New Prediction-Augmented Classical Least Squares (PACLS) Methods: Application to Unmodeled Interferents *

David M. Haaland and David K. Melgaard

Sandia National Laboratories

Albuquerque, New Mexico 87185-0342

Abstract

A significant improvement to the classical least squares (CLS) multivariate analysis method has been developed. The new method, called prediction-augmented classical least squares (PACLS), removes the restriction for CLS that all interfering spectral species must be known and their concentrations included during the calibration. We demonstrate that PACLS can correct inadequate CLS models if spectral components left out of the calibration can be identified and if their "spectral shapes" can be derived and added during a PACLS prediction step. The new PACLS method is demonstrated for a system of dilute aqueous solutions containing urea, creatinine, and NaCl analytes with and without temperature variations. We demonstrate that if CLS calibrations are performed using only a single analyte's concentrations, then there is little, if any, prediction ability. However, if pure-component spectra of analytes left out of the calibration are independently obtained and added during PACLS prediction. then the CLS prediction ability is corrected and predictions become comparable to that of a CLS calibration that contains all analyte concentrations. It is also demonstrated that constant-temperature CLS models can be used to predict variable-temperature data by employing the PACLS method augmented by the spectral shape of a temperature change of the water solvent. In this case, PACLS can also be used to predict sample temperature with a

DISCLAIMER

DISCLAIMER

Portions of this document may be illegible in electronic image products. Images are produced from the best available original document.

standard error of prediction of 0.07 °C even though the calibration data did not contain temperature variations. The PACLS method is also shown to be capable of modeling system drift to maintain a calibration in the presence of spectrometer drift.

Key Words: Classical least squares (CLS); Prediction-augmented classical least squares (PACLS); Multivariate calibration; Near IR spectroscopy; Aqueous solutions; Temperature calibration; Spectrometer drift.

*Sandia is a multi-program laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy under Contract DE-ACO4-94AL85000.

INTRODUCTION

Classical least squares (CLS) multivariate modeling has been used for the quantitative analysis of infrared spectra for over 20 years.[1,2,3,4,5,6,7,8] The CLS calibration and prediction algorithms are based upon explicit linear additive models, e.g., Beer's law, that require the quantitative knowledge of all spectrally active components in the calibration sample set. With CLS modeling, it has not been possible to accurately account for spectral variations resulting from spectrometer drift, sample insertion effects, or system nonlinearities since explicit equations required to model these effects are not known. The introduction of partial least squares (PLS)[9,10,11] and principal component regression (PCR)[12] factor analysis methods provided the analyst with algorithms that could be used even if only the concentrations of a single analyte were known in the calibration sample set. PLS and PCR analysis methods could also empirically model spectral variations due to spectrometer drift, sample insertions, and unknown interferences in the calibration spectra. PLS and PCR are even capable of modeling nonlinearities in the data through the addition of factors that can approximate the nonlinear behavior. CLS methods were then relegated to the analysis of simple well-characterized linear systems or gas-phase samples[13] where Beer's law was followed and all spectrally interfering components were known. We have continued to use CLS methods for qualitative spectral interpretation since CLS always generates better pure-component spectral estimates than possible with either PLS or PCR.[14] However, for quantitative analysis of spectral data, we have generally used PLS or PCR because they exhibit superior quantitative prediction performance relative to CLS, except possibly in

the quantitative analysis of simple infrared gas-phase spectra[15] or inductively coupled plasma atomic emission spectra.[16]

To address the limitations of CLS, we have developed a new CLS-based algorithm that we have named prediction-augmented classical least squares (PACLS). PACLS can significantly improve the applicability and flexibility of CLS methods. With the PACLS algorithm, the detrimental effects of unknown components in the calibration, temperature variations, spectrometer drift, sample insertion related optical effects, and even nonlinearities in the CLS calibration model can be corrected during the CLS prediction phase of the analysis. To correct the harmful effects of the above sources of spectral variation, the spectral intensities or spectral shapes of the spectral variations not included during CLS calibration must be empirically measured and included in the CLS prediction portion of the analysis. We will show that adding the missing spectral shapes during CLS prediction compensates for the prediction errors generated when knowledge of their presence in the calibration data is not explicitly included as component concentrations in the CLS calibration. A variety of methods to empirically obtain the spectral shapes required to correct the detrimental effects will be discussed.

In this paper, we describe the new PACLS algorithm and demonstrate its use with near-infrared (NIR) spectra from a set of multi-component dilute aqueous solutions. An explanation of how the new PACLS method can produce accurate results in the presence of an inadequate model will be presented. The PACLS method will first be demonstrated by performing the CLS calibration after excluding some of chemical components from the model. The deficient CLS model will then be used with and without the spectral shapes of the missing components added during CLS prediction to compare the

prediction abilities of the two methods on unknown samples. In addition, a constant-temperature CLS model will be applied to sample spectra obtained at variable temperatures. The CLS predictions will be compared with and without the spectral shape of the effect of temperature changes added to the CLS prediction. We will also show that the new PACLS method allows for accurate solution temperatures to be predicted even when temperature variation was not a parameter that was included in the original CLS calibration data.

EXPERIMENTAL

The samples and NIR spectra used in this study have been described previously.[17,18] The samples consisted of 31 dilute solutions of urea, creatinine, and NaCl in a water solvent. The 31 compositions were obtained via a repetitive sampling scheme[19] that produced a pseudo D-optimal design with each of the three components separately varied at 16 levels over the concentration range from 0 to approximately 3000 mg/dL. The spectra of the samples, sealed in 10-mm pathlength cuvettes, were obtained in random order, and the spectra of three samples were obtained again at the end of each study. Spectra were collected initially at a constant temperature of 23°C and collected several days later with the samples varying over a temperature range of 20 to 25°C. All samples were maintained at the design temperature using a Hewlett Packard (HP) Peltier temperature controller that could maintain sample temperatures to 0.05°C ($\pm 1\ \sigma$). The HP temperature controller allowed 1000 rpm stirring with a Teflon-coated magnetic stirring bar sealed in the cuvette. In order to assure that the samples had equilibrated to the design temperature, long equilibration times ($\geq$ 8 min) were used. The total time of

data collection was 7 to 9 hr during a single day. Therefore, significant spectrometer drift was evident over the time of the data collection. In a separate experiment, variable-temperature spectra of pure water in a cuvette were obtained in random order at 0.5°C intervals from 20 to 25°C.

NIR spectra were collected on a Nicolet Model 800 Fourier transform infrared (FT-IR) spectrometer equipped with a liquid-N_2-cooled InSb detector, a quartz beam splitter, and a 75-W tungsten-halogen lamp. A total of 256 interferogram scans were signal averaged for each sample and background spectrum. Interferograms were Fourier transformed after applying Happ-Genzl apodization to obtain single-beam spectra at a nominal resolution of 16 cm^{-1}. Background spectra were collected of an empty cuvette after each sample spectrum. Best prediction results were obtained when using an averaged background for all samples rather than a separate background for each sample. Therefore, the single-beam sample spectra were ratioed to the average background spectrum and converted to absorbance.

The CLS and PACLS algorithms were programmed at Sandia National Laboratories using the Array Basic language of the GRAMS 32 software (Version 5.1). Spectra were analyzed over the spectral range from 7500 to 11000 cm^{-1}. Cross validation leaving out one sample at a time was employed to obtain cross-validated standard errors of prediction (CVSEP) for assessing prediction ability and to improve outlier detection. All spectra were included in the analyses since spectral F ratio[10] and Mahalanobis distance[19] outlier metrics did not indicate any outlier samples.

THEORY

The CLS calibration and prediction algorithms have been presented previously in various forms.[1-8] In this discussion, matrices are represented as upper-case bold letters; vectors are represented as column vectors using lower-case bold letters. Row vectors and transposed matrices are denoted by a superscript T. Lower-case letters in italics represent scalars. The CLS model can be written

$$\mathbf{A} = \mathbf{CK} + \mathbf{E}_A \qquad (1)$$

where $\mathbf{A}$ is the n x p matrix of absorbances for the n samples at the p frequencies, $\mathbf{C}$ is the n x m matrix of reference concentrations for the m components, $\mathbf{K}$ is the m x p matrix of pure-component spectra at unit concentration. Sample pathlength can be included in Eq. 1 by dividing the intensities for each spectrum (row) in $\mathbf{A}$ by the known pathlength of the sample. During calibration, we solve for the least squares solution of $\mathbf{K}$, i.e., $\hat{\mathbf{K}}$. The least-squares solution, $\hat{\mathbf{K}}$, is given by

$$\hat{\mathbf{K}} = (\mathbf{C}^T\mathbf{C})^{-1}\mathbf{C}^T\mathbf{A}. \qquad (2)$$

A variety of methods, including singular value decomposition,[20] can be employed to improve the numerical precision of the solution to Eq. (2). If all components are spectrally active in the spectral region analyzed and their concentrations are included in Eq. 2, then the data should not be mean centered since the $(\mathbf{C}^T\mathbf{C})$ matrix to be inverted in Eq. 2 will be nearly singular (it will be closer to singular for ideal solutions and as errors in the reference concentrations decrease). $\mathbf{A}$ and $\mathbf{C}$ can be mean centered if at least one component is not spectrally active in the spectral region being analyzed or if pathlengths of the samples are variable.[4] If sources of spectral variation are not represented by component concentrations in the $\mathbf{C}$ matrix, then the $\hat{\mathbf{K}}$ matrix will not accurately

represent the pure-component spectra. As will be demonstrated in the Results and Discussion Section, errors in the estimated pure-component spectra can result in significant prediction errors. Although we use a single continuous region of the spectra, all the methods presented here are also applicable to spectra with discontinuously selected spectral intensities.

During CLS prediction, we solve for the least-squares estimated component concentrations, $\hat{\mathbf{C}}_u$, of the m components in the n_u unknown samples to be predicted. The subscript u is used to indicate unknown samples. The CLS solution for $\hat{\mathbf{C}}_u$ is given by

$$\hat{\mathbf{C}}_u = \mathbf{A}_u\hat{\mathbf{K}}^T(\hat{\mathbf{K}}\hat{\mathbf{K}}^T)^{-1} \qquad (3)$$

where $\mathbf{A}_u$ represents the spectral matrix of the unknown samples to be predicted. We originally described[2-4] how the $\hat{\mathbf{C}}_u$ and $\hat{\mathbf{K}}$ matrices can be augmented to account for baseline variations in the data. The $\hat{\mathbf{K}}$ matrix can be augmented by a row of ones to represent a baseline offset and by a row of integers representing the index of the spectral data (e.g., indexed in order of spectral frequency for Fourier transform infrared data) to represent a linearly sloping baseline. The row of integers should be linearly mapped to the region from -1 to 1 to improve the condition of the $\hat{\mathbf{K}}\hat{\mathbf{K}}^T$ matrix to be inverted. Quadratic baselines can be added by simply adding a row that is the square of the row representing the linear baseline slope. Higher order baseline terms can be added by adding rows of the higher order transformations of the linear baseline slope. Alternatively, a set of orthogonal Legendre polynomials can be added to the $\hat{\mathbf{K}}$ matrix to represent polynomial baselines of any order. In a similar manner, any functional form of

spectral baselines can be fitted by augmenting the $\hat{K}$ matrix with rows representing the functional form of the baseline to be fitted. For each row added to the $\hat{K}$ matrix, coefficients representing the fitted magnitudes of the baseline components for each prediction sample must be added as columns to the sample concentration matrix $\hat{C}_u$. The augmentation of the $\hat{C}_u$ and $\hat{K}$ matrices provides for a simultaneous fit of the baseline components and the linear additive pure-component spectra. Unless the baseline variation is orthogonal to all other sources of spectral variation, a simultaneous least squares fit of these baseline spectral shapes is always preferable to simply baseline correcting the spectral data as a separate preprocessing step.

The prediction-augmented classical least squares (PACLS) method presented here is similar to the addition of the explicit baseline shapes during prediction to correct for simple baseline variations in the spectral data to be predicted. However for PACLS, empirically determined shapes are added in addition to the theoretical functional forms of the baselines. The new PACLS method is useful when all sources of spectral variation in the spectral region being analyzed are not known during the calibration phase of the CLS analysis. The exclusion from the calibration of spectral component concentrations or other parameters that cause spectral changes in the samples will result in estimated pure-component spectra that are each contaminated by the spectral variation of the unmodeled spectral components. The degree of contamination will depend on the design of the calibration sample set and the magnitude and number of independent sources of spectral variation left out of the CLS calibration. The use of contaminated CLS-estimated pure-component spectra during the CLS prediction will cause prediction errors to be larger relative to a CLS analysis that included representations of all sources of spectral variation

in the CLS calibration. The new PACLS algorithm allows for spectral shapes of components that were left out of the calibration to be added during CLS prediction in order to compensate for the absence of those spectral components during the CLS calibration. The requirement for the PACLS algorithm is that important spectral components left out of the CLS calibration have their spectral shapes or linear combinations of their shapes identified and included during the CLS prediction phase of the analysis. As in the case of baseline augmentation, the new PACLS algorithm uses the same equations as the CLS algorithm except that during CLS prediction, the $\hat{K}$ matrix in Eq. 3 is augmented with rows representing the spectral shapes of those spectral components that were not included in the CLS calibration. If $\hat{\hat{K}}$ represents the augmented matrix with spectral shapes added as rows in the original $\hat{K}$ matrix, then the PACLS prediction becomes

$$\hat{\hat{C}}_u = A_u \hat{\hat{K}}^T (\hat{\hat{K}} \hat{\hat{K}}^T)^{-1} \qquad (4)$$

where $\hat{\hat{C}}_u$ represents the matrix of CLS-estimated concentrations that has been augmented with corresponding columns of parameters to estimate the least-squares contribution of each augmented shape to each prediction sample spectrum contained in the matrix of unknown sample spectra to be predicted, A_u. The PACLS algorithm applied to a single sample spectrum is depicted diagrammatically in Fig. 1. Figure 1 demonstrates the case where the CLS-estimated pure-component spectra of the molecular species (urea, creatinine, NaCl, and water) were obtained from the constant-temperature data and the spectral shape of a temperature change in the solutions is added in PACLS prediction along with an offset term and a linear term to represent a simultaneous linear

baseline fit. The augmented spectral shapes are represented by dashed lines to indicate the spectral shapes added with the PACLS prediction. Corresponding least-squares estimated concentration elements are added to the concentration vector to complete the PACLS equations. The addition of spectral shapes both changes and corrects concentration estimates relative to predictions without added spectral shapes. The $\mathbf{K}$ matrix can be augmented during the creation of the cross-validated calibration model and the augmented model can be saved for prediction of unknown samples. By including the augmentation during cross-validation, more realistic estimates of prediction ability can be obtained and outlier detection sensitivity is improved. Alternatively, the augmentation can be performed before true prediction on unknown samples. Both types of augmentation, which yield identical concentration predictions for analytes included in the calibration, will be discussed in the Results and Discussion Section.

The fact that the PACLS method can correct for inaccurate estimated pure-component spectra when sources of spectral variation are left out of the CLS model can be understood by examining either the CLS regression coefficients or the net-analyte signals (NAS).[21] The vector of p regression coefficients for each of the m components included in the CLS calibration are contained as the m columns of $\hat{\mathbf{K}}^T(\hat{\mathbf{K}}\hat{\mathbf{K}}^T)^{-1}$ and $\hat{\hat{\mathbf{K}}}^T(\hat{\hat{\mathbf{K}}}^T\hat{\hat{\mathbf{K}}}^T)^{-1}$ in Eq. 3 and Eq. 4, respectively. The predicted analyte concentration is simply the dot product of the unknown sample spectrum and the regression coefficient vector for the analyte. Each CLS prediction regression coefficient vector is proportional to the NAS for the corresponding analyte. Therefore, both the CLS regression coefficient vector and the NAS for a given component represent that portion of the analyte spectral signal that is orthogonal to all other sources of spectral variance. The NAS is the only

portion of the analyte signal that is available for prediction.[22] We will demonstrate in the Results and Discussion Section that the regression coefficients for a given analyte are identical when 1) all the sources of spectral variation are included in the CLS calibration or when 2) sources of spectral variation are left out of the calibration but the spectral shapes of missing sources of spectral variation are added during the PACLS prediction. Since the NAS's are proportional to the regression coefficients, the same equivalence in the two cases is also true for the net analyte signals. The addition of the proper spectral shapes during PACLS prediction, therefore, corrects the regression coefficients for their absence during CLS calibration. Since the regression coefficient vector is corrected by this procedure, clearly the PACLS concentration prediction estimates will also be corrected. Empirical demonstration of this fact will be made in the Results and Discussion Section.

RESULTS AND DISCUSSION

Figure 2a and 2b present NIR spectra and the corresponding mean-centered spectra, respectively, of all variable-temperature samples for the entire data set. Much of the broad baseline variation present in the spectra is due to spectrometer drift during the day. The effect of the 5 °C temperature variation on the spectra is as great as the sum of all the chemical component changes in the samples. The spectra of the constant-temperature data have been presented previously.[18] The constant-temperature spectra exhibit a somewhat smaller magnitude of system drift spectral variations either due to the shorter time required for obtaining the constant-temperature data or due to better spectrometer stability during the day the constant-temperature data were collected.

The cross-validated CLS calibration prediction results for the constant- and variable-temperature data were calculated. During the cross-validation procedure, a single sample spectrum was removed during each rotation of the cross validation. Time of data collection was included in the CLS calibration to compensate for the linear portion of drift in the system with time,[14] and a quadratic spectral baseline fit was included in the prediction phase of the CLS analysis. Unlike the factor analysis methods of PLS and PCR, CLS does not require cross validation for factor selection. However, cross validation is desirable when implementing CLS calibration both to improve outlier detection and to obtain more realistic estimates of CLS prediction ability. The cross-validated prediction results (cross-validated standard error of prediction (CVSEP) and squared correlation coefficient (R^2)) for all three analytes in the constant- and variable-temperature data are given in Table I. The cross-validated results for temperature are also presented in Table I for the variable-temperature data. The cross-validated CLS calibration prediction results for the constant-temperature data are presented for urea in Fig. 3. The prediction results in Table I and Fig. 3 are not as precise as those achieved with PLS[14] since insertion errors, nonlinear components of spectrometer drift, small uncontrolled temperature variations (±0.1°C), and potential nonlinearities are not explicitly included in the CLS analysis. However, the purpose of this study was to demonstrate prediction improvements using PACLS compared to CLS. Future papers will demonstrate how to achieve CLS predictions that are competitive with PLS methods.

Figure 4 shows the cross-validated prediction results for a CLS analysis of the constant-temperature data when only urea concentration and time of data collection (to compensate for linear drift) are included in the concentration matrix during calibration.

The poor prediction results demonstrate why CLS has not been used when all interfering analyte concentrations are not available during calibration. Table II gives the cross-validated CLS predictions for each analyte when the CLS model only includes time of data collection and concentrations for the single analyte being predicted. Cross-validated prediction results are very poor for all three analytes. However, we can use the new PACLS algorithm to improve predictions if the spectral shapes of the various analytes whose concentrations are left out of the model can be obtained independently and added during prediction.

Estimates of these analyte spectral shapes were derived from the variable-temperature data using a CLS calibration model that includes all chemical components along with sample temperature and time of data collection. The CLS-estimated pure-component spectral shapes for the three analytes, the water solvent, and temperature are shown in Fig. 5. It is interesting to note that the estimated pure-component spectrum of NaCl is due to the interaction of NaCl with the solution rather than due to spectral features of NaCl since NaCl is ionic in solution. Yet this interaction is both adequate and sufficiently unique for accurate CLS predictions to be obtained for NaCl. Urea and creatinine have their own spectral features due to molecular vibrations of the molecules, but they also interact with the water solvent to yield additional spectral features in the CLS estimated pure-component spectra.

The spectral shapes of the two analytes and water solvent left out of the CLS calibrations were added to a PACLS prediction step during cross-validation of the constant-temperature data. The improved PACLS predictions are presented in Table II next to the prediction results of the deficient CLS models. The results in Table II

demonstrate that the concentration predictions are corrected with the addition of estimated pure-component shapes in the PACLS algorithm. These PACLS predictions are comparable to predictions obtained with standard CLS using all component concentrations (See Table I). Thus, we have empirical evidence that the concentration-deficient CLS models can be corrected with the PACLS algorithm by the addition of experimentally derived spectral shapes of the components whose concentrations were left out of the CLS models.

Spectral shapes representing the effect of the analytes on the solution spectra could be more readily obtained by spiking a calibration sample with the analyte and performing a CLS analysis (Eq. 2), on the spectra before and after spiking. Alternatively, the pure-component estimate could be obtained by subtracting the spectrum of the sample without spiking from the spectrum obtained after spiking the sample. In this latter case, the difference spectrum will be the spectral shape of the analyte with displacement of the solution. If the spectral shape added during the PACLS analysis is this difference spectrum, then all data should be mean centered during CLS calibration.

Figure 6 compares the CLS-estimated pure-component spectrum of urea when all component concentrations are included in the CLS analysis (spectrum **b**) to the spectrum when only urea concentrations are included in the analysis (spectrum **a**). Clearly, the urea pure-component spectrum **a** is contaminated by the unmodeled spectral variation of the other interfering spectral components. Since water is the dominant component in these calibration samples, the CLS-estimated urea pure-component spectrum **a** is almost identical to that of water. Two calculations of the net analyte spectra for urea are also included in Fig. 6. One calculation is based on CLS predictions with all component

concentrations and times of data collection included in the CLS model. The other calculation is for PACLS where only urea concentrations are included in the CLS model, but CLS estimates of the creatinine, NaCl, H_2O, temperature, and linear drift pure-component spectra from the variable temperature solution data were included in the PACLS model. The NAS vector is the same within the numerical precision of the calculation in both cases (NAS vectors are displaced for clarity). Clearly, if the NAS of urea is the same in each case, then predicted urea concentrations must be identical in each case. Thus, the effects of the inaccurate shape of urea in the second case are exactly corrected by the addition of the appropriate spectral shapes during the PACLS prediction. It is interesting to note that the NAS and regression coefficients are not affected by the magnitude of the shapes added during prediction augmentation. Therefore, quantitative determination of the added spectral shapes is not required.

The prediction results discussed above make it clear that the PACLS method reduces the restriction that the concentrations of spectrally active species must be known during CLS calibration. A further advantage of the PACLS algorithm is its ability to accommodate the presence of unmodeled components that may appear in the prediction samples that were not present during calibration. This advantage of PACLS over the standard CLS algorithm can be demonstrated for the case where a constant-temperature CLS model is applied to the spectra of samples of varying temperature.

Figure 7 shows the prediction results for urea when a conventional CLS model built upon the constant-temperature data is applied to the spectra of the samples collected several days later at variable temperatures between 20 and 25°C. Table III presents the prediction results for all three analytes in this case where the constant-temperature CLS

model is applied to the variable-temperature spectra. Included in Table IV are the standard error of prediction (SEP), the bias-corrected SEP (BCSEP), the bias, and the squared correlation coefficient (R^2) for the prediction results. The predictions exhibit a significant bias and loss of precision due primarily to the spectral variations of the unmodeled temperature spectral component. The advantage of the PACLS method in true prediction mode can be demonstrated by using this same example.

We should be able to improve the CLS prediction results by augmenting the PACLS prediction step with the spectral shape of a temperature change. Table IV summarizes the prediction results for urea, creatinine, and NaCl when the spectral shapes of a temperature change and linear drift estimated from the independently obtained variable-temperature pure-water solvent data (20° to 25°C temperature range) are added to $\hat{\mathbf{K}}$ from the constant-temperature CLS calibration. The PALCS prediction results in Table IV indicate significant improvements in prediction over those in Table III, but the bias and prediction precision are not as good as the original CLS cross-validated calibration predictions of the constant- or variable-temperature data (see Table I). However, another source of spectral variation not included in this temperature-augmented PACLS model is the variation due to unmodeled long-term drift between the two sets of data and the short-term drift during the 9-hour collection of the variable-temperature solution data. The absence of these sources of spectral variation in the PACLS model causes inflated prediction errors.

In order to accommodate both temperature changes and short- and long-term spectrometer drift, we must take an approach that is somewhat different than presented above. The PACLS algorithm has the opportunity to model temperature variations and

complex short- and long-term drift through the use of a set of subset samples measured during the collection of both constant- and variable-temperature spectral data sets. We first select five samples that span the concentration range of the calibration and cover the 20 - 25 °C temperature range in the variable-temperature data set. Spectral differences are generated from each pair of samples measured in both the constant- and variable-temperature data sets. This set of spectral differences represents linear combinations of the effects of temperature variations and long- and short-term drift differences between measurements of the same physical specimens for different spectrometer and temperature conditions. If these spectral differences are very similar, then adding these spectral shapes could cause a matrix condition problem. We can solve any matrix condition problem by performing an eigenvector analysis of the spectral differences and retaining only the significant eigenvectors. The prediction results for PACLS obtained by adding the 5 spectral differences from the subset spectral pairs are presented in Fig. 8 for urea. The prediction results for all three analytes are given in Table V. In order to avoid overfitting, the prediction results in both Fig. 8 and Table V are based on only the 26 samples not selected as subset samples. The prediction results in Table V are now even better than the prediction abilities of the original CLS calibration model for either the constant- or variable-temperature data. The improvements over the original calibrations can be attributed to the inclusion of the effects of temperature variation and long- and short-term spectrometer drift that are present in the spectral shapes of the difference spectra added during in the PACLS analyses. These improved prediction results can be obtained without re-measuring the entire sample set at variable temperature followed by recalibration. Simply measuring the effect of temperature and drift on the sample

solution with the use of subset sample spectra with the PACLS algorithm corrects the deficient CLS model without extensive recalibration.

If the spectral shape of the unmodeled component is not present in the calibration spectra, then the PACLS method can even be used to quantify the concentration of the unmodeled component in the unknown samples. In the example presented in this paper, temperature is nearly constant (±0.1°C) in the calibration sample set but varies considerably in the prediction sample spectra (±2.5°C). The elements in $\hat{C}_u$ that correspond to the added shape of the effect of temperature during PACLS prediction will represent the temperature estimates for the unknown samples. PACLS estimated temperatures for the variable-temperature spectral data are plotted in Fig. 9 as a function of the measured reference temperature for these samples. In order to account for spectrometer drift while also predicting temperature, a set of 5 sample spectra measured at 23°C were taken from both the constant- and variable-temperature data sets. The corresponding difference spectra of the respective samples in the two data sets yield the spectral shapes needed to correct for the system drift. The spectral shape of temperature changes was obtained from the variable-temperature spectra of pure water. Both the spectral shape of a temperature change in water and the drift spectral shapes from the 23°C subset samples were added during the PACLS prediction of the variable-temperature spectra in order to both correct for temperature variations and to predict temperature. As before, the spectra of the five subset samples were not predicted to avoid potential overfitting of the data. The PACLS SEP for temperature demonstrated in Fig. 9 is 0.07°C, which is not much greater than the ability of the temperature controller to control the temperature of the sample solutions ($\sigma = 0.05$°C). Thus with quantitatively

obtained spectral shapes, PACLS has the added advantage that it allows the unmodeled component to be quantified. Accurate temperature estimates are obtained without the requirement for redeveloping multivariate spectral calibration models using calibration data containing temperature variations. Therefore, significant improvements in efficiency, cost, and time accrue from the new PACLS method. If temperature had significantly varied during the calibration and sample temperatures were not included in the C matrix for the CLS calibration model, then temperature predictions would not be accurate with the PACLS method because the temperature variations would significantly contaminate all the pure-component spectra. Temperature predictions based solely on the pure temperature shape would, therefore, be in error. The ±0.1°C variation in the original calibration data apparently is not sufficiently large to greatly degrade the PACLS temperature predictions of the 5°C variation in the variable-temperature solution data.

CONCLUSIONS

A significantly improved CLS method has been presented that greatly increases the accuracy and applicability of CLS calibration and prediction methods. The ability of the new PACLS method to correct CLS model deficiencies during CLS prediction allows increased flexibility for CLS modeling. Any source of spectral variation that is not included in the concentration matrix during CLS calibration can be accommodated in CLS prediction if the spectral shape of the unmodeled spectral variation can be obtained. Therefore, PACLS is ideally suited to improve CLS calibration models when 1) spectrally active species are present in the calibration but their concentrations are not known or included during calibration, 2) unmodeled components are present in the

unknown prediction samples, 3) maintaining a calibration on a single spectrometer, 4) transferring a calibration model between spectrometers, or 5) to correct for changes in spectrally active purge-gas components. The source of each spectral interferent must be identified and its spectral shape obtained and included in the PACLS prediction. Generally the spectral shape of the interferent is obtained empirically, but its shape could also be obtained from other sources, e.g., spectral libraries. Greatest accuracy is expected if the spectral shape is derived from the same spectrometer as used when collecting calibration or unknown sample spectra.

It is also possible that the PACLS algorithm can be used to accommodate nonlinearities. Modeling nonlinearities could be accomplished by varying the analyte over a range of concentrations in a representative sample. Quadratic or higher order concentration terms can be added to the concentration matrix in the CLS estimate of the analyte pure component to approximate the effects of the nonlinearity. Alternatively, interaction terms (e.g., concentration cross-product terms) can be included in the concentration matrix of samples spiked with variable amounts of the analyte. Thus, pure-component spectra of the interaction terms, quadratic terms, etc. can be obtained and added during PACLS prediction.

Since the PACLS predictions are independent of the magnitude of the added shapes, the augmented shapes do not have to be obtained quantitatively. In addition, linear combinations of the augmented shapes can be used in PACLS. This latter fact allowed us to use the subset difference spectra that contained linear combinations of temperature and drift variations in varying proportions. In order for linear combinations of spectral shapes to perform properly in the PACLS algorithm, there should be at least as

many vectors available for augmentation as the number of underlying sources of unmodeled spectral variation and these underlying variations must be present in varying proportions in the added vectors. The ability to use linear combinations of spectral shapes in the PACLS algorithm allows us to propose other important applications of the PACLS method. For example, PACLS might be used to accommodate spectrometer drift by the use of multiple spectra obtained from a repeat sample. If the repeat sample is constant in concentration, then changes in the spectra of multiple repeated measurements of the sample will represent linear combinations of both spectral variations due to spectrometer drift and insertion effects. Since spectrometer drift can interact with the sample spectrum, the effect of drift on the spectra can be sample dependent. Thus, it might be preferable to have the repeat sample at the target concentration of the calibration. Alternatively, multiple repeat samples at varying concentrations could be measured. Mean-centered repeat sample spectra obtained separately for each repeat sample will represent the spectral shapes of the effect of spectrometer drift on different sample spectra. Since linear combinations of the spectral shapes can also be used in PACLS independent of their magnitude, noise filtering of the spectral shapes might be accomplished by performing an eigenvector analysis of the repeat spectra. The highest signal-to-noise ratio eigenvectors can be selected as the spectral shapes to add during PACLS prediction to optimally model spectrometer drift.

Of course the accuracy of PACLS predictions will be degraded if there are errors in the empirically derived spectral shapes. The effect of errors in the empirical shapes on the PACLS predictions will be investigated in the future through Monte Carlo simulations. In addition, each shape added will reduce the net analyte signal (unless it is

orthogonal to the shape of the analyte) which will serve to degrade analysis sensitivity. However, the degradation is expected to be no more than if the same source of spectral variation were included in the original calibration.

We have implemented the software such that the spectral shapes can be added both in cross validation as well as in true prediction. Adding all known shapes during cross-validated calibration allows the most realistic estimates of PACLS prediction ability to be determined. In addition, outlier detection is enhanced since spectral residuals are reduced when adding spectral shapes during cross-validated calibrations. Smaller spectral residuals will improve the spectral F tests by reducing the denominator of the F ratio making the F test more sensitive to smaller changes in spectral residuals. Finally, the PACLS can be made even more useful and flexible if the spectral and concentration residuals in CLS calibration are passed to a PLS algorithm to model all those sources of spectral variation that are not included in the CLS calibration or the PACLS predication. The resulting PACLS/PLS hybrid algorithm offers a potentially significant improvement to the PACLS algorithm that will be described and demonstrated in future papers.

ACKNOWLEDGMENTS

The authors would like to acknowledge Howland D. T. Jones for making the samples and collecting the spectral data and Edward V. Thomas for providing the Latin Hypercube experimental design for the sample concentrations.

[1] M. K. Antoon, J. H. Koenig, and J. L. Koenig, Appl. Spectrosc. **31**, 518 (1977).

[2] D. M. Haaland and R. G. Easterling, Appl. Spectrosc. **34**, 539 (1980).

[3] D. M. Haaland and R. G. Easterling, Appl. Spectrosc. **36**, 665 (1982).

[4] D. M. Haaland, R. G. Easterling, and D. A. Vopicka, Appl. Spectrosc. **39**, 73 (1985).

[5] D. M. Haaland, "Multivariate Calibration Methods Applied to Quantitative FT-IR Analyses," in Practical Fourier Transform Infrared Spectroscopy, J. R. Ferraro and K. Krishnan, Eds., (Academic Press, New York, 1989) Chap. 8, pp. 396-468.

[6] D. M. Haaland, "Methods to Include Beer's Law Nonlinearities in Quantitative Spectral Analysis," in *Computerized Quantitative Infrared Analysis*, ASTM Special Technical Publication, G. L. McClure, Ed., STP Vol. **934**, American Society of Testing Materials, Philadelphia, Pennsylvania, 1987, p. 78.

[7] P. Saarinen and J. Kauppinen, Applied Spectroscopy **45**, 953 (1991).

[8] P. Jaakkola, J. D. Tate, M. Paakkunainen, J. Kauppinen, and P. Saarinen, Applied Spectroscopy **51**, 1159 (1997).

[9] W. Lindberg, J.-A. Persson, and S. Wold, Anal. Chem. **55**, 643 (1983).

[10] D. M. Haaland and E. V. Thomas, Anal. Chem. **60**, 1193 (1988).

[11] D. M. Haaland and E. V. Thomas, Anal. Chem. **60**, 1202 (1988).

[12] P. M. Fredericks, J. B. Lee, P. R. Osborn, and D. A. Swinkels, App. Spectrosc. **39**, 303 (1985).

[13] D. W. T. Griffith, Appl. Spectrosc. **50**, 59 (1996).

[14] D. M. Haaland, L. Han, and T. M. Niemczyk, Appl. Spectrosc. **53**, 390 (1999).

[15] B. R. Stallard, R. K. Rowe, M. J. Garcia, D. M. Haaland, L. H. Espinoza, and T. M. Niemczyk, "Trace Water Vapor Determination in Corrosive Gasses by Infrared Spectroscopy," Sandia Report SAND93-4026 (Sandia National Laboratories, Albuquerque, NM), December, 1993.

[16] D. M. Haaland, W. B. Chambers, M. R. Keenan, and D. K. Melgaard, "Improved Multivariate Calibration Methods for Quantitative ICP-AES Analyses," submitted to Appl. Spectrosc. (1999).

[17] D. M. Haaland and H. D. T. Jones, "Multivariate Calibration Applied to Near-Infrared Spectroscopy for the Quantitative Analysis of Dilute Aqueous Solutions," in *Proceedings of the 9th International Conference on Fourier Transform Spectroscopy*, J. E. Bertie and H. Wisser, Eds., SPIE Vol. 2089 (SPIE, Bellingham, Washington, 1993), p. 448.

[18] D. M. Haaland, "Synthetic Multivariate Models to Accommodate Unmodeled Interfering Spectral Components During Quantitative Spectral Analyses," accepted for publication in Appl. Spectrosc. (Feb. 2000).

[19] H. Mark, Anal. Chem. **59**, 790 (1987).

[20] C. L. Lawson and R. J. Hanson, "*Solving Least Squares Problems*," Prentice-Hall, Englewood Cliffs, NJ (1974).

[21] A. Lorber, Anal. Chem. **58**, 1167 (1986).

Table I. Cross-validated CLS predictions of analytes in constant- and variable-temperature calibration data including all analytes, water, and time in the CLS calibration.

	Constant-Temperature Data		Variable-Temperature Data	
Analyte	CVSEP (mg/dL)	R^2	CVSEP (mg/dL)	R^2
Urea	66	0.9954	60	0.9962
Creatinine	54	0.9960	36	0.9982
NaCl	59	0.9959	51	0.9978
Temperature	NA	NA	0.11 °C	0.9961

Table II. Cross-validated CLS predictions of analytes in constant-temperature calibration data including only single analyte and time in the CLS calibration vs. PACLS with other two analytes and water pure components added.

	CLS		PACLS	
Analyte	CVSEP (mg/dL)	R^2	CVSEP (mg/dL)	R^2
Urea	603	0.628	62	0.9960
Creatinine	1386	0	58	0.9956
NaCl	324	0.880	54	0.9966

Table III. CLS predictions of analytes in variable-temperature data using a constant-temperature CLS model that included all three analytes, water, and time in the calibration.

	Variable-Temperature Predictions			
Analyte	SEP (mg/dL)	BCSEP	Bias	R^2
Urea	599	498	345	0.602
Creatinine	187	115	-149	0.948
NaCl	163	160	44	0.967

Table IV. PACLS predictions of variable-temperature spectra using a PACLS model that includes all three analytes, water, and time in the CLS calibration with the PACLS prediction step augmented by the CLS estimates of temperature and linear drift spectral shapes from variable-temperature pure-water solvent data.

	Variable-Temperature Predictions			
Analyte	SEP (mg/dL)	BCSEP	Bias	R^2
Urea	60	60	11	0.9960
Creatinine	108	82	-71	0.9830
NaCl	94	60	-73	0.9889

Table V. PACLS predictions of 26 variable-temperature spectra using a PACLS model that includes all three analytes, water, and time in the CLS calibration with the PACLS prediction step augmented by the spectral shapes of the 5 subset difference spectra (variable-temperature minus constant-temperature spectra).

	Variable-Temperature Predictions			
Analyte	SEP (mg/dL)	BCSEP	Bias	R^2
Urea	36	36	-10	0.9985
Creatinine	27	27	5	0.9990
NaCl	47	47	2	0.9971

FIGURE CAPTIONS

Figure 1. Schematic diagram of the PACLS prediction for a single unknown sample spectrum. Augmented concentrations are noted with underlining. Augmented spectral shapes are represented as dashed lines. Note that the spectra on the right-hand-side of the diagram are presented as columns (vertical) since the matrix is transposed.

Figure 2. Spectra of the 31 variable-temperature samples, a) absorbance spectra and b) mean-centered absorbance spectra.

Figure 3. Cross-validated CLS predictions for urea for the 31 constant-temperature samples including urea, creatinine, NaCl, H_2O, and time of data collection in the CLS calibration. Solid line is line of identity.

Figure 4. Cross-validated CLS predictions for urea for the 31 constant-temperature samples including only urea and time of data collection in the CLS calibration. Solid line is line of identity.

Figure 5. CLS-estimated pure-component spectra from the 31 variable-temperature samples including urea, creatinine, NaCl, H_2O, temperature, and time of data collection in the CLS calibration. The pure-component spectra are: **a)** H_2O, **b)** urea, **c)** creatinine, **d)** NaCl, and **e)** temperature. The four analytes are estimates of the pure-component spectra at concentrations of 1 mg/dL and the temperature pure-component estimate represents the spectral changes in the solution for a 1 millidegree C change in temperature.

Figure 6. CLS estimated pure-component spectrum of urea from variable-temperature data for **a)** case with only urea concentrations included in the CLS calibration (scaled by 0.01), **b)** case with urea, creatinine, NaCl, H2O, and time of data collection included in

the CLS calibration. c) Net analyte signal calculated for urea using urea pure component in Fig. 6a and d) Net analyte signal calculated for urea using urea pure component in Fig. 6b. All spectra except d are shifted for clarity.

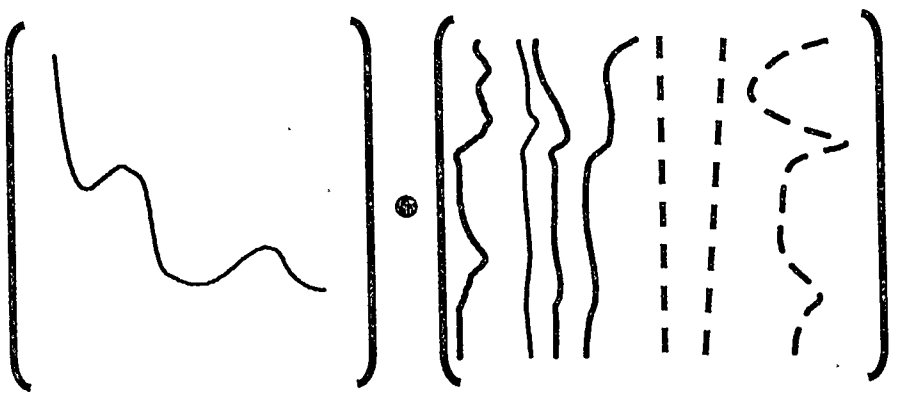
Figure 7. CLS predictions for urea for the 31 variable-temperature samples based on a CLS calibration of the 31 constant-temperature sample data including urea, creatinine, NaCl, H_2O, and time of data collection in the CLS calibration. Solid line is line of identity.

Figure 8. PACLS predictions for urea for the 31 variable-temperature samples based on a CLS calibration of the 31 constant-temperature sample data including urea, creatinine, NaCl, H_2O, and time of data collection in the CLS calibration and 5 subset sample spectral differences added in PACLS prediction. Solid line is line of identity.

Figure 9. PACLS predictions for temperature for the 31 variable-temperature samples based on a CLS calibration of the 31 constant-temperature sample data including urea, creatinine, NaCl, H_2O, and time of data collection in the CLS calibration and 5 subset sample spectral differences at 23 °C and the temperature of water pure component added in PACLS prediction. Solid line is line of identity.

CLS model: $\mathbf{a}_u = \mathbf{c}_u \tilde{\hat{\mathbf{K}}} + \mathbf{e}_a$

CLS prediction: $\hat{\mathbf{c}}_u = \mathbf{a}_u \tilde{\hat{\mathbf{K}}}^T (\tilde{\hat{\mathbf{K}}} \tilde{\hat{\mathbf{K}}}^T)^{-1}$

$[\hat{c}_1\ \hat{c}_2\ \hat{c}_3\ \hat{c}_4\ \hat{c}_5\ \hat{c}_6\ \hat{c}_7\ ] =$  $\cdots$

$\hat{\mathbf{c}}_u = \mathbf{a}_u \cdot \tilde{\hat{\mathbf{K}}}^T \cdots$

Figure 1

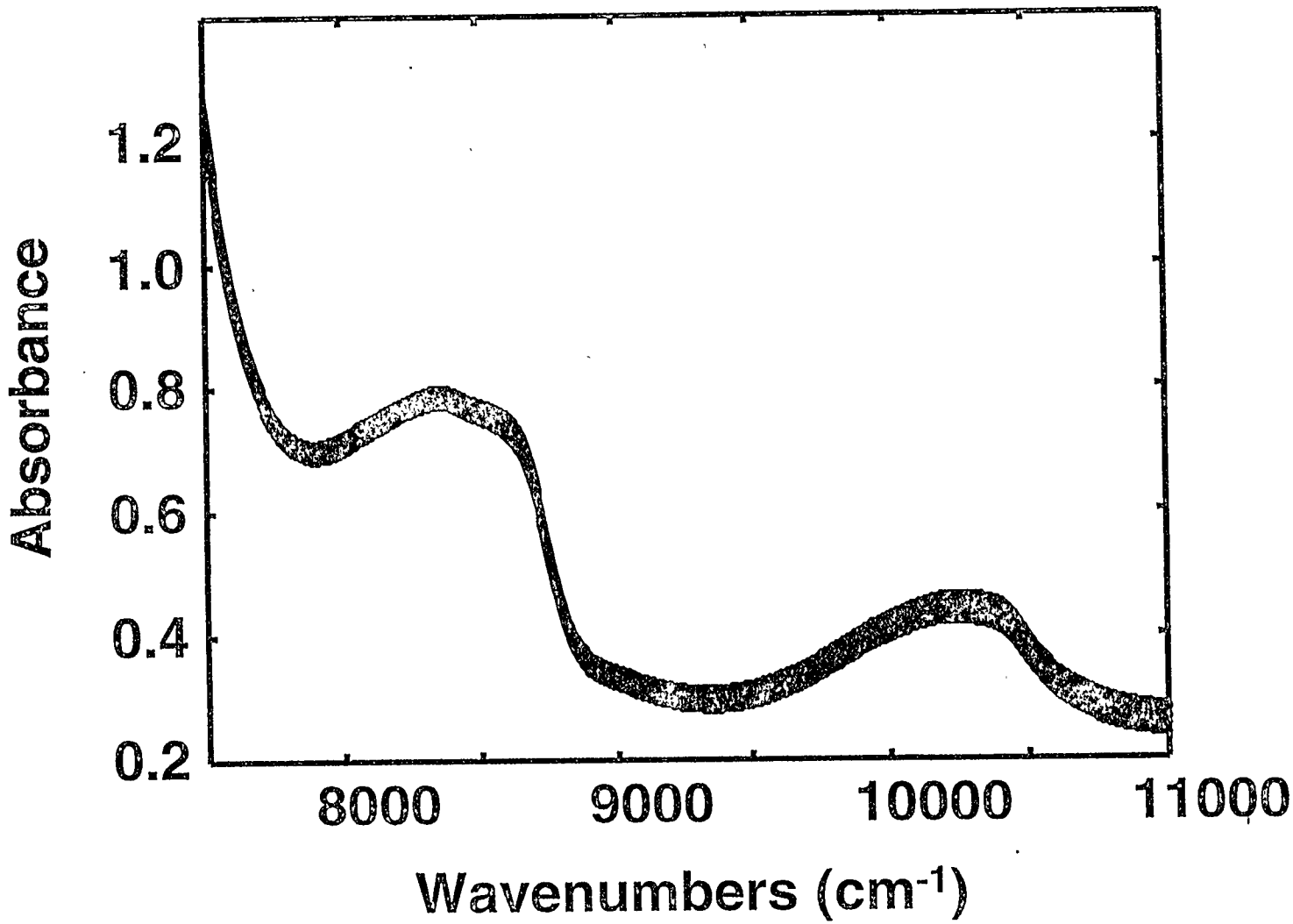

Absorbance
1.2
1.0
0.8
0.6
0.4
0.2
8000
9000
10000
11000
Wavenumbers (cm-1)

Figure 2a

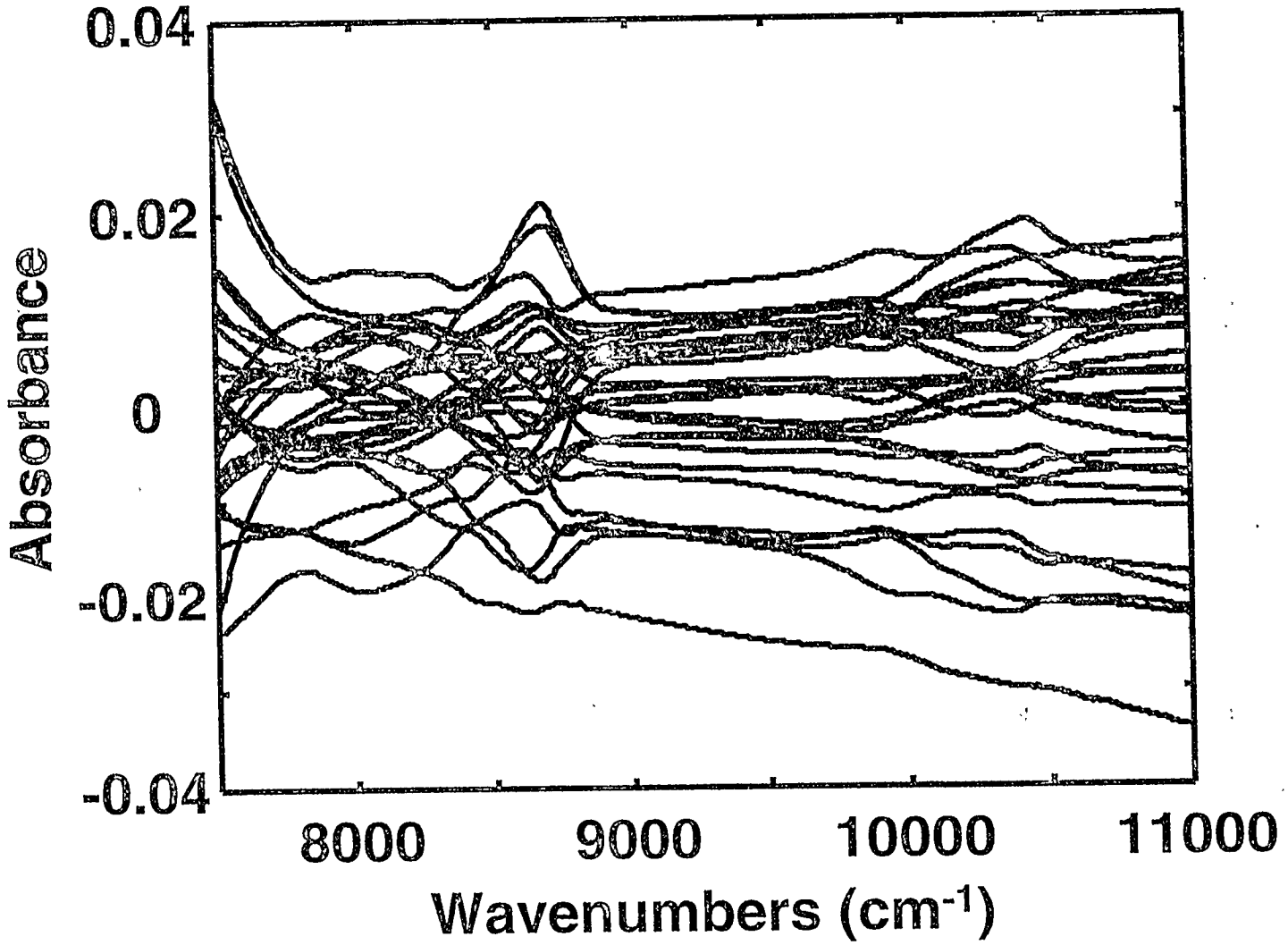

0.04
0.02
0
-0.02
-0.04
Absorbance
8000
9000
10000
11000
Wavenumbers (cm-1)

Figure 2b

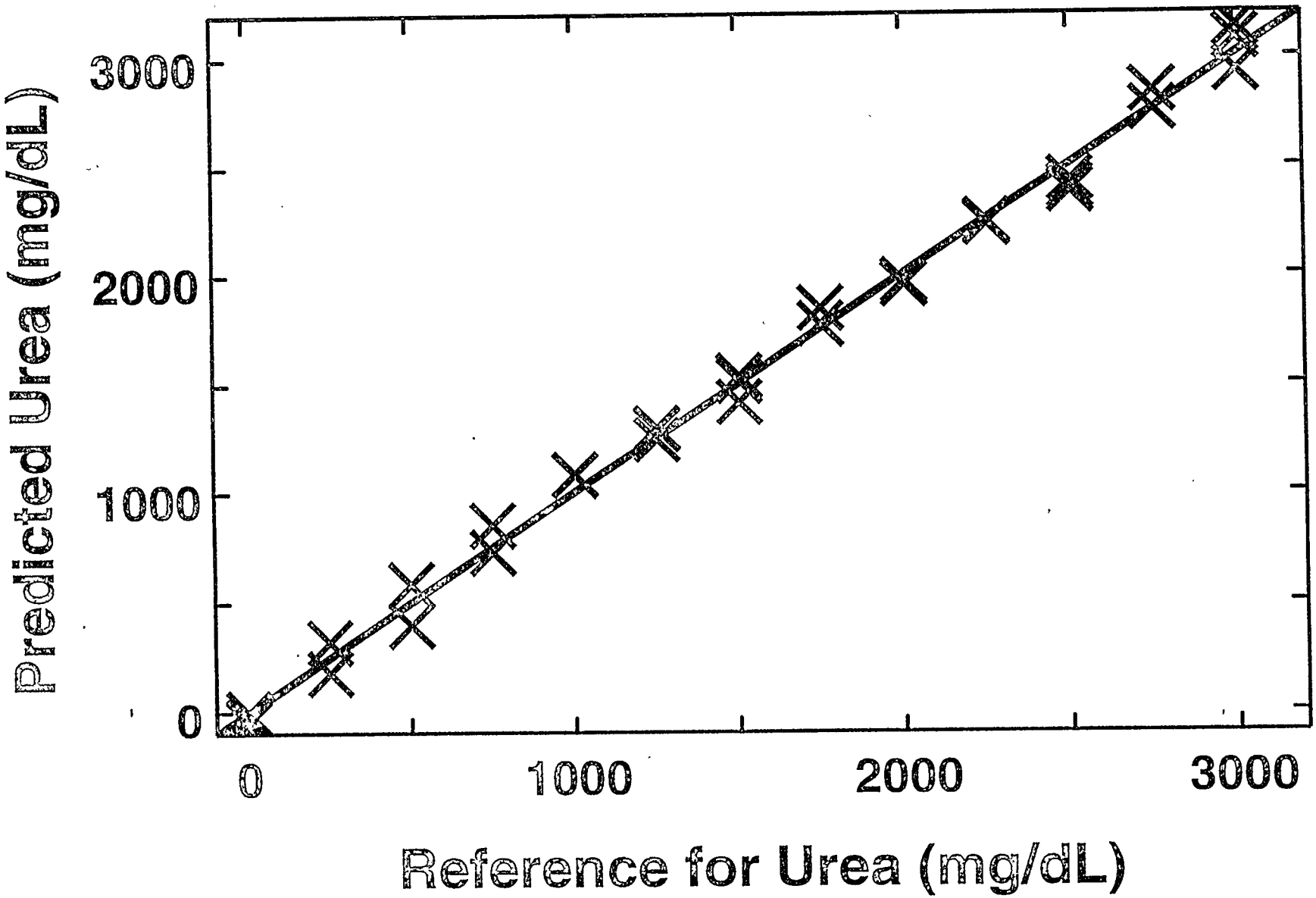

Predicted Urea (mg/dL)
3000
2000
1000
0
0
1000
2000
3000
Reference for Urea (mg/dL)

Figure 3

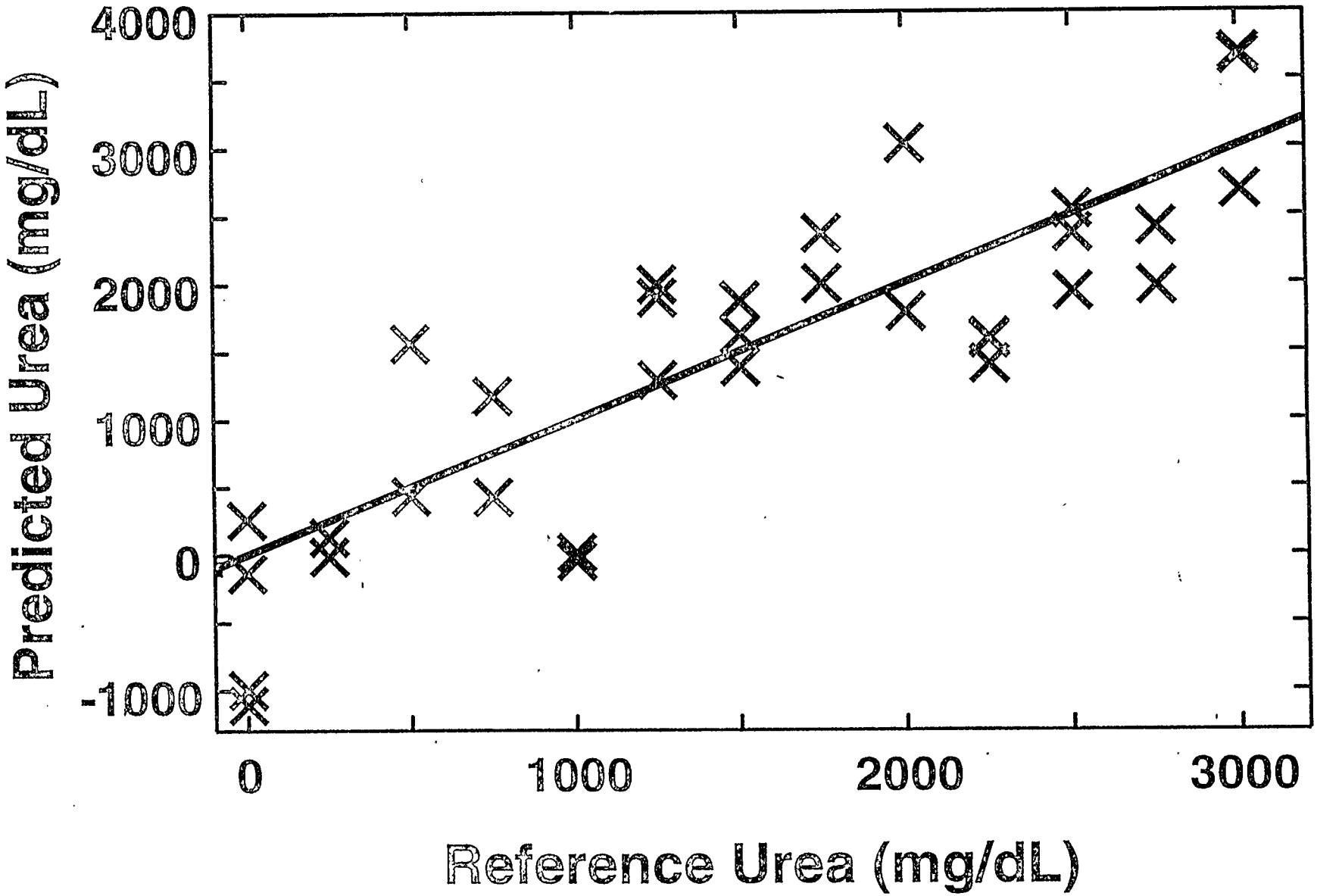

Predicted Urea (mg/dL)
4000
3000
2000
1000
0
-1000
0
1000
2000
3000
Reference Urea (mg/dL)

Figure 4

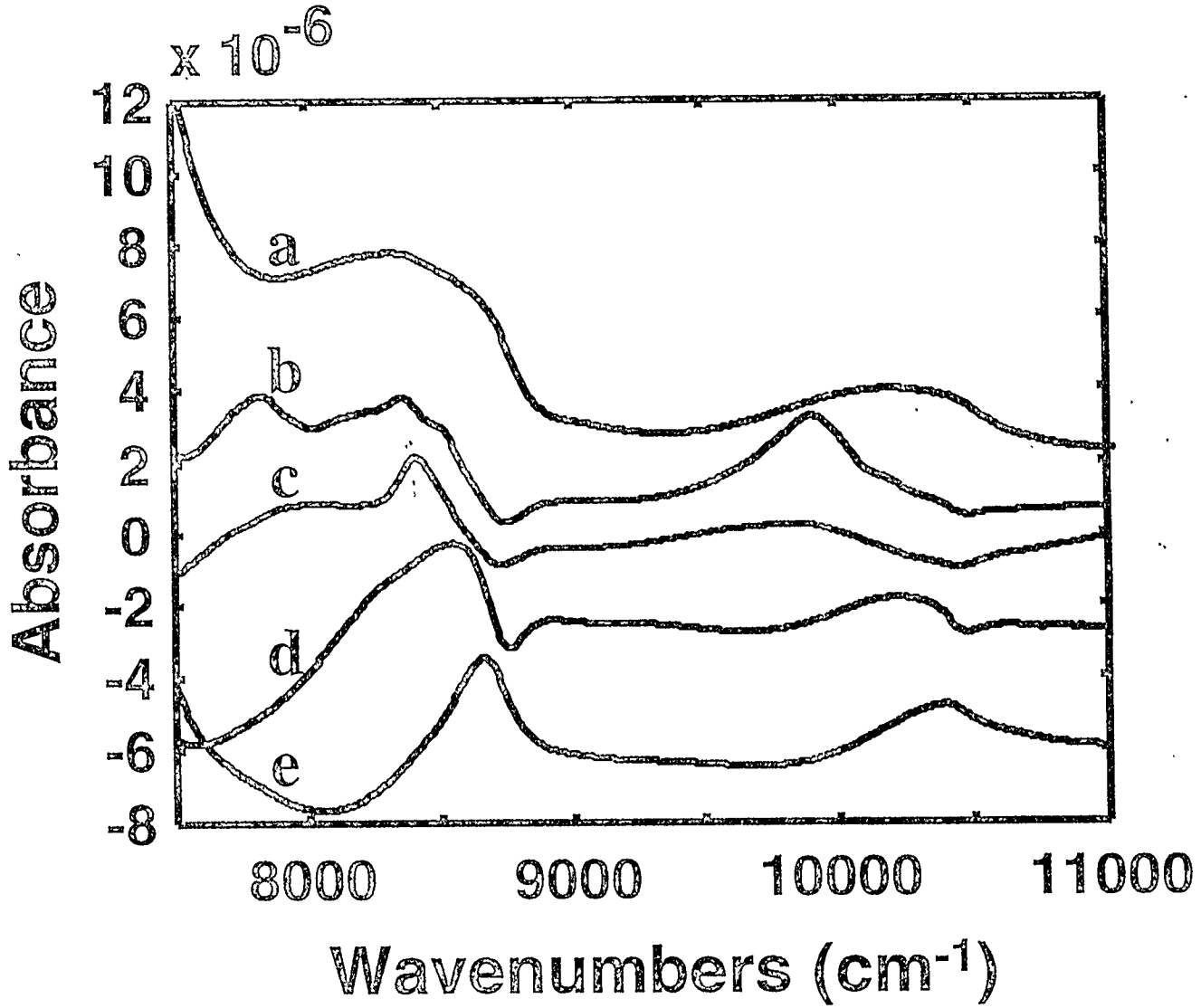

x 10-6
12
10
8
6
4
2
0
-2
-4
-6
-8
Absorbance
a
b
c
d
e
8000
9000
10000
11000
Wavenumbers (cm-1)

Figure 5

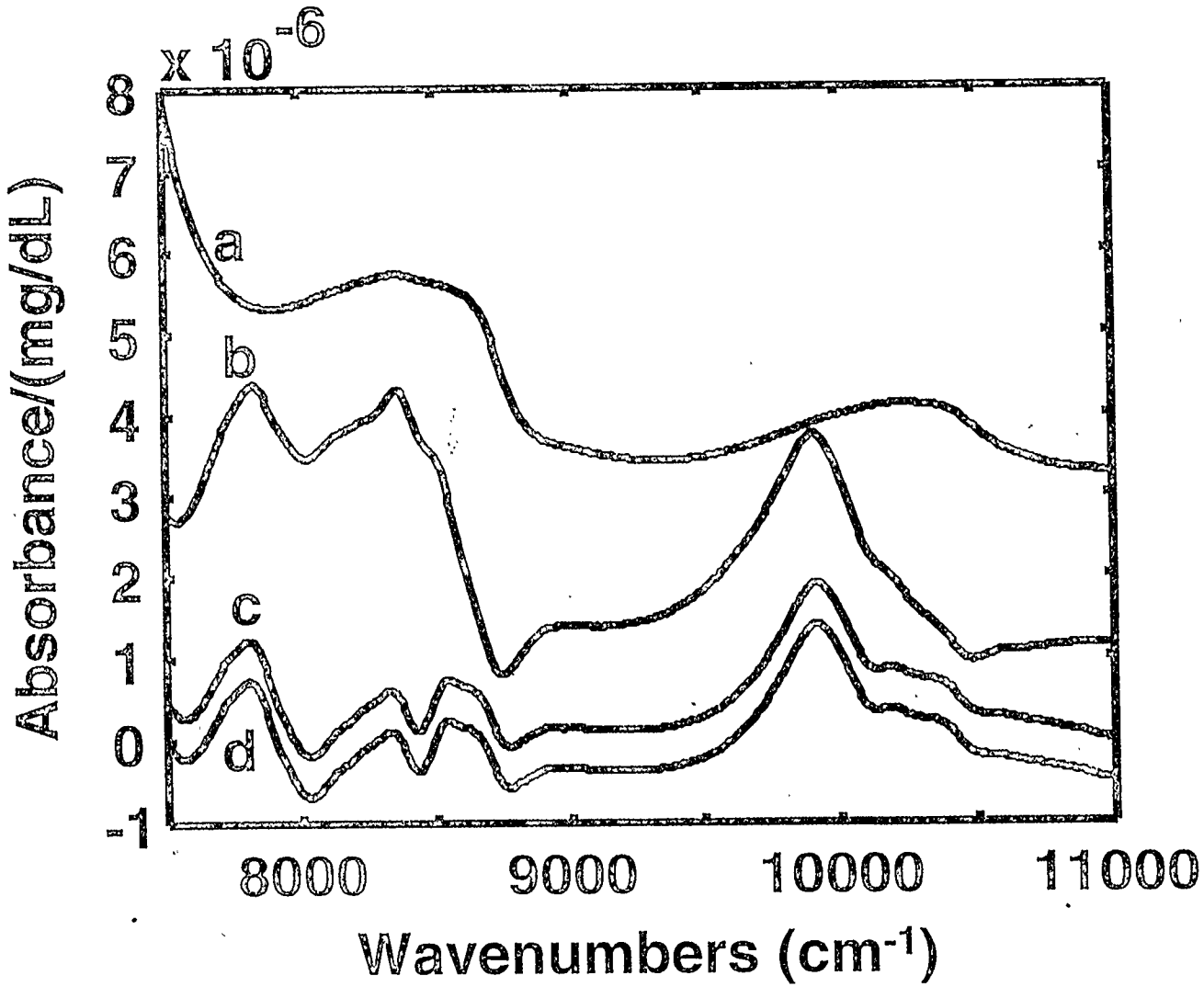

x 10-6
8
7
6
5
4
3
2
1
0
-1
Absorbance/(mg/dL)
a
b
c
d
8000
9000
10000
11000
Wavenumbers (cm-1)

Figure 6

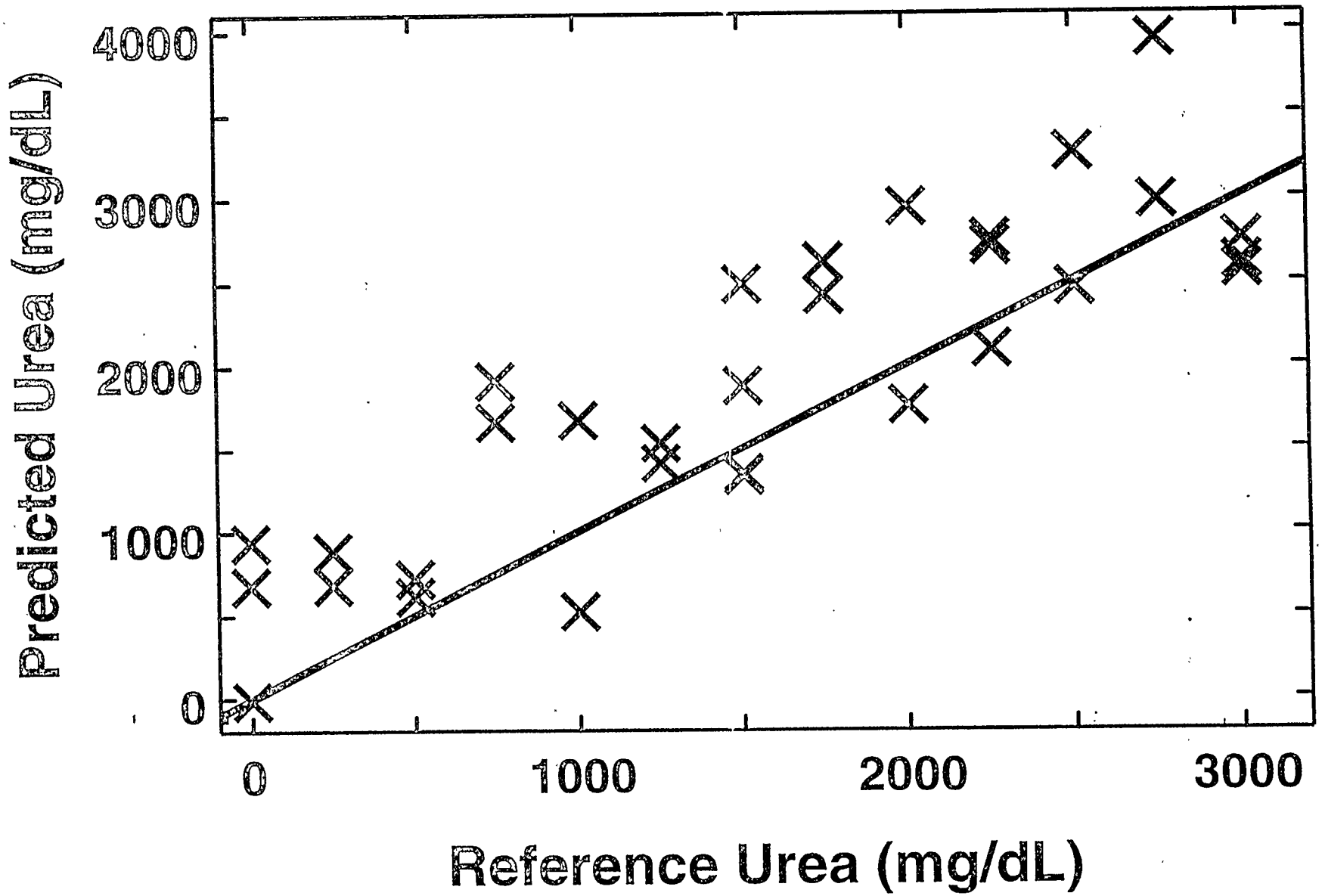

Predicted Urea (mg/dL)
4000
3000
2000
1000
0
0
1000
2000
3000
Reference Urea (mg/dL)

Figure 7

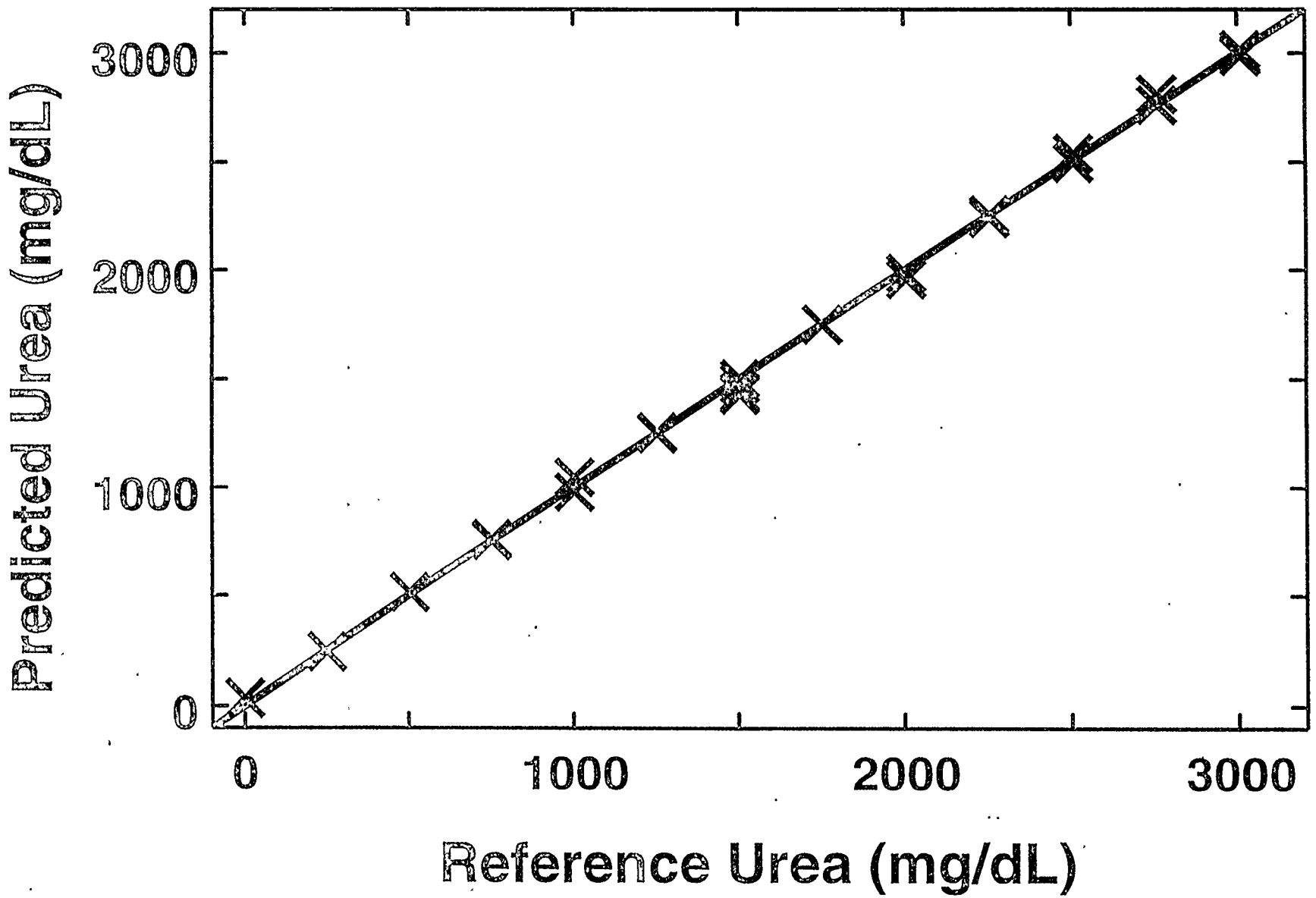

Predicted Urea (mg/dL)
3000
2000
1000
0
0
1000
2000
3000
Reference Urea (mg/dL)

Figure 8

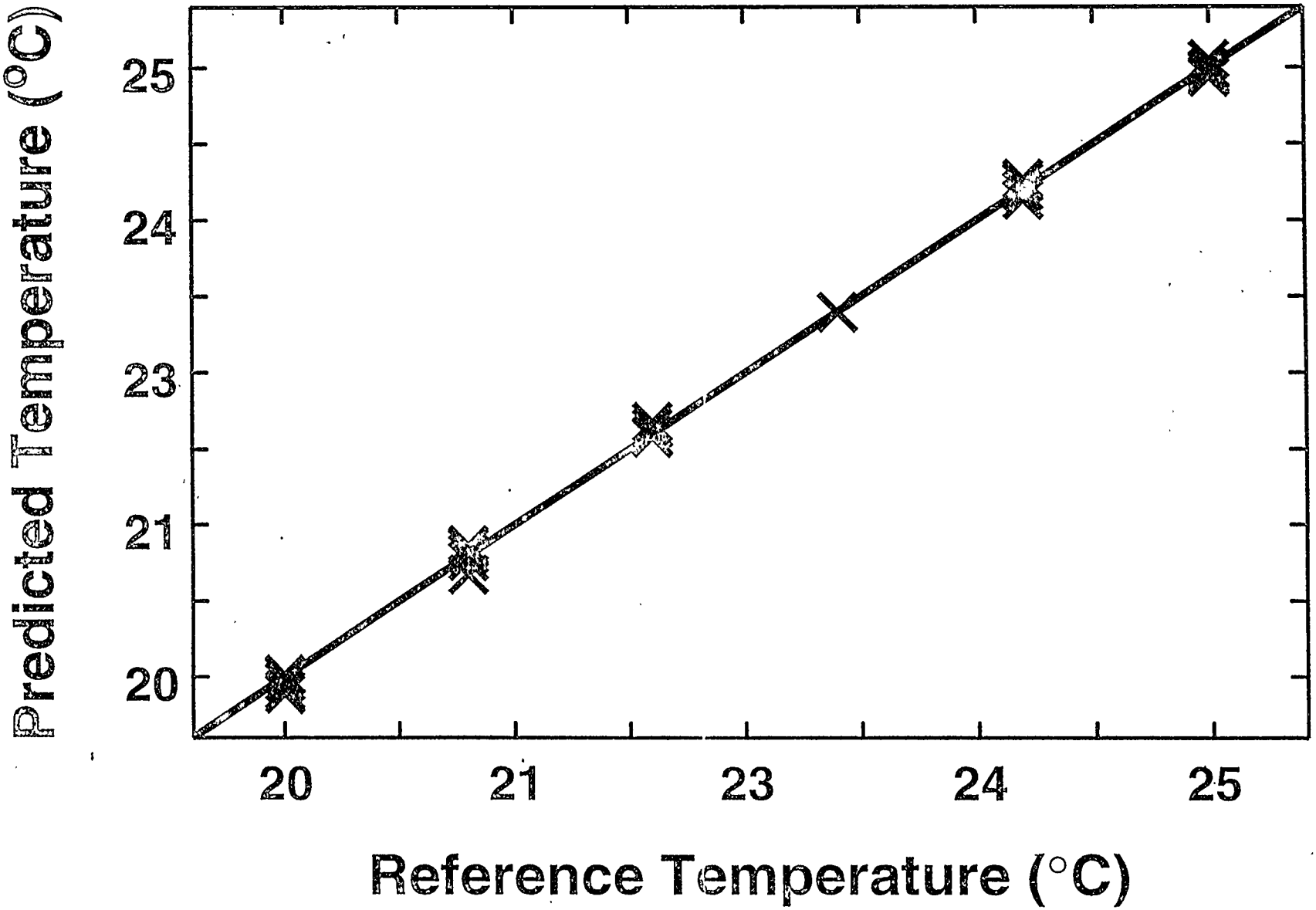

Predicted Temperature (°C)
25
24
23
21
20
20
21
23
24
25
Reference Temperature (°C)

Figure 9