Anal. Chem. **2007,** *79,* 234−239

Constrained Regularization: Hybrid Method for Multivariate Calibration

Wei-Chuan Shih,* Kate L. Bechtel, and Michael S. Feld

G. R. Harrison Spectroscopy Laboratory, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Room 6-014, Cambridge, Massachusetts 02139

We present a hybrid multivariate calibration method, constrained regularization (CR), and demonstrate its utility via numerical simulations and experimental Raman spectra. In this new method, multivariate calibration is treated as an inverse problem in which an optimal balance between model complexity and noise rejection is achieved with the inclusion of prior information in the form of a spectral constraint. A key feature is that the constraint is incorporated in a flexible manner, allowing the minimization algorithm to arrive at the optimal solution. We demonstrate that CR, when used with an appropriate constraint, is superior to methods without prior information, such as partial least-squares, and is less susceptible to spurious correlations. In addition, we show that CR is more robust than methods in which the constraint is rigidly incorporated, such as hybrid linear analysis, when the exact spectrum of the analyte of interest as it appears in the sample is not available. This situation can occur as a result of experimental or sample variations and often arises in complex or turbid samples such as biological tissues.

Multivariate calibration is a powerful analytical technique for extracting analyte concentrations in complex chemical systems that exhibit linear response.[1−3] Multivariate techniques are particularly well suited to analysis of spectral data because information about all of the analytes can be collected simultaneously at many wavelengths. The goal of multivariate calibration is to obtain a spectrum of regression coefficients, **b**, such that an analyte's concentration, c, can be accurately predicted by taking the scalar product of **b** with a prospective experimental spectrum, **s**:

$$c = \mathbf{s}^{\mathrm{T}} \cdot \mathbf{b} \tag{1}$$

(Lowercase boldface type denotes a column vector, uppercase boldface type a matrix; and the superscript T denotes transpose.) The regression vector, **b**, is unique in an ideal noise-free linear system without constituent correlations. Under realistic experimental conditions, however, only an approximation to **b** for the experimental system of interest can be found.

Explicit and implicit multivariate calibration methods have their own advantages and limitations. Explicit calibration methods are often used when all of the constituent spectra can be individually measured or precalculated.[4] Examples are ordinary least squares (OLS) and classical least squares. Explicit methods provide transparent models with easily interpretable results. However, highly controlled experimental conditions, high-quality spectra, and accurate concentration measurements of each of the constituent analytes (or equivalent information) may be difficult to obtain, particularly in biomedical applications.

When all of the individual constituent spectra are not known, implicit calibration methods are often adopted. Principal component regression (PCR)[5] and partial least squares (PLS)[6] are two frequently used methods in this category. Implicit methods require only high-quality calibration spectra and accurate concentration measurements of the analyte of interest—the calibration data—greatly facilitating experimental design. However, unlike explicit methods, the performance of implicit methods cannot be simply judged by conventional statistical measures such as goodness of fit. As pointed out in the literature,[7] spurious effects such as system drift and covariations among constituents can be incorrectly interpreted as legitimate correlations. Furthermore, implicit methods such as PCR and PLS lack the ability to incorporate additional information beyond the calibration data about the system or analytes. Such prior information has the potential to improve implicit calibration and limit spurious correlations.

The incorporation of prior information into models has been extensively pursued in fields such as pattern recognition, machine learning, and inverse problems. The use of prior information generally helps stabilize and enhance deconvolution, classification, or inversion algorithms. In multivariate calibration, methods combining explicit and implicit schemes have been explored by Haaland,[8] Wentzell,[9] and, in our laboratory, by Berger.[10] Owing to prior information about model constituents, measurement error variance, or the analyte of interest, these methods in principle

* To whom correspondence should be addressed. E-mail: wshih@mit.edu.

(1) Martens, H.; Naes, T. *Multivariate Calibration*; John Wiley & Sons: New York, 1989.
(2) Geladi, P.; Kowalski, B. R. *Anal. Chim. Acta* **1986**, *185*, 1−17.
(3) Kowalski, B. R.; Lorber, A. *Abstr. Pap. Am. Chem. Soc.* **1988**, *196*, 100-Anyl.
(4) Haaland, D. M.; Thomas, E. V. *Anal. Chem.* **1988**, *60*, 1193−1202.
(5) Gunst, R. F.; Mason, R. L. *Regression Analysis and its Applications*, 6th ed.; Marcel Dekker Inc.: New York, 1980.
(6) Wold, S.; Martin, H.; Wold, H. *Lecture Notes in Mathematics*; Springer-Verlag: Heidelberg, 1983.
(7) Arnold, M. A.; Burmeister, J. J.; Small, G. W. *Anal. Chem.* **1998**, *70*, 1773−1781.
(8) Haaland, D. M.; Melgaard, D. K. *Appl. Spectrosc.* **2001**, *55*, 1−8.
(9) Wentzell, P. D.; Andrews, D. T.; Kowalski, B. R. *Anal. Chem.* **1997**, *69*, 2299−2311.
(10) Berger, A. J.; Koo, T. W.; Itzkan, I.; Feld, M. S. *Anal. Chem.* **1998**, *70*, 623−627.

10.1021/ac060732v CCC: $37.00
Published on Web 11/18/2006

outperform those without prior information. However, depending on how prior information is incorporated, these methods may lack robustness due to inaccuracy in the prior information, especially for methods incorporating known analyte spectra, such as hybrid linear analysis (HLA).[10]

HLA utilizes a separately measured spectrum of the analyte of interest together with the calibration data and outperforms methods without prior information such as PLS. However, because HLA relies on the subtraction of the analyte spectrum from the calibration data, it is highly sensitive to the accuracy of the spectral shape and its intensity. For complex turbid samples in which absorption and scattering are likely to alter the analyte spectral features in unknown ways, we find that the performance of HLA is impaired. Motivated by advancing transcutaneous measurement of blood analytes in vivo, we have developed a method that is more robust against inaccuracies in the previously measured pure analyte spectra.

This paper presents the new method to merge prior spectral information with calibration data in an implicit calibration scheme. Starting with the inverse mixture model as the forward problem, we define the inverse problem with solution **b**. Instabilities associated with the inversion process are removed by means of a technique known as regularization,[11] and prior information is included by means of a spectral constraint. We thus call the method constrained regularization (CR). We study the effectiveness of CR using numerical simulations and demonstrate its performance using experimental Raman spectra. We show that with CR the standard error of prediction (SEP) is lower than methods without prior information, such as PLS, and is less affected by analyte covariations. We further show that CR is more robust than our previously developed hybrid method, HLA, when there are inaccuracies in the applied constraint, as often occurs in complex or turbid samples such as biological tissues.

It should be mentioned that the terms prior information and spectral constraints are used interchangeably for both CR and HLA in this paper.

THEORY

Multivariate calibration can be viewed as an inverse problem. Regularization methods,[11] also known as ridge regression in the statistical literature,[12] are mostly used on ill-conditioned inverse problems such as tomographic imaging, inverse scattering, and image restoration. These methods seek to obtain a source distribution in the presence of noisy (system-corrupted) data. In our application, the noise is assumed to be uncorrelated, which simplifies the analysis.

Implicit calibration schemes require a set of calibration spectra, **S**, with each spectrum occupying a column of **S**, associated with several known concentrations of the analyte of interest that are expressed as a column vector, **c**, the jth element of which corresponds to the jth column of **S**. Developing an accurate regression vector, **b**, requires accurate values of **c** and **S**. The forward problem for our calibration method is defined by the linear inverse mixture model for a single analyte:

$$\mathbf{c} = \mathbf{S}^{\mathrm{T}}\mathbf{b}. \tag{2}$$

The goal of the calibration procedure is to use the set of data [**S**,**c**] to obtain an accurate **b** by inverting eq 2. The resulting **b** can then be used in eq 1 to predict the analyte concentration, c, of an independent prospective sample by measuring its spectrum, **s**. The "accuracy" of **b** is usually judged by its ability to correctly predict concentrations prospectively via eq 1.

There are two primary difficulties in directly inverting eq 2. First, the system is usually underdetermined, i.e., there are more variables (e.g., wavelengths) than equations (e.g., number of calibration samples). Thus, direct inversion does not yield a unique solution unless truncation of principal components or factors is carried out. Second, even if a pseudoinverse exists and results in a unique solution, such a solution tends to be unstable because all measurements contain noise and error. That is, small variations in **c** or **S** can lead to large variations in **b**. Therefore, a more robust solution is required.

The inversion process may be viewed in terms of singular value decomposition (SVD),[13] in which the spectra of the sample set, **S**, are decomposed into principal component directions, $\mathbf{v}_j$, with amplitudes given by their respective singular values, σ_j. Most of the information in **S** is captured in the principle components with large σ_j. The singular values with small amplitudes, although potentially important, are the main cause of instability.[14] Methods to alleviate such instabilities are based on reducing the influence of these small singular values,[14,15] accomplished by means of a regularization parameter, Λ. The regularized solution for **b** is given by

$$\mathbf{b} = \sum_{j=1}^{p} f_j \frac{\mathbf{u}_j^{\mathrm{T}}\mathbf{c}}{\sigma_j}\mathbf{v}_j \tag{3a}$$

with

$$f_j(\Lambda) = \frac{\sigma_j^2}{\sigma_j^2 + \Lambda^2} \tag{3b}$$

$\mathbf{u}_j$ and $\mathbf{v}_j$ the eigenvectors of $\mathbf{S}^{\mathrm{T}}\mathbf{S}$ and $\mathbf{S}\mathbf{S}^{\mathrm{T}}$, respectively, and p the rank of **S**. Note that for $\sigma_j \gg \Lambda$, $f_j \cong 1$, and for $\sigma_j \ll \Lambda$, $f_j \cong \sigma_j^2/\Lambda^2$. Thus, one can interpret regularization as providing a smoothing filter f_j that limits the importance of the small singular values. For $\Lambda = 0$, eq 3 reduces to the least-squares solution for **b**. In PCR, $\Lambda = 0$ and only the k largest singular values ($k < p$) are used. In Wiener filtering,[16] Λ is chosen to be the noise-to-signal ratio.

Equation 3 is the regularized solution of eq 2; i.e., no prior information is included except by forcing the solution to be finite.

(11) Tikhonov, A. N.; Arsenin, V. I. F. A. F. *Solutions of ill-posed problems*; Winston: Washington DC, 1977.
(12) Hastie, T.; Tibshirani, R.; Friedman, J. H. *The elements of statistical learning: data mining, inference, and prediction*; Springer: New York, 2001.
(13) Strang, G. *Introduction to linear algebra*, 2nd ed.; Wellesley-Cambridge Press: Wellesley, MA, 1998.
(14) Bertero, M.; Boccacci, P. *Introduction to inverse problems in imaging*; Institute of Physics Pub.: Bristol, UK; Philadelphia, PA, 1998.
(15) Hansen, P. C. *Rank-deficient and discrete ill-posed problems: numerical aspects of linear inversion*; SIAM: Philadelphia, 1997.
(16) Wiener, N. *Extrapolation, interpolation, and smoothing of stationary time series, with engineering applications*; Technology Press of the Massachusetts Institute of Technology: Cambridge, MA, 1949.

However, eq 3 can be modified to incorporate prior information. A convenient way to accomplish this is by viewing regularization as the minimization of a quadratic cost function, Φ:[14]

$$\Phi(\Lambda,\mathbf{b}_0) = ||\mathbf{S}^{\mathrm{T}}\mathbf{b} - \mathbf{c}||^2 + \Lambda||\mathbf{b} - \mathbf{b}_0||^2 \qquad (4)$$

with $||\mathbf{a}||$ the Euclidean norm (i.e., magnitude) of $\mathbf{a}$, and $\mathbf{b}_0$ a spectral constraint that introduces prior information about $\mathbf{b}$. The first term of Φ is the model approximation error, and the second term the norm of the difference between the solution and the constraint, which controls the smoothness of the solution and its deviation from the constraint. If $\mathbf{b}_0$ is zero, the solution to minimize Φ is given by eq 3. As mentioned above, for $\Lambda = 0$ the least-squares solution is then obtained. In the other limit, in which Λ goes to infinity, the solution is simply $\mathbf{b} = \mathbf{b}_0$. In the following, we adopt a calibration method in which regularization with a properly chosen spectral constraint, $\mathbf{b}_0$, is employed, hence, the name constrained regularization.

The CR solution, a generalization of eq 3, can be analytically derived in SVD form as[15]

$$\mathbf{b} = \sum_{j=1}^{p}\left\{f_j(\Lambda)\frac{\mathbf{u}_j^{\mathrm{T}}\mathbf{c}}{\sigma_j} + (1 - f_j(\Lambda))\mathbf{v}_j^{\mathrm{T}}\mathbf{b}_0\right\}\mathbf{v}_j \qquad (5)$$

A reasonable choice for $\mathbf{b}_0$ is the spectrum of the analyte of interest because that is the solution for $\mathbf{b}$ in the absence of noise and interferents. Another choice is the net analyte signal[17] calculated using all of the known pure analyte spectra. Such flexibility in the selection of $\mathbf{b}_0$ is owing to the manner in which the constraint is incorporated into the calibration algorithm. For CR, the spectral constraint is included in a nonlinear fashion through minimization of Φ, and is thus termed a "soft" constraint. On the other hand, there is little flexibility for methods such as HLA, in which the spectral constraint is algebraically subtracted from each sample spectrum before performing PCA. We term this type of constraint a "hard" constraint. In the Methods section, we use CR and HLA as examples to show that the type of constraint affects the robustness of hybrid methods concerning the accuracy of the constraint.

Once $\mathbf{b}_0$ is chosen, application of CR is straightforward, as eq 5 is a direct solution of $\mathbf{b}$ and easy to evaluate. A trial value of Λ is selected and $\mathbf{b}$ is calculated from eq 5 using leave-one-out cross-validation[12] on the calibration data set to obtain a trial prediction residual error sum of squares (PRESS):

$$\mathrm{PRESS} = \sum_{\mathrm{i}}(c_{\mathrm{i}} - \hat{c}_{\mathrm{i}})^2 \qquad (6)$$

where c_{i} and $\hat{c}_{\mathrm{i}}$ are reference and predicted concentrations, respectively, and i denotes the sample index. Λ is then varied until the minimum PRESS value is obtained. The resulting value of Λ is then used with the full calibration data set, $[\mathbf{S},\mathbf{c}]$, to calculate $\mathbf{b}$. This regression vector can then be used to predict the concentrations of prospective samples with SEP values calculated by the following formula:

$$\mathrm{SEP} = \sqrt{\frac{\sum_{i=1}^{n}||c_{\mathrm{i}} - \hat{c}_{\mathrm{i}}||^2}{n}} \qquad (7)$$

with n the number of samples in the prospective data set. Because we compare several methods in this paper, it is convenient to denote the $\mathbf{b}$ vector obtained from a particular method as $\mathbf{b}_{\mathrm{method}}$.

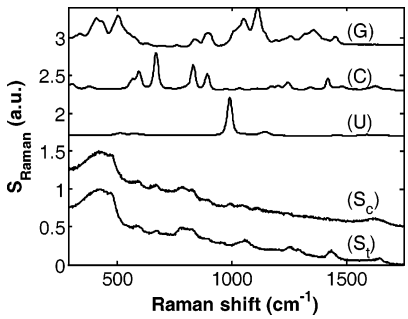

Figure 1. Measured Raman spectra of pure analytes dissolved in water and typical experimental mixture spectra in clear and turbid samples: (G) glucose, (C) creatinine, and (U) urea, (S_C) representative clear sample spectrum, and (S_T) representative turbid sample spectrum. For the turbid samples, the only clearly identifiable analyte peak is of creatinine at ~680 cm^{-1}. Traces are normalized and offset for clarity.

METHODS

In all studies, glucose and creatinine are chosen as the analytes of interest, while urea is present as an additional active Raman spectral interferent.

Numerical Simulations. Numerical spectra were generated by forming linear combinations of constituent analyte spectra of glucose (G), creatinine (C), and urea (U) as measured in our Raman instrument[18] (Figure 1). Spectra from 280 to 1750 cm^{-1} occupying 1051 CCD pixels were binned every 2 adjacent pixels to produce Raman spectra of 525 data points each, reducing the size of the data set for more rapid computation. Random concentrations uniformly distributed between 0 and 10 were used to generate 60 mixture sample spectra, with zero-mean Gaussian white noise generated by MATLAB superimposed on the spectra. The signal-to-noise ratio (SNR), defined here as the ratio of the major Raman peak magnitude to the mean noise magnitude, was ~9. The uniform noise across the spectra and the SNR are consistent with typical Raman spectra used for these types of analytical measurements. Half of the noise-added spectra formed the calibration set and the other half the prospective set. Different calibration methods were applied to the calibration set to generate the $\mathbf{b}$ vectors by minimizing the respective PRESS through leave-one-out cross-validation. The $\mathbf{b}$ vectors were then used to calculate the SEP among the prospective set. Repeating this entire procedure, we obtained average SEP values and $\mathbf{b}$ vectors for different methods. In all calibrations, three factors were needed to obtain

(17) Lorber, A. *Anal. Chem.* **1986**, *58*, 1167−1172.

(18) Enejder, A. M. K.; Scecina, T. G.; Oh, J.; Hunter, M.; Shih, W. C.; Sasic, S.; Horowitz, G. L.; Feld, M. S. *J. Biomed. Opt.* **2005**, *10*, 031114.

optimal prediction in PLS and HLA. The respective pure analyte spectrum was used as the spectral constraint for CR and HLA. Additionally, because all sample-generating constituent analytes were known, OLS was used to establish the best achievable prediction.

Two numerical simulations have been performed to evaluate the different methods under uncorrelated and correlated conditions. In the first simulation, all analyte concentrations varied randomly. In the second simulation, the glucose concentrations correlated to creatinine concentrations with $R^2 \sim 0.5$. Other implementation details are provided in the Results and Discussion section.

Experimental Mixture Spectra. *Clear Samples—Uncorrelated.* In the first experiment, Raman spectra were acquired from 84 water-dissolved mixture samples composed of glucose, creatinine, and urea, each with randomized concentration profiles from 0 to 50 mM, with respective mean ~25 mM. Half of the samples were acquired on day 1 and the rest on day 2 to allow instrumental drifts to be incorporated into the model. All samples were measured in a 1-cm-path length quartz cuvette using a Raman instrument described previously.[18] Each spectrum was acquired in 2 s with laser power equivalent to ~12 mW/mm^2 and a 1-mm^2 spot size at the sample. A total of 90 spectra of each water-dissolved analyte and of water were acquired and averaged for better SNR. Pure analyte spectra were obtained by subtracting the water plus quartz spectrum from the water-dissolved analyte spectra. A representative sample spectrum (S_C) is displayed in Figure 1. For data analysis, 21 samples randomly chosen from each day formed the calibration set, and the other 42 samples formed the prospective set. **b** vectors obtained using different calibration methods were applied to the prospective set to calculate SEP, and the randomized calibration−prediction procedure was repeated 400 times for each method. In all calibrations with leave-one-out cross-validation, five factors were needed to obtain optimal predictions in both PLS and HLA. The pure analyte spectra were used as the spectral constraints for both CR and HLA. Because of measurement errors in the pure analyte concentrations (estimated <1%), as well as to fully exploit HLA, we allowed the amplitude of the pure analyte spectra to vary within 1%.

Clear Samples—Correlated. In the second experiment, Raman spectra were acquired from 84 water-dissolved mixture samples composed of glucose, creatinine, and urea. Analyte concentrations were varied between 0 and 50 mM with mean ~25 mM. In 42 samples, the glucose concentrations correlated to creatinine concentrations with $R^2 \sim 0.5$, and in the other 42, they varied randomly. The urea concentration was random in all 84 samples. Half of the correlated samples (21) and the random samples (21) were acquired on day 1 and the rest on day 2 to allow instrumental drifts to be incorporated into the model. For data analysis, the 42 samples with the design correlation formed the calibration set and the 42 random samples formed the prediction set. Owing to the limited number of correlated samples, no randomized calibration−prediction sets were attempted. Other details are similar to the first experiment.

Turbid Samples. In the third experiment, the same protocol as in the first experiment was followed, but with the addition of intralipid and India ink to increase turbidity. The analyte concentrations were uncorrelated. Raman spectra were acquired from 84 water-dissolved mixture samples composed of glucose, creatinine, urea, India ink, and intralipid with randomized concentration profiles. Analyte concentrations were varied between 0 and 50 mM with mean ~25 mM. The concentration of India ink was varied such that the sample absorption coefficients ranged from 0.1 to 0.2 cm^{-1} with mean ~0.15 cm^{-1}. The concentration of intralipid was varied such that the sample scattering coefficients ranged from 35 to 75 cm^{-1} with mean ~55 cm^{-1}. The range of optical property changes agree well with reported values measured from human skin.[19] A representative sample spectrum (S_T) is displayed in Figure 1. In all calibrations with leave-one-out cross-validation, no more than six factors were needed to obtain optimal prediction in both PLS and HLA.

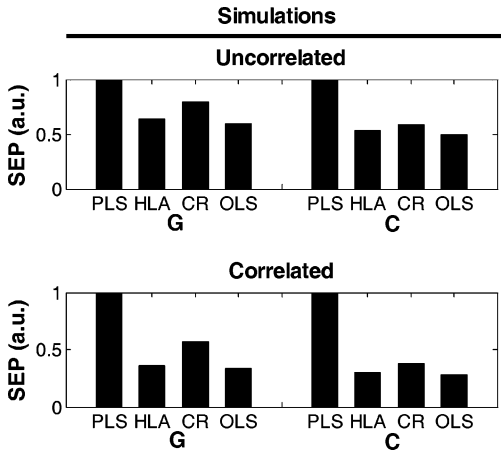

Figure 2. SEP values normalized to PLS results for glucose (G) and creatinine (C) obtained from various methods in the first (uncorrelated) and second (correlated) numerical simulations. See text for details.

It should be mentioned that using prediction error (SEP) to compare results from different methods rather than cross-validated error can effectively avoid false interpretation based on chance correlations and overfitting.

RESULTS AND DISCUSSION

All reported SEP values are normalized to the PLS SEP value for better comparison among methods.

Numerical Simulations. As mentioned in the Methods section, two numerical simulations were performed on spectra generated from measured constituent analyte spectra. The first simulation, in which analyte concentrations were uncorrelated, demonstrates that CR significantly outperforms PLS when all analyte concentrations vary in a random fashion. The results, summarized in Figure 2 (uncorrelated), show that with the aid of prior information, CR generates lower SEP values than PLS. The reason for this is that $\mathbf{b}_{CR}$ better converges to $\mathbf{b}_{OLS}$, therefore improving prediction over PLS. It is expected that HLA is only slightly inferior to OLS because the constraints are absolutely correct in simulations.

The second simulation, in which correlations between analytes were introduced, demonstrates that CR is less susceptible than PLS to spurious correlations among covarying analytes. We modified the calibration data set such that the concentration of

(19) Doornbos, R. M. P.; Lang, R.; Aalders, M. C.; Cross, F. W.; Sterenborg, H. J. C. M. *Phys. Med. Biol.* **1999**, *44*, 967−981.

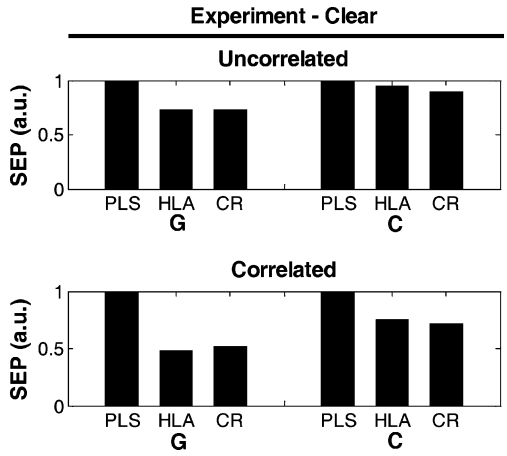

Figure 3. SEP values normalized to PLS results for glucose (G) and creatinine (C) obtained from various methods for clear sample experiments without (uncorrelated) and with (correlated) analyte correlations. See text for details.

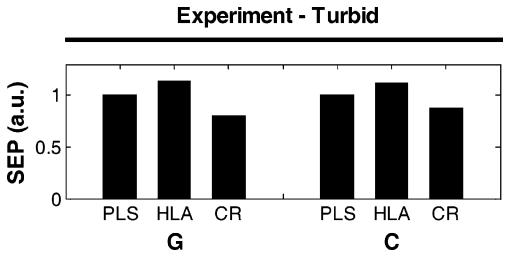

Figure 4. SEP values normalized to PLS results for glucose (G) and creatinine (C) obtained from various methods for the turbid sample experiment. See text for details.

glucose correlated to creatinine with $R^2 \sim 0.5$. The prospective set remained uncorrelated. The results are displayed in Figure 2 (correlated), in which CR possesses a much lower SEP value relative to PLS. Again, it is expected that HLA is little affected by analyte correlations because the constraints are absolutely correct in simulations, and therefore, any correlations are broken after removing the pure analyte contributions (discussed below).

It should be mentioned that the simulations allow us to study and compare various methods under perfect control; however, quantitative comparison with our experimental study is not intended.

Experimental Mixture Spectra. *Clear Samples—Uncorrelated.* Mean SEP values for glucose and creatinine obtained from PLS, HLA, and CR in the first experiment are summarized in Figure 3 (uncorrelated). OLS results are not listed because the three-constituent model does not account for all experimental variations, e.g., low amounts of fluorescence from the quartz cuvette; therefore, OLS no longer provides the best achievable performance. Among the implicit calibration techniques, substantial improvement over PLS is observed using the hybrid methods. CR and HLA generate similar SEP values, suggesting that these two methods have comparable performance under highly controlled experimental conditions with clear samples and without analyte correlations. The calculated 99% confidence intervals for the differences in means are SEP_{PLS-CR} (0.28, 0.33) and SEP_{HLA-CR} (−0.02, 0.02) for glucose, and SEP_{PLS-CR} (0.06, 0.13) and SEP_{HLA-CR} (0.02, 0.09) for creatinine, indicating that the results in comparison to PLS are statistically significant.

Clear Samples—Correlated. Mean SEP values for glucose and creatinine obtained from PLS, HLA, and CR in the second experiment are summarized in Figure 3 (correlated). Among the implicit calibration techniques, substantial improvement over PLS is observed using the hybrid methods. CR and HLA generate similar SEP values, suggesting that these two methods have comparable performance under highly controlled experimental conditions with clear samples and with analyte correlations. In principle, HLA should be less affected by analyte correlations than CR; however, this is not observed in this experiment. Possible explanations include imperfect experimental conditions and the higher sensitivity of HLA to inaccurate constraints (discussed below).

Turbid Samples. Mean SEP values for glucose and creatinine obtained from PLS, HLA, and CR in the third experiment with turbid samples are summarized in Figure 4. Substantial improvement over both PLS and HLA is observed using CR. The performance of HLA is significantly impaired as a result of the turbidity-induced sampling volume variations of the analyte of interest. In HLA, the analyte of interest is assumed to be present in the data according to the reference concentrations. This assumption leads to the first and most important step: the removal of the spectral contribution of the analyte of interest from the data by subtracting the known spectrum of the analyte according to its concentration in each sample. As a result, the performance critically depends on the "accuracy" of the constraint, as well as the legitimacy of the assumption. In CR, however, the constraint only guides the inversion, allowing the minimization algorithm to arrive at the optimal solution, thereby reducing its dependency on the accuracy of the constraint. Further, unlike HLA, which models the residual data after removing the analyte contribution, CR retains data fidelity and is unlikely to produce false built-in analyte spectral features in the **b** vector. The calculated 99% confidence intervals for the differences in means are SEP_{PLS-CR} (0.18, 0.23) and SEP_{HLA-CR} (0.31, 0.37) for glucose and SEP_{PLS-CR} (0.09, 0.15) and SEP_{HLA-CR} (0.32, 0.38) for creatinine, indicating that the results are statistically significant.

The results presented here demonstrate that there is a tradeoff between maximizing prior information utilization and robustness concerning the accuracy of such information. Multivariate calibration methods range from explicit methods with maximum use of prior information (e.g., OLS, least robust when accurate model is not obtainable), hybrid methods with a hard constraint (e.g., HLA), hybrid methods with a soft constraint (e.g., CR), and implicit methods with no prior information (e.g., PLS, most robust, but is prone to be misled by spurious correlations). We believe CR achieves the optimal balance between these ideals in practical situations.

CONCLUSION

Constrained regularization is a new hybrid method for multivariate calibration. Strictly speaking, it should be categorized as an implicit calibration method with one additional piece of information, the spectrum of the analyte of interest. In the broader context, regularization methods may perform somewhat better

than either PLS or PCR[20] for certain data structures. A heuristic explanation is that regularization provides a continuous "knob" and, therefore, can be used to find a better balance between model complexity and noise rejection. Our results show that, in addition to this plausible intrinsic advantage, solid improvement can be obtained by incorporating a meaningful solution constraint.

CR significantly outperforms methods without prior information such as PLS and is less susceptible to spurious correlations with covarying analytes. Compared to HLA, CR has superior robustness with inaccurate spectral constraints. This robustness is crucial for hybrid methods because it is difficult, if not impossible, to quantify high-fidelity pure analyte spectra in complex systems such as biological tissues. Further, CR naturally extends to situations in which pure spectra of more than one constituent are also known. In that case, a better choice of constraint ($\mathbf{b}_0$) might be the net analyte signal calculated from all the known pure spectra. CR is thus able to include more prior information without sacrificing the principal advantage of implicit calibration: that only the reference concentrations of the analyte of interest are required in addition to the calibration spectra.

(20) Frank, I. E.; Friedman, J. H. *Technometrics* **1993**, *35*, 109−135.

ACKNOWLEDGMENT

This work was performed at the MIT Laser Biomedical Research Center and supported by the NIH National Center for Research Resources, Grant P41-RR02594, and a grant from Bayer Health Care, LLC. We thank Thomas Scecina for helpful discussions. We thank professor George Barbastathis for valuable discussions and W.-C.S. acknowledges a fellowship from the Martin Family Society of Fellows for Sustainability at the Massachusetts Institute of Technology.

Received for review April 18, 2006. Accepted October 17, 2006.

AC060732V