Prediction range estimation from noisy Raman spectra with robust optimization

Olga Lyandres,[a] Richard P. Van Duyne,[b] Joseph T. Walsh,[a] Matthew R. Glucksberg[a] and Sanjay Mehrotra*[c]

Received 10th March 2010, Accepted 16th May 2010
DOI: 10.1039/c0an00134a

Inferences need to be drawn in biological systems using experimental multivariate data. The number of samples collected in many such experiments is small, and the data are noisy. We present and study the performance of a robust optimization (RO) model for such situations. We adapt this model to generate a minimum and a maximum estimation of analyte concentration for a given sample, producing a prediction range. The calibration model was applied to sets of Raman spectra. In particular we used normal Raman measurements of pyridine/deuterated pyridine mixtures and spectra from a more complex glucose detection system based on surface-enhanced Raman spectroscopy. The results from the RO model were compared with prediction intervals estimated from partial least squares (PLS) method. We find that the RO prediction ranges included the actual concentration value of the sample more consistently than the 99% prediction intervals built with PLS methods.

Introduction

The clinical success of biological sensors depends on the accuracy of their measurements. Despite many years of innovation, calibration remains a challenging aspect of sensor design due to the noisy data arising from the complexity of the biological environment, variations in experimental conditions, and instrument stability. While sensors and calibration methods are initially characterized in laboratory settings where parameters are carefully controlled and close-to-ideal conditions can be achieved, in complex biological environments, it is rarely the case. Providing a more realistic, even worst-case prediction for a patient might have a significant impact on treatment decisions. An example of this situation includes potential hypoglycemia among insulin treated diabetics, which could be fatal if goes unchecked. In this case, a good estimate of the lower threshold of predicted glucose concentration is more important than a mean estimate. Furthermore, because of physical limitations, samples are available for estimation in only a few such situations.

Multivariate calibration techniques have been widely utilized in analytical chemistry in an attempt to model the relationship between a variable and the effect of this variable on the corresponding measurements. Among the methods commonly used is partial least squares (PLS), which has been utilized in proteomics, in a variety of spectroscopic techniques, as well as in analyzing process data.[1–9] In the case of spectroscopic sensors, a regression vector is computed that relates the spectral measurements to the concentration of the analyte in the sample and can be used to predict concentration of analyte in unknown samples. These methods do not require individual spectra of each system variable to be known, and thus are referred to as implicit calibration methods.

The computation of PLS regression vector involves non-linear systems, therefore computing prediction intervals is challenging. Different methods such as bootstrapping, cross-validation, and local linearization methods have been proposed.[10] There is also an ASTM standard E1655, based on ordinary least squares estimation of regression coefficients, that gives a formula for computing the variance of prediction variables.[11] Faber and Kowalski proposed an approximation method that is very similar to the ASTM standard.[12] Aside from a term that arises due to mean-centering of data, the methods compute the same estimates. Several methods based on local linearization have also been studied.[10,13,14] These require the computation of a Jacobian matrix of the regression vector. While all of them are computationally expensive, based on empirical evidence, Denham recommends computing prediction intervals for PLS using a local linearization method over boot-strapping and other approaches.[10] As an example of computational costs, Denham's algorithm computes $p^2 \times p$ and $p^2 \times n$ matrices, where p is the number of pixels in the Raman spectrum, and n is the number of concentrations for which calibration spectra are collected. An improved algorithm requiring calculation of a $p \times p$ matrix to compute this Jacobian matrix is given by Serneels *et al.*[13] Phatak *et al.*[14] propose an alternative algorithm to compute the Jacobian matrix of the regression vector. Furthermore, estimation of the degrees of freedom is needed for calculation of prediction intervals using t-statistics. The conclusion of a paper published in 2009 states "Results from this study seem to suggest that none of these algorithms generate accurate coverage rates in all cases. Note that there are large uncertainties in the coverage probability due to the use of a finite set of test data. Particularly there is a need in developing an uncertainty estimation method consistent with the reduced rank nature of the PLS method, using parameters which are part of the PLS model."[15]

We propose a robust optimization (RO) approach for calibration of spectral data as a complement to the PLS based calibration. The concept of robust modeling has received greater

[a]*Department of Biomedical Engineering, Northwestern University, Evanston, Illinois, 60208, USA*
[b]*Chemistry Department, Northwestern University, Evanston, Illinois, 60208, USA*
[c]*Department of Industrial Engineering and Management Sciences, Northwestern University, Evanston, Illinois, 60208, USA. E-mail: mehrotra@iems.northwestern.edu*

attention recently in optimization.[16–18] In robust modeling an uncertainty set is used to define data noise, and a worst-case optimization is performed over this uncertainty set. Problems of varying computational difficulty result depending on the imposed model of uncertainty.[18] Following this modeling paradigm, the notion of robust calibration proposed here signifies a worst-case scenario prediction based on an uncertainty model imposed from the available calibration data. Furthermore, we adapt the model to produce a prediction range, rather than a single value prediction for a measurement. We compare RO determined prediction ranges with 99% prediction intervals computed from PLS. We find that RO generated prediction range more consistently includes the true value when compared with the PLS generated prediction interval. Since the 95% prediction intervals for PLS are narrower, the same conclusions also apply to these prediction intervals.

Theory

Robust optimization prediction model

Let y_j, $j = 1,\ldots, n$, be the concentrations for which we have collected spectra. The i^{th} ($i = 1,\ldots, m$) calibration spectra for the j^{th} concentration is represented by vectors $\boldsymbol{x}_{ij} \in R^p$, where p is the number of elements in $\boldsymbol{x}_{ij}$. For the concentration y_j we build the spectra uncertainty set

$$U_j = \left\{ \boldsymbol{x}_j \middle| \boldsymbol{x}_j = \sum_{i=1}^{m} \lambda_{ij}\boldsymbol{x}_{ij}, \sum_{i=1}^{m} \lambda_{ij} = 1,\ \lambda_{ij} \geq 0 \right\}, \quad j = 1, \ldots, \ n \tag{1}$$

The vector $\boldsymbol{x}_j$ defined in eqn (1) is generated by taking a non-negative weighted sum of the observed spectra $\boldsymbol{x}_{ij}$, where the weights λ_{ij} sum up to one. This is called taking a convex combination of ($\boldsymbol{x}_{1j},\ldots, \boldsymbol{x}_{mj}$). Since any non-negative weights summing up to one are allowed, we obtain a set U_j. Such a set U_j is called the convex hull of ($\boldsymbol{x}_{1j},\ldots, \boldsymbol{x}_{mj}$). Each concentration has its associated uncertainty set. The underlying assumption here is that we may observe any spectra in the set U_j when collecting data for y_j. Alternative model may use a 'pixel-wise' uncertainty set, or quadratic uncertainty set, such as the one proposed by Ben-Tal and Nemirovski for each concentration.[16] However, these models give a larger uncertainty set and add to the complexity of the model.† Note that the mean spectrum $\bar{\boldsymbol{x}}_j \equiv \frac{1}{m}\sum_{i=1}^{m} \boldsymbol{x}_{ij}$ is in the set U_j.

For a spectrum $\boldsymbol{x}^{\text{u}}$ of unknown concentration y^{u} we associate a linear predictive model

$$y^{\text{u}} := \sum_{j=1}^{n} \beta_j y_j \tag{2}$$

where the prediction coefficients β_j, $j = 1,\ldots, n$ are to be determined. The standard least-squares model with no uncertainty in the calibration spectra (*i.e.*, $\boldsymbol{x}_j^0 := \boldsymbol{x}_{1j} = \ldots = \boldsymbol{x}_{mj}$, for any j) determines β_j by considering the minimization problem

$$\min_{\beta_1,\ldots,\beta_n} \left\| \boldsymbol{x}^{\text{u}} - \sum_{j=1}^{n} \beta_j \boldsymbol{x}_j^0 \right\|^2 \tag{3}$$

and uses the solution of eqn (3) in eqn (2) for concentration estimation. In the presence of data uncertainty we build the optimization model for the lower limit estimator y_{lo}* as:

$$y_{\text{lo}}^* = \min \sum_{j=1}^{n} \beta_j y_j \tag{4}$$

$$\text{subject to } \left\| \boldsymbol{x}^{\text{u}} - \sum_{j=1}^{n} \beta_j \boldsymbol{x}_j \right\|^2 \leq z \tag{5}$$

$$\boldsymbol{x}_j \in U_j, j = 1,\ldots, n \tag{6}$$

$$\sum_{j=1}^{n} \beta_j = 1, \ \beta_j \geq 0 \tag{7}$$

This model assumes that the unknown concentration y^{u} is a convex combination of the known concentrations y_j, *i.e.*, $y^{\text{u}} \in [y_1, y_n]$ where without loss of generality $y_1 \leq y_2 \leq \ldots \leq y_n$. The parameter z controls the feasible set for β. We will discuss its choice below. Similar to the model for the lower limit prediction given in eqn (4)–(7) we have a model for the maximum concentration estimator as

$$y_{\text{up}}^* = \max \sum_{j=1}^{n} \beta_j y_j \tag{8}$$

$$\text{subject to } \left\| \boldsymbol{x}^{\text{u}} - \sum_{j=1}^{n} \beta_j \boldsymbol{x}_j \right\|^2 \leq z \tag{9}$$

$$\boldsymbol{x}_j = \sum_{i=1}^{m} \lambda_{ij}\boldsymbol{x}_{ij}, \ \sum_{i=1}^{m} \lambda_{ij} = 1, \ \lambda_{ij} \geq 0, \ \ j = 1, \ldots, \ n \tag{10}$$

$$\sum_{j=1}^{n} \beta_j = 1, \ \beta_j \geq 0$$

where we have written the uncertainty requirement, given previously in eqn (6), explicitly in eqn (10). The range [y_{lo}*,y_{up}*] is called the *prediction range*.

Proposition 1 gives an equivalent convex optimization formulation of upper limit prediction model eqn (8)–(10). The lower limit prediction model given by eqn (4)–(7) can be formulated in a similar manner. This model, which has significantly fewer constraints, remodels non-linear constraint specified by eqn (9) as a convex quadratic constraint specified below in eqn (12), making it more tractable. Algorithms for solving

† The convex hull of a set of points S is the smallest convex set containing S. In this sense any alternative model for building uncertainty set that maintains convexity and includes the observed spectra necessarily gives larger uncertainty sets, hence providing more conservative estimates than the ones obtained using U_j. An alternative way to build larger (or smaller) uncertainty sets than U_j in the current framework is as follows. Let $\boldsymbol{d}_{ij} = \boldsymbol{x}_{ij} - \bar{\boldsymbol{x}}_j$ and $\hat{\boldsymbol{x}}_{ij} = \bar{\boldsymbol{x}}_j + \delta_{ij}\boldsymbol{d}_{ij}$, $\delta_{ij} \geq 0$. Then a larger set $\hat{U}_j$ can be defined by taking $\delta_{ij} > 1$ and letting $\hat{U}_j := \text{conv}(\hat{\boldsymbol{x}}_{1j},\ldots, \hat{\boldsymbol{x}}_{mj})$. A smaller uncertainty set is defined if $0 \leq \delta_{ij} < 1$ is used in defining $\hat{U}_j$.

quadratically constrained convex quadratic programs in polynomial time are known.[19–21]

Proposition 1: if a solution (β_j*, λ_j*, $j = 1,\ldots, n$, $i = 1,\ldots, m$) is optimum for upper limit prediction model eqn (8)–(10) then $\gamma_{ij}^* = \lambda_{ij}^*/\beta_j^*$, $j = 1,\ldots, n$, $i = 1,\ldots, m$ is optimum for

$$\max \sum_{j=1}^{n} \sum_{i=1}^{m} \gamma_{ij} y_{ij} \tag{11}$$

$$\text{subject to } \left\| \boldsymbol{x}^{\mathrm{u}} - \sum_{j=1}^{n} \sum_{i=1}^{m} \gamma_{ij} \boldsymbol{x}_j \right\|^2 \leq z \tag{12}$$

$$\sum_{j=1}^{n} \sum_{i=1}^{m} \gamma_{ij} = 1, \quad \gamma_{ij} \geq 0, \tag{13}$$

where $y_{ij} = y_j$, $i = 1,\ldots, m$, $j = 1,\ldots, n$. Furthermore, if γ_{ij}^*, $j = 1,\ldots, n$, $i = 1,\ldots, m$ is optimum for eqn (11)–(13), then $\beta_j^* = \sum_{i=1}^{m} \gamma_{ij}^*$, $\lambda_{ij}^* = \gamma_{ij}^*/\beta_j^*$, is optimum for the upper limit prediction model eqn (8)–(10). Finally, the optimum objective value of eqn (11)–(13) is y_{up}*.

Proof follows from making a direct substitution of constructed solution and comparing the corresponding objective values.

We now specify the value of z used in our computations. We consider the 'average' model

$$z^* \equiv \min_{\beta_1,\ldots,\beta_n} \left\| \boldsymbol{x}^{\mathrm{u}} - \sum_{j=1}^{n} \overline{\beta}_j \overline{x}_j \right\|^2, \quad \sum_{j=1}^{n} \overline{\beta}_j = 1, \quad \overline{\beta}_j \geq 0 \tag{14}$$

that one may use for a point estimate for analyte concentration in the absence of a robust model. Let z* be the optimum objective value in eqn (14). We take $z = z$*. This choice of z ensures that eqn (11)–(13) is feasible, since $\gamma_{1j} = \ldots = \gamma_{mj} = \dfrac{\overline{\beta}_j}{m}$ satisfies constraints (12) and (13) whenever $\bar{\alpha}_j$ is feasible for eqn (14). For this choice of z the upper limit prediction y_{up}* gives the largest analyte concentration that would be possible under model eqn (2) if the prediction spectra $\boldsymbol{x}^{\mathrm{u}}$ can be obtained by taking a convex combination of known spectra $\boldsymbol{x}_{ij}$ in a way not to exceed error z*. The corresponding minimization problem has an analogous interpretation.‡

Partial least squares

Since there is no clear choice of a method for a comparison, we used all of them. Prediction intervals reported below were determined by four different methods: boot-strapping, Faber's, Serneels', and Phatak's algorithms.

The performance of PLS models further depends on the number of latent variables (LVs) used. The appropriate number is determined by root-mean-squared error of cross-validation (RMSECV). Typically, the number of LVs corresponding to minimum RMSECV is used. It should be noted that the determination of the optimal number of LVs is empirical and depends on the system in question. The number of LVs used in the model also affects the calculation of prediction intervals and should, therefore, be considered. Computations were performed for LVs = 1,…,10 to evaluate how the prediction interval changes as a function of LVs used. The data were mean-centered and the optimal number the LVs was determined by the software package used for all PLS calculation with leave-one-out cross-validation. The software recommendations were consistent for both data sets and further verified with Venetian blinds and random cross-validation methods. We present the results for 99% prediction intervals based on a 7 LVs model. The use of 7 LVs is in agreement with previous PLS calibration models constructed for the SERS glucose sensor where 6–8 LVs were utilized.[22,23] Furthermore, with a 7 LVs model, the resulting prediction intervals are comparable to average prediction ranges computed by RO for both the glucose and pyridine systems. Use of a 99% prediction interval is consistent with a 100% inclusion that RO strives for. Furthermore, the method to estimate the degrees of freedom greatly influences the performance of the algorithm.[15] We utilized generalized degrees of freedom in the prediction interval estimation as recommended by Zhang *et al.*[15]

Results and discussion

Tables 1 and 2 (Column 2) give the RO prediction ranges and point predictions for a set of pyridine data and SERS glucose sensor data, respectively. Point estimates and corresponding prediction intervals generated from PLS for pyridine and glucose are given in Tables 1 and 2 (Column 3–6). The last three rows in Tables 1 and 2 give the mean prediction range, relative error (calculated when the actual value is not included in the range, as the difference between the value and the closest boundary of the range, averaged for all concentration values that were not included in the prediction interval), and the root mean squared error of prediction (RMSEP) for pyridine and glucose, respectively. Fig. 1 shows the RO prediction ranges and the PLS prediction intervals for pyridine. Fig. 2 gives the same information for glucose.

Comparison of calibration models—pyridine system

The choice of pyridine and pyridine-D_5 mixtures is justified by the fact that the two components are independent and have very distinct narrow bands. Since pyridine exhibits strong Raman scattering, the signal-to-noise of the spectra in this system is much higher compared to the glucose system. As a result, we expect to have prediction ranges from RO and the PLS prediction intervals to be quite narrow for this system. The actual concentration is included in the range for 11 out of 15 samples for RO method as shown in Fig. 1a. Prediction intervals determined from the PLS models vary in their accuracy depending on the method used. When 7 LVs are used, only 3 out of 15 prediction intervals include the true concentration of the sample for bootstrap, Faber96, and Phatak methods and 7 out of 15 for the Serneels method even though the prediction intervals are comparable to or slightly larger in size than the RO mean prediction range (Fig. 1b). The other

‡ We point out that the objective value of upper limit prediction in eqn (8)–(10) is a non-decreasing function of z, hence a larger (respectively, smaller for minimization problem) upper limit value will result when the value of z is increased until constraint eqn (12) becomes redundant.

Table 1 Summary of the RO prediction ranges and PLS prediction intervals, for pyridine

Actual concentration	Robust optimization (RO) prediction range	Partial least squares (PLS) prediction − 99% prediction interval			
		Bootstrap method	Faber96 method	Serneels method	Phatak method
y_{meas}	$y_{lo} - y_{up}/y_{lsq}$	$y_{lo} - y_{up}/y_{PLS}$	$y_{lo} - y_{up}/y_{PLS}$	$y_{lo} - y_{up}/y_{PLS}$	$y_{lo} - y_{up}/y_{PLS}$
Pyridine concentrations (% v/v)					
5	3.14 − 8.66/3.97	−11.81 − −6.03/−8.92[a]	−11.63 − −6.21/−8.92[a]	−13.38 − −4.46/−8.92[a]	−11.63 − −6.21/−8.92[a]
10	8.6 − 17.37/8.87	−0.52 − 4.48/1.98[a]	−0.49 − 4.45/1.98[a]	−0.62 − 4.58/1.98[a]	−0.49 − 4.45/1.98[a]
15	14.3 − 19.85/17.66[b]	10.41 − 15.32/12.87	10.42 − 15.31/12.87	10.41 − 15.32/12.87	10.42 − 15.31/12.87
25	24.96 − 26.7/25.87	25.08 − 29.90/27.49[a]	25.09 − 29.90/27.49[a]	24.97 − 30.01/27.49	25.09 − 29.90/27.49[a]
30	30.19 − 32.57/31.47[a]	33.18 − 38.08/35.63[a]	33.21 − 38.05/35.63[a]	32.89 − 38.36/35.63[a]	33.21 − 38.05/35.63[a]
35	35.75 − 36.88/36.16[a]	35.27 − 40.09/37.68[a]	35.27 − 40.08/37.68[a]	35.24 − 40.11/37.68[a]	35.27 − 40.08/37.68[a]
45	44.32 − 48.64/44.98	40.15 − 45.30/42.72	40.27 − 45.18/42.72	39.55 − 45.90/42.72	40.27 − 45.18/42.72
50	47.47 − 54.64/50.24	44.41 − 49.60/47.01[a]	44.41 − 49.60/47.01[a]	43.07 − 50.94/47.01	44.41 − 49.60/47.01[a]
55	52.62 − 59.35/54.85	49.91 − 54.97/52.44[a]	50.01 − 54.87/52.44[a]	49.60 − 55.28/52.44	50.01 − 54.87/52.44[a]
65	61.65 − 67.98/65.47	59.48 − 64.55/62.02[a]	59.58 − 64.45/62.02[a]	59.07 − 64.96/62.02[a]	59.58 − 64.45/62.02[a]
70	68.08 − 75.62/71.93	65.43 − 70.80/68.11	65.62 − 70.60/68.11	65.40 − 70.83/68.11	65.62 − 70.60/68.11
75	71.96 − 77.61/75.43	69.64 − 74.66/72.15[a]	69.75 − 74.54/72.15[a]	69.59 − 74.70/72.15[a]	69.75 − 74.54/72.15[a]
85	80.52 − 85.66/83.25	79.58 − 84.90/82.24[a]	79.82 − 84.66/82.24[a]	79.44 − 85.03/82.24	79.82 − 84.66/82.24[a]
90	86.50 − 89.31/87.2[a]	83.79 − 88.85/86.32[a]	83.89 − 88.74/86.32[a]	83.43 − 89.21/86.32[a]	83.89 − 88.74/86.32[a]
95	90.08 − 93.41/91.16[a]	88.08 − 93.06/90.57[a]	88.13 − 93.00/90.57[a]	87.82 − 93.31/90.57[a]	88.13 − 93.00/90.57[a]
Mean range	4.38	5.1	4.93	5.84	4.93
Relative error	0.81	2.04	2.12	2.95	2.12
RMSEP	1.7	5.1	5.1	5.1	5.1

[a] Indicates when actual value is not included in prediction range or interval. [b] Coefficients initialized to values determined by least squares solution, for all other samples coefficients were initialized to 0.1.

Table 2 Summary of RO prediction ranges and PLS prediction intervals for glucose

Actual concentration	Robust optimization (RO) prediction range	Partial least squares (PLS) prediction − 99% confidence interval			
		Bootstrap method	Faber96 method	Serneels method	Phatak method
y_{meas}	$y_{lo} - y_{up}/y_{lsq}$	$y_{lo} - y_{up}/y_{PLS}$	$y_{lo} - y_{up}/y_{PLS}$	$y_{lo} - y_{up}/y_{PLS}$	$y_{lo} - y_{up}/y_{PLS}$
Glucose concentrations/mg dL^{-1}					
15	11.94 − 120.18/14.22	−40.38 − 95.65/27.64	−38.32 − 93.60/27.64	−40.79 − 96.06/27.64	−38.53 − 93.80/27.64
50	31.42 − 108.96/39.04	−10.95 − 130.86/59.96	−7.30 − 127.21/59.96	−11.66 − 131.58/59.96	−7.55 − 127.46/59.96
80	73.41 − 245.65/136.47	48.88 − 180.07/114.47	49.50 − 179.45/114.47	48.78 − 180.17/114.47	49.47 − 179.48/114.47
120	65.69 − 196.83/100.52	71.99 − 205.43/138.71	73.16 − 204.26/138.71	71.73 − 205.70/138.71	72.77 − 204.65/138.71
200	182.73 − 303.59/229.1	143.54 − 277.72/210.63	145.27 − 276.00/210.63	143.25 − 278.01/210.63	144.93 − 276.33/210.63
300	131.91 − 334.21/201.9	165.87 − 300.91/233.39	167.48 − 299.30/233.39[a]	165.56 − 301.22/233.39	167.20 − 299.57/233.39[a]
Mean range	135.39	135.28	131.67	135.98	132.17
Relative error	—	—	0.7%	—	0.4%
RMSEP	48.58	32.53	32.53	32.53	32.53

[a] Indicates when actual value is not included in the prediction range or confidence interval.

two metrics, relative error and RMSEP, are 0.8 and 1.7, respectively, in RO analysis. On the other hand, both relative error and RMSEP are higher in PLS, with values of 2.3 and 5.1, respectively. Moreover, when the concentration values are small, a negative concentration lower bound having no physical meaning is often given by PLS. The prediction interval computation (results not included here) using PLS was also performed for the pyridine data set by truncating it in the spectral sub-regions (from within the larger spectral range) with vibrational features of the molecules. No changes were seen in the performance of the PLS generated prediction intervals in including the true concentration.

It is apparent from pyridine data set that even a simple system can have experimental uncertainty, which will cause errors in quantitative measurements. Both the RO and PLS require assumptions (PLS prediction intervals are valid under normality assumption). With only a few calibration spectra per concentration, this is difficult to build or verify accurately. It is neither feasible nor practical to have to calibrate a sensor with a large number of measurements. While experimental setup can be modified to reduce the variation in the system, it is clear that a large degree of uncertainty will remain largely due to the nature of the sensor and the biological environment where it will be placed. The results indicate that the RO model is superior at incorporating data uncertainty by using a 'worst-case' model than the PLS derived prediction intervals. Thus, an alternative method such as RO for generating prediction intervals and providing information regarding the accuracy of the point prediction is a valuable addition to the toolbox of data analysis methods.

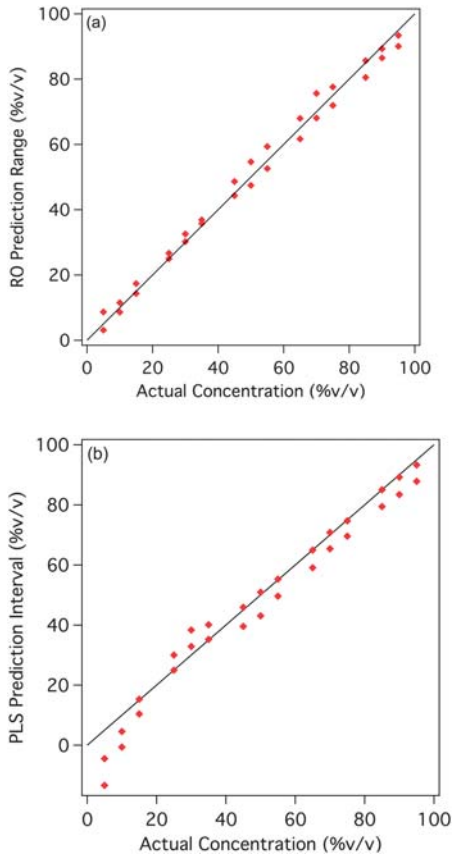

Fig. 1 Calibration of normal Raman pyridine measurements: (a) robust optimization (RO) prediction ranges and (b) PLS prediction intervals based on a model constructed with 7 latent variables.

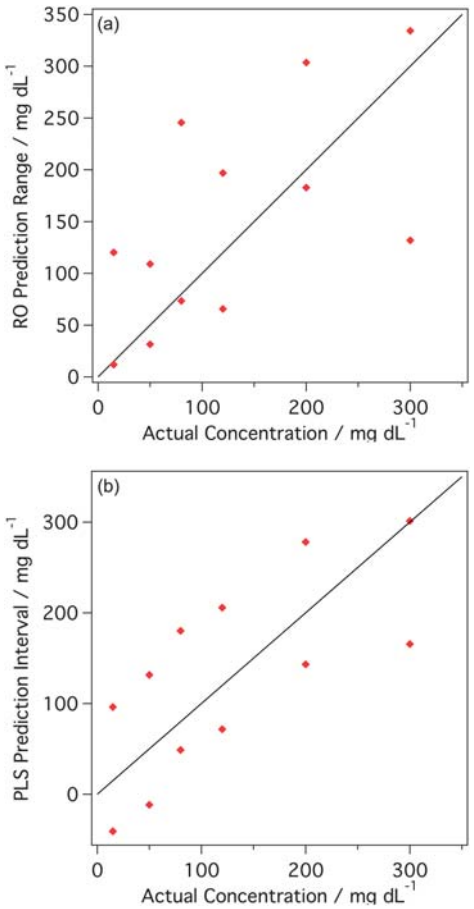

Fig. 2 Calibration of SERS glucose measurements: (a) robust optimization (RO) prediction ranges and (b) PLS prediction intervals based on a model constructed with 7 latent variables.

Comparison of calibration models—glucose system

Even though the glucose detection system has one explicit variable, it is significantly more complex. In addition to the analyte of interest, glucose, the sensor is composed of a surface-enhancing Raman substrate, which is functionalized by a self-assembled monolayer to facilitate glucose interaction with the surface. The detection scheme has been previously described in detail.[22] The sensor surface is a dynamic system that responds to changes in temperature, humidity, and incubation time, although the discussion of sensor stability is beyond the scope of this paper. The issues of data complexity and inability of PLS to incorporate prior information, *i.e.* spectrum of the analyte of interest, have been noted before by Feld and coworkers.[24,25] Hybrid linear analysis (HLA) and constrained regularization (CR) were explored in order to include prior information into a PLS model.[24,25] While prediction errors are diminished by applying CR, HLA does not actually perform better in systems with turbid samples, an attribute critical for biological sensing.[24] Other methods for substantiating PLS performance and reducing the risk of spurious correlations have been utilized by Arnold and coworkers. Spectral ranges optimal for calibration are carefully analyzed.[26–28] Furthermore, to assess the selectivity and sensitivity of the model, pure component selectivity analysis (PCSA)[29] and net analyte signal (NAS)[30] calculations are performed, respectively. First proposed by Lorber *et al.*, NAS is a generalized figure of merit for calibration models.[31] While all these methods improve the results of PLS calibration, none of them are 'truly' robust with respect to uncertainties in the underlying data. They attempt to incorporate *a priori* information or require additional spectra that may not be easily attainable or be of needed quality to produce meaningful results. The robust multivariate calibration and prediction problem are further complicated by the fact that very few spectra are available per concentration. Hence, the assumptions underlying PLS/PCSA analysis are not fully met or validated. Furthermore, these techniques focus only on a point prediction rather than estimation of a prediction interval.

RO analysis was applied in a similar way to a set of SERS glucose sensor data. The actual concentration value is included in the prediction range generated from the RO model for all samples as shown in Fig. 2a. Overall, the average prediction interval from PLS computed with 7 LVs over all samples is comparable to the prediction interval computed from RO (Fig. 2b). At lower concentrations, prediction intervals include negative values, which have no physical meaning. For the Faber96 and Phatak methods, the sample at 300 mg dL^{-1} is not included in the prediction interval. RMSEP for PLS predictions is 32.5 while RO predictions result in a RMSEP of 48.6. For a real system, with noisy data and limited calibration samples,

such as a SERS glucose sensor, these results demonstrate that RO prediction ranges may outperform PLS prediction intervals when information is limited. All the intervals/ranges are indeed large, however, this is a weakness of the sensor design rather than the model.

Experimental

Materials

All chemicals were reagent grade or better, and used as purchased. Silver pellets (99.99%) were purchased from Kurt J. Lesker Company (Clairton, PA). Titanium was obtained from McMaster-Carr (Chicago, IL) and cut into 18 mm diameter disks. To clean substrates, NH_4OH, H_2O_2, and CH_3CH_2OH were used from Fisher Scientific (Fairlawn, VA). Surfactant-free, white carboxyl-substituted latex polystyrene nanosphere suspensions (390 ± 19.5 nm diameter, 4% solid) were purchased from Duke Scientific Corporation (Palo Alto, CA). Ultrapure water (18.2 MΩ cm^{-1}) from a Millipore system (Marlborough, MA) was used for substrate and solution preparation. Glucose, pyridine, and deuterated pyridine (pyridine-D_5) were purchased from Sigma (St Louis, MO). Decanethiol ($CH_3(CH_2)_9SH$) and 6-mercapto-1-hexanol ($HS(CH_2)_6OH$) were purchased from Aldrich (Milwaukee, WI).

Surface fabrication procedure

The titanium substrates were cleaned by sonicating in 5 : 1 : 1 H_20/30% H_2O_2/NH_4OH. Approximately 10 μL of nanosphere solution were drop-coated onto a clean copper substrate and allowed to dry at room temperature. Then, 200 nm thick Ag films were deposited onto and through the nanosphere mask using the Kurt J. Lesker electron beam deposition system (Clairton, PA) to form Ag film over nanosphere (AgFON) substrates. The mass thickness and deposition rate (2 Å s^{-1}) of the Ag metal were measured by 6 MHz gold plated quartz-crystal microbalance purchased from Sigma Instruments (Fort Collins, CO). AgFON substrates were functionalized with a two component self-assembled monolayer (SAM). They were first incubated in 1 mM decanethiol in ethanol for 45 minutes and then transferred to 1 mM mercaptohexanol in ethanol for at least 12 hours. Then the SAM-functionalized surfaces were mounted into a small volume flow cell for spectra collection.

Instrumentation

A Spectra-Physics model Millennia Vs laser (λ_{ex} = 532 nm) and a Renishaw diode laser (λ_{ex} = 785 nm) were used in the experiments; the laser spot size on the sample was less than 0.5 mm in diameter. The measurement system includes an interference filter, an edge filter (Semrock, Rochester, NY), a model VM-505 single-grating monochromator with the entrance slit set at 100 μm (Acton Research Corp., Acton, MA), and a LN_2-cooled CCD detector (Roper Scientific, Trenton, NJ). A collection lens with magnification 5 was used to collect the scattered light.

System parameters

Two different systems were used to examine the performance of the model: a surface-enhanced Raman spectroscopy (SERS) based glucose sensor and normal Raman measurements of pyridine/pyridine-D_5 mixtures at varying concentrations. The SERS sensor and glucose detection mechanism were described in previous reports.[22] The SERS glucose sensor consists of a supporting metal substrate with a SERS-active nanostructured Ag film. The Ag film is functionalized with a self-assembled monolayer (SAM), which consists of two components to provide an appropriate hydrophilic/hydrophobic surface chemistry and facilitate interactions of the surface with the analyte. The pyridine system was used as proof of concept to confirm the results obtained with the SERS glucose sensor. Pyridine and pyridine-D_5 are two independent components that have narrow Raman bands at different frequencies. The spectral range of the data used in the analysis is 525–1800 cm^{-1} corresponding to 1340

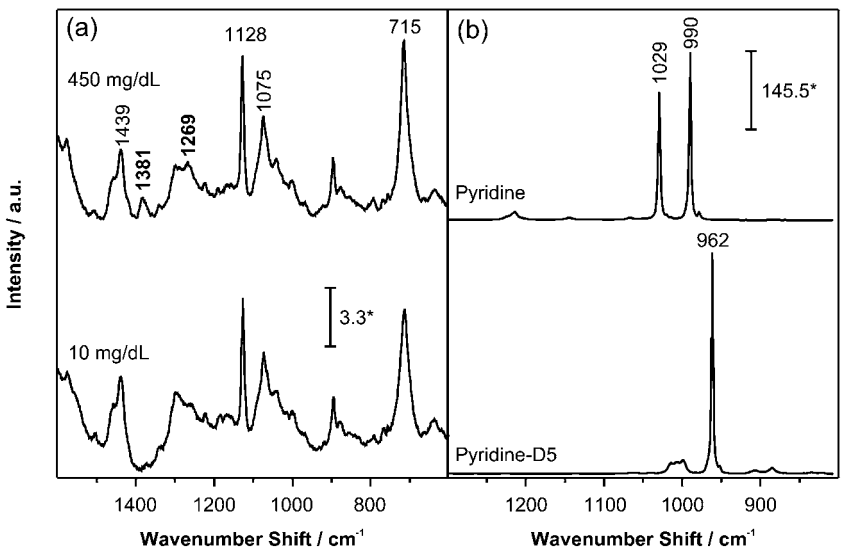

Fig. 3 (a) Representative surface-enhanced Raman glucose spectra used in data analysis. Lower concentration (10 mg dL^{-1}) and higher concentration (450 mg dL^{-1}) are shown. (b) Normal Raman spectra of pyridine and deuterated pyridine used in data analysis. * denotes adu mW^{-1} s^{-1}.

CCD pixels for each spectrum. This results in 1340 data points ($p = 1340$) for each spectrum used in the analysis. Glucose solutions were prepared in phosphate buffered saline (pH ~7.4) with concentrations 10, 20, 40, 60, 100, 150, 250, 350, and 450 mg dL^{-1} used for calibration (9 samples), and 15, 50, 80, 120, 200, and 300 mg dL^{-1} used for validation of the model (6 samples). A total of 10 spectra were acquired for each sample. Thus, $n = 9$ and $m = 10$ for the implementation of the robust prediction model. The PLS regression vector was computed based on 10 calibration spectra for each concentration. The point estimates and the corresponding prediction intervals were determined for each concentration value based on a mean spectrum (average of 10 spectra) at that concentration.

The acquisition time, t_{ac}, for each spectrum was 2 min, with $P = 10$ mW at the sample, and excitation wavelength, λ_{ex}, of 532 nm. Spectra were collected for each solution after it was injected into a small-volume flow cell that contained the sensor and allowed to incubate for 1 minute. The flow cell was cleared, but not rinsed before next solution was introduced into the flow cell. Representative spectra are shown in Fig. 3a.

For pyridine measurements, spectral range of the data used in the analysis is 800–1330 cm^{-1}. The spectral range changed due to the change in the excitation wavelength used in the experiments, however, the number of data points remained fixed since the same detection system was used for both. Equal volume mixtures of pyridine and pyridine-D_5 were made by changing the ratio of the two components at 5% intervals, from 0% to 100% (a total of 21 samples). Mixtures were placed in cylindrical vials for measurements. Solutions with pyridine concentrations 0, 20, 40, 60, 80, and 100 were used for calibration ($n = 6$ samples) and the remaining fifteen concentrations were used for validation of the model. Similarly to glucose, a total of 10 spectra ($m = 10$) were acquired for each sample. The acquisition time, t_{ac}, for each spectrum was 1 second, with $P = 275$ mW at the sample, and excitation wavelength, λ_{ex}, of 785 nm. Normal Raman spectra of pyridine and pyridine-D_5 are shown in Fig. 3b. The data for calibration and validation were partitioned in the same way as for all methods.

Analysis

The MATLAB (MathWorks, INC., Natick, MA) platform was used for all calculations. MATLAB Optimization Toolbox was used for solving the RO model. PLS models were calculated using PLS_Toolbox 5.5 (Eigenvector Research, INC., Wenatchee, WA). Prior to analysis all SER glucose spectra were normalized. The slowly varying background, commonly seen in SERS experiments, was removed by subtracting a fourth-order polynomial fit. This method greatly reduced varying background levels with minimum effect on the SERS peaks. Raw normal Raman spectra of pyridine/pyridine-D_5 mixtures were used, *i.e.*, they were not preprocessed prior to analysis.

Conclusion

We presented an RO model to relate spectral measurements to quantitative concentrations of analytes being measured. The model is tractable, and it produces a range, *i.e.* minimum and maximum prediction values for each concentration. We demonstrate the performance of the model on a set of data acquired from mixtures of pyridine and pyridine-D_5 at various concentrations as well as a SERS based glucose sensor. To evaluate the performance of the model, we compare the RO prediction ranges to prediction intervals generated from the PLS regression. For both the glucose and the pyridine systems, we find that the RO model provides better prediction estimates. The prediction range from RO either included the actual concentration value, or missed it by a very small error. Simple multivariate linear regression was also performed and the prediction estimates and prediction intervals were computed. This occasionally generated negative values at lower concentrations and ranges that produce negative values. The RO model presented here provides a new approach to sensor range prediction that can be applied in this field. The large prediction ranges/prediction intervals (*e.g.* in glucose) can be taken as an indication of desired further improvements in sensor accuracy.

Acknowledgements

Matlab code for calculating PLS prediction intervals and degrees of freedom was generously provided by Lin Zhang and Salvador Garcia-Munoz. Funding for this work was provided by the NIH (4 R33 DK066990-02) and ONR (N00014-09-1-0518).

References

1 P. Bhandare, Y. Mendelson, R. A. Peura, G. Janatsch, J. D. Krusejarres, R. Marbach and H. M. Heise, *Appl. Spectrosc.*, 1993, **47**, 1214–1221.
2 D. Ebrahimi, E. Chow, J. J. Gooding and D. B. Hibbert, *Analyst*, 2008, **133**, 1090–1096.
3 M. L. Griffiths, R. P. Barbagallo and J. T. Keer, *Anal. Chem.*, 2006, **78**, 513–523.
4 S. Serneels and P. J. Van Espen, *Anal. Chim. Acta*, 2005, **544**, 153–158.
5 K. E. Shafer-Peltier, C. L. Haynes, M. R. Glucksberg and R. P. Van Duyne, *J. Am. Chem. Soc.*, 2003, **125**, 588–593.
6 K. S. Lilley and P. Dupree, *J. Exp. Bot.*, 2006, **57**, 1493–1499.
7 B. Norden, P. Broberg, C. Lindberg and A. Plymoth, *Chem. Biodiversity*, 2005, **2**, 1487–1494.
8 A. J. Berger, I. Itzkan and M. S. Feld, *Spectrochim Acta, Part. A*, 1997, **53**, 287–292.
9 L. A. Marquardt, M. A. Arnold and G. W. Small, *Anal. Chem.*, 1993, **65**, 3271–3278.
10 M. C. Denham, *J. Chemom.*, 1997, **11**, 39–52.
11 *Molecular Spectroscopy Standards and Separation Science Standards*, ASTM, West Conshohocken, 1998, vol. E, pp. 1555–1505.
12 K. Faber and B. R. Kowalski, *Chemom. Intell. Lab. Syst.*, 1996, **34**, 283–292.
13 S. Serneels, P. Lemberge and P. J. Van Espen, *J. Chemom.*, 2004, **18**, 76–80.
14 A. Phatak, P. M. Reilly and A. Penlidis, *Anal. Chim. Acta*, 1993, **277**, 495–501.
15 L. Zhang and S. Garcia-Munoz, *Chemom. Intell. Lab. Syst.*, 2009, **97**, 152–158.
16 A. Ben-Tal and A. Nemirovski, *Oper. Res. Lett.*, 1999, **25**, 1–13.
17 L. E. Ghaoui, F. Oustry and I. Herve, *SIAM J. Matrix Anal.*, 1997, **4**, 1035–1064.
18 A. Ben-Tal and A. Nemirovski, *Math. Program.*, 2008, **112**, 125–158.
19 D. Goldfarb and S. Liu, *Math. Program.*, 1991, **49**, 325–340.
20 S. Mehrotra and J. Sun, *SIAM J. Numer. Anal.*, 1991, **28**, 529–544.
21 Y. Nesterov and A. Nemirovski, *Interior-Point Polynomial Algorithms in Convex Programming*, Society for Industrial & Applied Mathematics, Philadelphia, 1994.
22 O. Lyandres, N. C. Shah, C. R. Yonzon, J. T. Walsh Jr., M. R. Glucksberg and R. P. VanDuyne, *Anal. Chem.*, 2005, **77**, 6134–6139.

23 O. Lyandres, J. M. Yuen, N. C. Shah, R. P. VanDuyne, J. T. Walsh and M. R. Glucksberg, *Diabetes Technol. Ther.*, 2008, **10**, 257–265.
24 W. C. Shih, K. L. Bechtel and M. S. Feld, *Anal. Chem.*, 2007, **79**, 234–239.
25 A. J. Berger, T.-W. Koo, I. Itzkan and M. S. Feld, *Anal. Chem.*, 1998, **70**, 623–627.
26 K. H. Hazen, M. A. Arnold and G. W. Small, *Anal. Chim. Acta*, 1998, **371**, 255–267.
27 M. J. Mattu, G. W. Small and M. A. Arnold, *Appl. Spectrosc.*, 1997, **51**, 1369–1376.
28 G. W. Small, M. A. Arnold and L. A. Marquardt, *Anal. Chem.*, 1993, **65**, 3279–3289.
29 M. A. Arnold, G. W. Small, D. Xiang, J. Qui and D. W. Murhammer, *Anal. Chem.*, 2004, **76**, 2583–2590.
30 J. Chen, M. A. Arnold and G. W. Small, *Anal. Chem.*, 2004, **76**, 5405–5413.
31 A. Lorber, K. Faber and B. R. Kowalski, *Anal. Chem.*, 1997, **69**, 1620–1626.

