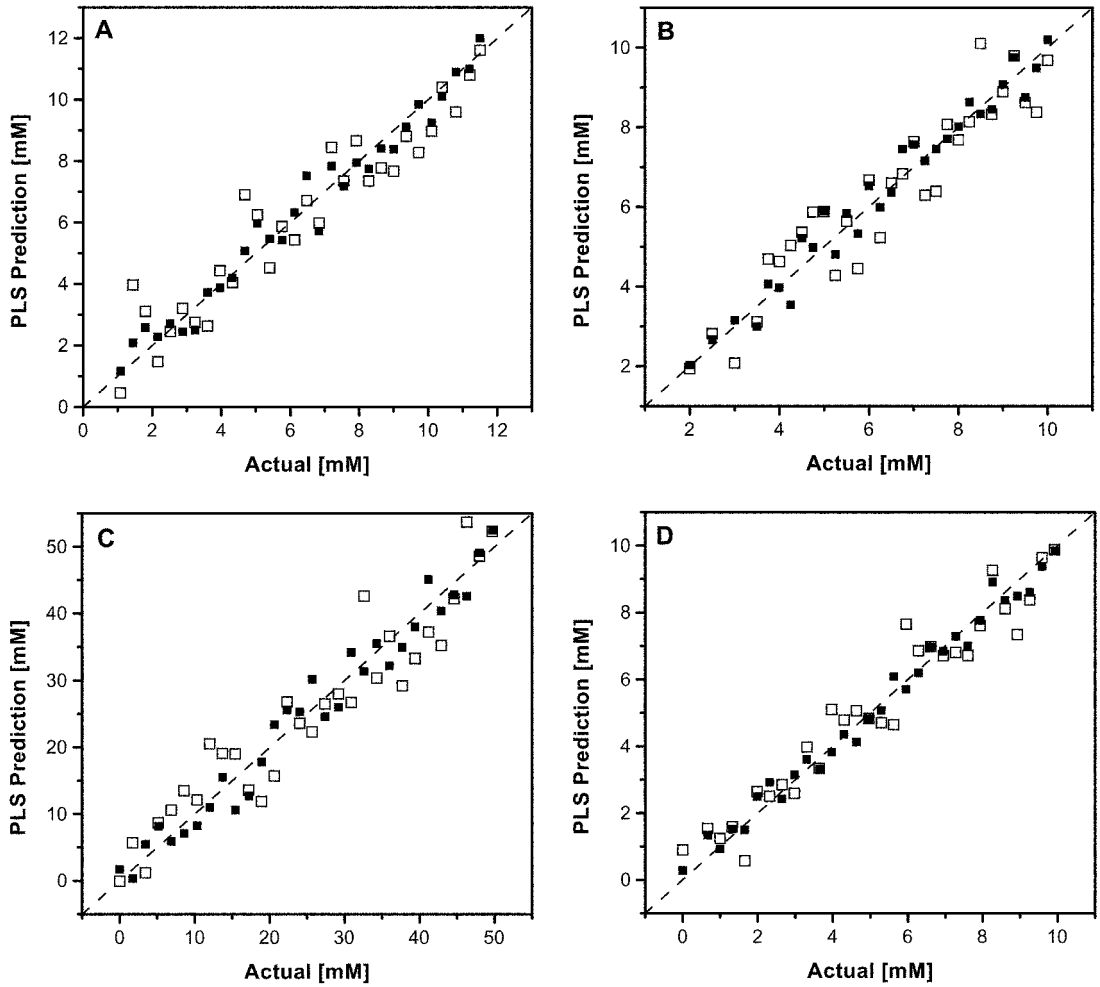

FIG. 8. MFPLS and PLS prediction plots for (**A**) glucose, (**B**) urea, (**C**) ethanol, and (**D**) acetaminophen. The solid squares represent the MFPLS predictions and the open squares represent the PLS predictions.

In conclusion, we have demonstrated in this study that MFPLS is a method superior to the standard PLS. The results indicate that the interference confounding the PLS calibration can be effectively reduced by applying the multiple band-pass filtering to the raw data, and the prediction accuracy has been increased. The MFPLS uses the information concerning the desired analyte and the pure spectrum as the guide for identifying the frequency components of high SBR and designing the optimal multiple band-pass filter. This advantage makes the optimization procedure more effective and the optimal filter more intuitive and robust. It should be noted that only the frequency characteristics of the the pure signal are required for development of the optimal filter. The error in the measurement of the amplitude of pure signal does not affect the performance of MFPLS. Here, the information extracted from the pure spectrum is used to guide the search for the optimal multiple band-pass filter. The method is thus not sensitive to the slight modulation in characterization of the pure spectrum by the solution matrix, as is HLA.

The MFPLS requires the clear difference in frequency characteristics between the desired signal and interfering signal to build a multiple band-pass filter to enhance the SBR. In general, the small and simple molecules with a few sharp and distinct Raman peaks, such as those investigated in this study, consist of more high frequency components than the uninformative background. Thus, the Raman spectroscopy satisfies the requirements of MFPLS. However, there may be some limitations to extending the MFPLS method to other spectroscopic techniques that provide smooth and broadened molecular spectra. For instance, the IR spectra of most molecules are broad, which mostly contributes to low frequency bands as interfering background. It may be more difficult to identify the bands of high SBR. The efficacy of MFPLS calibration on IR spectra may then be affected.

TABLE I. Comparison of RMSEP and r^2 for PLS and MFPLS.

	RMSEP values (mM)			r^2 values		
	PLS	MFPLS	Ref.[a]	PLS	MFPLS	Ref.[a]
Glucose	1.0	0.52	0.41	0.90	0.97	0.99
Urea	0.80	0.42	0.35	0.88	0.97	0.98
Ethanol	4.9	2.7	2.1	0.89	0.97	0.98
Acetaminophen	0.74	0.36	0.3	0.93	0.99	0.99

[a] The RMSEP and r^2 values obtained from the PLS calibration on the spectra measured in the mixtures of glucose, urea, ethanol, and acetaminophen in PBS.

ACKNOWLEDGMENT

This work was supported by Hong Kong Research Council grant HKUST6207/98P.

1. D. M. Haaland and E. V. Thomas, Anal. Chem. **60,** 1193 (1988).
2. H. Martens and T. Naes, *Multivariate Calibration* (John Wiley and Sons, New York, 1989), Chap. 3.
3. B. K. Lavine, Anal. Chem. **72,** 91R (2000).
4. J. N. Y. Qu, B. C. Wilson, and D. Suria, Appl. Opt. **38,** 5491 (1999).
5. R. E. Shaffer, G. W. Small, and M. A. Arnold, Anal. Chem. **68,** 2663 (1996).
6. M. J. Mattu, G. W. Small, and M. A. Arnold, Anal. Chem. **69,** 4695 (1997).
7. D. Jouan-Rimbaud, B. Walczak, R. J. Poppi, O. D. Noord, and D. L. Massart, Anal. Chem. **69,** 4317 (1997).
8. J. H. Kalivas, N. Roberts, and J. M. Sutter, Anal. Chem. **61,** 2024 (1989).
9. C. B. Lucasius and G. Kateman, Trends Anal. Chem. **10,** 254 (1991).
10. U. Horchner and J. H. Kalivas, Anal. Chim. Acta **311,** 1 (1995).
11. D. Jouan-Rimbaud, D. Massart, R. Leardi, and O. D. Noord, Anal. Chem. **67,** 4295 (1995).
12. A. S. Bangalore, R. E. Shaffer, G. W. Small, and M. A. Arnold, Anal. Chem. **68,** 4200 (1996).
13. M. J. McShane, B. D. Cameron, G. L. Coté, M. Motamedi, and C. H. Spiegelman, Anal. Chim. Acta **388,** 251 (1999).
14. C. H. Spiegelman, M. J. McShane, M. J. Goetz, M. Motamedi, Q. L. Yue, and G. L. Coté, Anal. Chem. **70,** 35 (1998).
15. A. J. Berger, T. W. Koo, I. Itzkan, and M. S. Feld, Anal. Chem. **70,** 623 (1998).
16. C. Lentner, *Geig Scientific Tables* (Ciba-Geigy, New Jersey, 1981), Chap. 3, pp. 89–115.
17. N. W. Tietz, *Clinical Guide to Laboratory Tests* (W. B. Saunders Company, Philadelphia, 1995), pp. 787–897.
18. N. J. Loy, *An Engineer's Guide to FIR Digital Filters* (Prentice Hall, New Jersey, 1988).