# Investigation of constrained regularization for the development of a robust and clinically accurate multivariate calibration procedure

18.086: Computational Science and Engineering II
Spring 2008

Ishan Barman
G.R. Harrison Spectroscopy Laboratory
Massachusetts Institute of Technology

# 1. INTRODUCTION

Blood analytes provide valuable information for the diagnosis of many fatal diseases and abnormal health conditions. Development of painless and convenient methods for measurement of such analytes has received considerable attention. Glucose detection[1], in particular, has been studied extensively over the last couple of decades as it has widespread implications in the control and management of diabetes. As diabetes has no known cure, tight control of glucose levels is critical for the prevention of such complications[2]. Given the necessity for regular monitoring of blood glucose, development of non-invasive glucose detection devices is essential to improve the quality of life of diabetic patients.

Our laboratory has pioneered the development of a non-invasive glucose sensor based on the principles of NIR (near-infrared) Raman spectroscopy[3,4]. NIR Raman Spectroscopy combines the substantial penetration depth of NIR light with the excellent chemical specificity of Raman spectroscopy. Additionally, it enables the simultaneous determination of multiple blood analytes. The underlying principle of this technology is that the backscattered Raman photons, obtained by focusing a monochromatic source of light on biological tissue, have characteristic signatures of the analytes present in the tissue. The analytes of interest could be cholesterol, fats, proteins and glucose among a host of others. NIR Raman spectroscopy thus provides an excellent tool to meet the challenges involved in not only monitoring glucose levels but also diagnosing various pathophysiological conditions, such as cancer and atherosclerotic plaque.

A number of major technical challenges, however, impede the development of a viable NIR Raman spectroscopic glucose sensor. Significant among these is the lack of robust and accurate information extraction algorithms, which can be applied to the spectra acquired *in vivo*. The poor signal to noise ratio of the Raman features makes a difficult task even more so. Further more, while the ability of Raman spectroscopy to detect multiple analytes simultaneously is a tremendous advantage, the extraction of analytical information about each of the constituents is not trivial – as in practice, most of these analytes tend to give overlapping features which do not readily lend themselves to quantitative predictions.

In order to determine the concentration of the various analytes in a complex chemical system, multivariate calibration techniques are usually employed. Multivariate calibration algorithms, which can be utilized in a wide range of possible scenarios in terms of knowledge of the system under consideration, take the full-range spectrum into account[5]. This is critical as the complex spectra that are acquired *in vivo* cannot provide useful information if only a limited number of wavelengths are selected for analysis.

The existing set of calibration algorithms can be broadly classified into explicit and implicit schemes. The explicit calibration procedures provide highly accurate models, but require complete knowledge of the constituents of the system and their corresponding (Raman) spectra. This limitation renders it of little value in most real life biomedical applications, where delineating the system constituents is a major task in itself. Implicit calibration, on the other hand, does not require knowledge of the constituent spectra, and can be used in applications where concentration information about the analyte of interest is known (or can be determined) in a set of reference samples. These calibration methods, however, are unable to distinguish between legitimate correlations between spectra and concentrations and spurious correlations,

such as that obtained by system drift and high degree of (unrelated) covariance between constituents. Nevertheless, these techniques have gained widespread acceptability and function as the gold standard of the day.

To alleviate the problems associated with the implicit calibration techniques, our laboratory has recently developed two hybrid calibration schemes, namely hybrid linear analysis (HLA) and constrained regularization (CR)[6,7]. CR, which provides more flexibility in the incorporation of the prior information than HLA, has been shown to significantly outperform the implicit calibration techniques in certain studies, where a reasonably high degree of correlation between at least two constituents of the samples is intentionally maintained (called correlated samples herein).

In this study, our aim is to investigate the applicability of CR in more general situations, where sample constituents are uncorrelated (termed as uncorrelated samples). This would more closely mimic the prevalent situation in any glucose clamping or point of care clinical validation study. In this context, the concept of confidence maximization has been introduced. Confidence maximization is defined here as the weighted selection of samples and wavelengths in the calibration procedure, where larger weights are assigned to those samples and wavelengths that have greater probability of providing reliable and accurate constituent-specific information.

In this article, we review the basic principles of the various multivariate calibration schemes that are pertinent to the introduction of the constrained regularization method. The theory of CR is then extended to include the formalism of confidence maximization. Next, an experimental study is presented to investigate the relative advantages in the application of CR over existing implicit methods for uncorrelated samples. Finally, we present results showing the substantial reduction in the prediction error that can be obtained by incorporating confidence maximization principles into the current CR formalism, especially for uncorrelated samples.

## 2. THEORY

Establishment of relationships between measurements made on a system and the underlying state of the system is a key component of any experimental science. In chemistry, this idea has developed into a whole field of study, known as chemometrics. Chemometric analysis and interpretation of instrumental data utilizes well-established mathematical and statistical methods, such as design of experiments, calibration and pattern recognition, to name a few.

In spectroscopy, the primary application of chemometrics is in calibration, and normally involves using one type of measurement to predict the value of an underlying property or parameter. The traditional method of calibration was univariate calibration, which involved calibration of a single variable (e.g. spectroscopic intensity at a single wavelength) to another variable (e.g. concentration). However, univariate calibration proves to be wholly inadequate for analyzing complex chemical systems where the spectrum of each constituent contributes to the overall spectra of the sample. This motivated the use of multivariate calibration, where instead of using a single variable, several variables (e.g. spectroscopic intensities at 100 wavelengths) are calibrated to one or more variables. These methods provide an improvement in the estimate of the underlying parameter or property due to the effect of averaging (of the noise and errors). Although extension from a

univariate to a multivariate scheme might seem like a natural extension of the space in which the formalism is defined, a new class of techniques is actually required to deal efficiently with the problem.

It is pertinent to note, however, that irrespective of the specific method that is being used to tackle the problem, the basic procedure remains the same and can be stated in the following manner (Fig. 1). From a reference mixture of compounds of known concentrations (typically called a calibration or training set), one seeks to establish a relationship between these known concentrations and the measured spectra. Once this model is established, it can be used, in conjunction with the acquired spectra, to predict the unknown concentrations of the same compounds in future samples (also called the prediction or test samples/set). Given this framework, it is not surprising that for any calibration procedure to be effective in prospective prediction, the range of samples used to develop the calibration set must be sufficiently representative of all future samples that may have to be analyzed by this model.

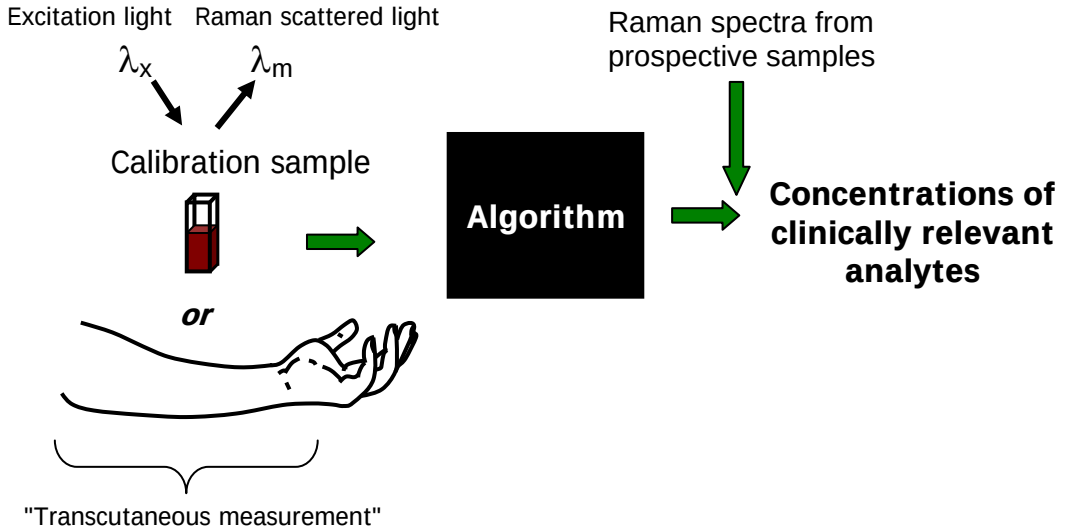


Fig. 1: Basic outline of a multivariate calibration algorithm in our experimental framework.

## 2.1. Multivariate Calibration

Multivariate calibration is a powerful analytical technique for extracting analyte concentrations in complex chemical systems that exhibit linear response. The "linear response" prerequisite appears as multivariate analysis employs fundamental linear algebra techniques. In spectroscopy, it has been shown linear additivity works well and this is especially true in analytical chemistry calibration[8].

The framework for calibration described in the above paragraphs and shown in Fig. 1 can be reduced to set of linear equations as represented in matrix form below:

$$\mathbf{S}_{j\times\lambda} = \mathbf{C}_{j\times p}\mathbf{P}_{p\times\lambda} + \mathbf{E}_{j\times\lambda} \quad [1]$$

where S is the matrix of acquired spectra from the sample set,
C is the concentration matrix of the sample set,
P is the pure component spectra of the constituents in the sample and
E is the error matrix associated with the spectral block,

j, λ, p are the number of samples, wavelengths and constituents respectively. (In this article, all matrices are denoted by uppercase boldface type, vectors by lowercase boldface type and scalars by lowercase letters.)

Eqn (1) is just a statement of the linear combination assumption, where the weight of each constituent is equal to its concentration in the specific sample. The representation of the calibration model in the form of Eqn (1) is called the classical calibration model, where the spectra are related to the concentrations and not vice versa. This form has the advantage of a direct physical interpretation but has two important drawbacks. The first difficulty is that our goal is to be able to predict concentrations from the spectra and not the other way around, as represented in Eqn. (1). Secondly, and more importantly, the classical calibration model assumes that all the errors are in the spectra block. However, it is recognized in the scientific community that the greatest source of error is generally in sample preparation such as dilution, weighting and extraction, rather than instrumental reproducibility, so the measurement of a concentration is likely to be less certain than the measurement of spectral intensity. The inverse calibration model addresses these problems (Eqn. (2)). This is specifically more appropriate for implicit calibration models where only the analyte of interest is considered (rather than all the constituents present in the sample) and is written here as such:

$$\mathbf{c}_{j\times 1} = \mathbf{S}_{j\times\lambda}\,\mathbf{b}_{\lambda\times 1} + \mathbf{e}_{j\times 1} \qquad [2]$$

where b is the regression vector and

e is the vector of errors associated with the concentration measurements.

As the b-vector correlates the acquired spectra to the concentrations of *the* analyte of interest, it is expected that this vector will contain features of the pure spectrum of the analyte of interest. However, due to the presence of spectral interferents, it will not be exactly the same as the pure spectrum.

It is to be noted that despite the problems associated with the classical model, it is frequently used due to its physically intuitive nature. We will come across both the models in the following sections as we review the various multivariate calibration strategies.

## 2.1.1. Explicit Calibration

The explicit calibration techniques require that all pure component spectra must either be known or pre-calculated before the determination of the concentrations of the constituents in the samples is undertaken. The whole class of techniques revolves around multiple linear regression (MLR) and can be thought of as a natural extension to univariate linear regression.

If the spectra of all pure components are known (if all the rows of the P matrix can be populated), then ordinary least squares (OLS) is usually applied. From the classical calibration model of Eqn (1), one can then obtain a least squares solution for the concentrations of the constituents in the samples in the following manner:

$$\mathbf{C}_{LS} = \mathbf{S}\mathbf{P}^{T}(\mathbf{P}\mathbf{P}^{T})^{-1} \qquad [3]$$

In fact, one can observe that this solution does not really have a calibration step in contradiction to the general calibration procedure laid out earlier. This solution exists for $\lambda \geq p$ – otherwise $\mathbf{P}\mathbf{P}^{T}$ is singular. In almost all applications, however, the number of wavelengths sampled (~1000) is greater than the number of constituents in the sample (~10) and consequently, OLS can be applied whenever the spectra of a unit concentration of the pure constituents (basis spectra) are known.

To demonstrate the use of this methodology, an experiment was performed with two tissue mimicking phantoms where each of the constituents was well characterized spectrally. After the OLS inversion was performed to obtain the concentrations, the least squares fits to the actual spectra were calculated using the aforementioned concentrations. This is shown in Fig. 2. The x and y-axis represent the intensity (photon count) and the Raman shift ($cm^{-1}$) respectively. It can be observed that the least squares fit matches very closely to the actual spectra observed – showing that the noise in the spectral measurements was minimal.

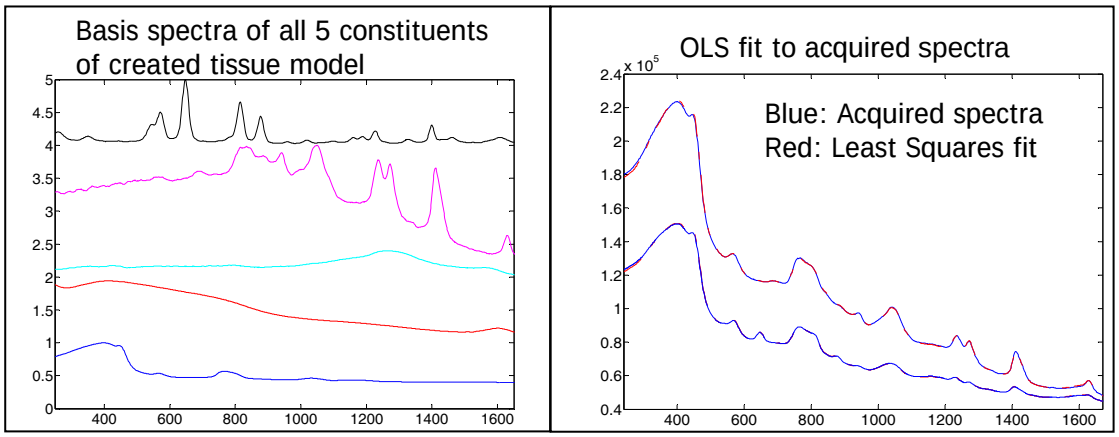


Fig. 2: Application of OLS to tissue mimicking phantoms created *in vitro*

However, the biggest disadvantage with this method is that it is very sensitive to any errors in the measurement of the basis spectra. Moreover, instrumental performance may vary from day to day so that the pure spectra measured on Day 1 may be different – sometimes substantially so – than the pure spectra measured on Day 2. In order to overcome these two issues, one can measure spectra of the mixtures in known concentrations and use these as a calibration set to obtain the basis spectra matrix, **P**. This is commonly known as the classical least squares (CLS) technique.

$$\mathbf{P}_{LS} = (\mathbf{C}^T\mathbf{C})^{-1}\,\mathbf{C}^T\mathbf{S} \qquad [4]$$

This exists only when j≥p, which again is true for most applications as the number of samples in the calibration set (~50) will exceed the number of pure components in each sample. The calculated $\mathbf{P}_{LS}$ from Eqn (4) can be used in place of **P,** in Eqn (3), to determine the concentrations of the constituents in unknown samples. In principle, the CLS strategy can be thought of as a derivative of OLS except that in the former the concentrations of all the constituents are known (in the calibration set) and in the latter the pure spectra of the constituents are likewise given.

Despite the simplicity of the above approaches, MLR is rarely used in biomedical studies, as briefly mentioned in Sec-1. This is due to the obvious limitation of having to know the concentrations of all significant compounds in the calibration set or the pure spectra of all the compounds. For example, if we have information on the concentrations of only m out of n constituents in a calibration set (where n>m) then the m predicted spectra will contain features of the spectra of the remaining n-m components, distributed among the m known components, and the concentration estimates will contain large and fairly unpredictable errors.

## 2.1.2. Implicit Calibration

The lack of complete knowledge of basis spectra motivates the use of implicit calibration strategies in a large number of analytical chemistry applications including almost all biomedical and clinical cases. The inverse calibration model of Eqn (2) is predominantly used as the fundamental idea is to obtain a regression vector, given the spectra of the mixture samples and the concentrations of only the analyte of interest in the calibration set. In other words, one would like to obtain the least squares solution for the b-vector by using the following formula:

$$\mathbf{b} = (\mathbf{S}^T\mathbf{S})^{-1}\mathbf{S}^T\mathbf{c} \quad [5]$$

Once the b-vector is obtained it can be used in conjunction with the spectrum of a future sample to predict the concentration of the analyte of interest in that sample using Eqn. (6).

$$c_{un} = \mathbf{s}_{un}.\mathbf{b} \quad [6]$$

The problem is that the inversion of Eqn. (5) can take place only for $j \geq \lambda$, i.e. the number of samples must be larger than the number of wavelengths for $\mathbf{S}^T\mathbf{S}$ to be invertible. As this is unrealistic in most situations, one must figure out an optimal way to compress the spectral data into fewer data points.

Let us assume that **S** can be expressed as a linear combination of base spectra, **Q**, which are not necessarily the basis spectra of the constituents of the system. Each column of **Q** represents a base spectrum. To write this in a mathematical form, we include weightings of each base spectrum for each sample in the form of the score matrix, **T**.

$$\mathbf{S} = \mathbf{T}\mathbf{Q}^T \quad [7]$$

The compression step can now be thought of as the transformation from $\mathbf{S}_{j\times\lambda}$ to $\mathbf{T}_{j\times p}$ such that the inversion step in Eqn (4) becomes straightforward. This is performed by projecting the original λ variables onto a new set of p axes, **W**. Each column of W forms one of the projection axes. The transformation can be written as:

$$\mathbf{T}_{j\times p} = \mathbf{S}_{j\times\lambda}\,\mathbf{W}_{p\times\lambda} \quad [8]$$

The rest of the analysis can then take place in the score-space than in the spectrum-space. Since linearity still holds, the concentrations can be predicted using the following equation:

$$\mathbf{c} = \mathbf{T}\,\mathbf{v} \quad [9]$$

where v is the regression vector in the score-space.

A least squares solution for v can now be obtained as $\mathbf{T}^T\mathbf{T}$ has a well-defined inverse as $j \geq p$. Using the regression vector with the score of the future sample, one can obtain the concentration of the constituent under consideration in the future sample.

$$c_{un} = \mathbf{t}_{un}\,\mathbf{v} = \mathbf{s}_{un}\mathbf{W}(\mathbf{T}^T\mathbf{T})^{-1}\,\mathbf{T}^T\mathbf{c} \quad [10]$$

While all implicit calibration methods use this formalism, they differ in the central idea of selection of **W**, i.e. the optimal way of data compression from λ to p variables. Without going into the details of any of them, we will visit the key concepts of two such widely used strategies, namely principal component regression (PCR) and partial least squares (PLS). These two methods surpass the capability of the direct method (inverse least squares (ILS)), where only p wavelengths are chosen to compress $\mathbf{S}_{j\times\lambda}$ to $\mathbf{T}_{j\times p}$ or all the wavelengths are 'binned' to p variables. Clearly, this reduces the effectiveness of the multivariate approach itself and yet does not guarantee that the transformed matrices are robustly invertible.

In PCR, principal component analysis (PCA) of the spectra is first undertaken followed by the regression step. For the set of spectra in the calibration set, one determines the set of p axes along which the spectral intensity variances are maximized. In other words, these axes (also called the principal components) explain most of the variation in the data. Not unexpectedly, the PCs are the eigenvectors of $\mathbf{S}^T\mathbf{S}$ matrix. The critical step here is to select the number of significant principal components, which should ideally correlate (or match) with the number of constituents in the sample. However, due to the presence of noise in the system it could vary from this value. Nevertheless, one must have certain *a priori* knowledge to make a decision of the number of significant components that should be used for building the calibration algorithm. To determine how many principal components are significant, one can apply a variety of methods, which all look towards optimizing the calibration model. For example, one could look for the p largest eigenvalues and select the corresponding eigenvectors (principal components). Once the PCs are chosen, the regression step (Eqn. (9)) can be employed to obtain the regression vector. However, due to this two-step approach, it minimizes the residual in spectrum fitting only without considering the variance in the concentration data of the calibration set. Stating it in another way, PCR assumes that all the errors are in the spectral block.

In contrast to PCR, PLS assumes that the errors in spectral and concentration block are of equal significance. This is a more reasonable assumption, given that the spectra in modern day labs are indeed more reproducible and accurate than concentration measurements. The formalism dictates that the residual in concentration fitting is minimized. In other words, instead of determining the axes along which the variance of the spectral intensity is maximized, we now look for the variables (here called the loading vectors) along which covariance between the spectral and the concentration blocks is maximized. Like PCR, however, PLS also forces the user to select the number of significant loading vectors based on *a priori* information or model validation strategies. Inclusion of less than ideal number of loading vectors introduces undesirable averaging of the spectral features and the inclusion of too many loading vectors makes it liable to 'over-fitting'. Over-fit models produce excellent results on the calibration set but, due to the incorporation of noise in the calibration set data, produce poor results on any arbitrary prediction set.

Given equal information regarding the number of significant PC/loading vectors, PLS outperforms PCR in all cases. In cases where the noise in both spectral and concentration data is negligible the two methods provide similar performances. As a consequence, PLS is currently the preferred method for calibration for chemometricians all over the world and forms the gold standard in this field of application.

One practical application of the PLS technique is shown in Fig. 3. A glucose clamping study was carried out on a dog for 3 hours and Raman spectra were collected over the duration. Every 10 minutes, a small volume of the dog's blood was withdrawn to test for the blood glucose level. Subsequently, the entire spectra and concentration data was divided into calibration and prediction sets. The regression vector was developed from the calibration set using the PLS technique and then applied on the prediction set.

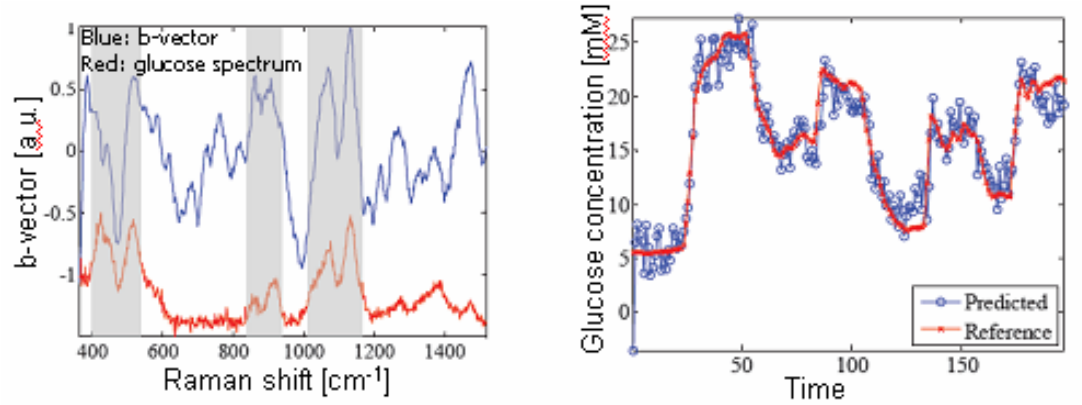


Fig. 3: (left) Comparison of b-vector (regression vector) and glucose spectrum; (right) Predicted and reference concentrations of blood glucose over three hour period

It is clear from Fig. 3 that the regression vector contains features from the Raman spectrum of pure glucose especially in the ranges highlighted. The other features are from spectral interferents. The prediction of glucose concentration closely resembles the actual reference concentration in the trends followed during the clamping study. At hypoglycemic levels (below 6 mM), the absolute errors in prediction are pretty significant. The prediction accuracy is reasonable in the hyperglycemic levels, as expected because at higher concentration levels, the role of spectral interferents will decrease appreciably.

Despite the reasonable prediction performance shown in Fig. 3, PLS suffers from some shortcomings, the most prominent of which is that it can be heavily influenced by spurious correlations arising from system drift and (unintentional) covariance among constituent concentrations in the calibration set[9]. Explicit calibration techniques, on the other hand, do not have such deleterious effects arising from system correlations due to the prior information that is fed into the model. This provides a strong motivation that if one could incorporate prior information into an implicit calibration technique, one could possibly get rid of the spurious correlations that significantly mar the prediction capability of the model. Hybrid calibration methods were developed to bridge the gap between robust (implicit) and accurate (explicit) calibration strategies.

## 2.1.3. Hybrid Calibration

Hybrid calibration techniques try to incorporate prior information that is available in the specific application area into the implicit calibration strategies. The aim is to enhance the accuracy of the prediction and minimize the effect of spurious correlation of an implicit calibration scheme. Hybrid schemes have been investigated by various research groups and several references can be found in the literature[6,10]. Hybrid linear analysis (HLA), for example, uses the pure spectrum of the analyte of interest to build the model. As the pure spectrum of glucose can be easily measured using a reference sample at unit concentration, the idea of incorporating the spectrum of analyte of interest is sound for our application. In HLA, the central idea is to remove from each acquired spectra the contribution due to the analyte of interest by subtracting from the former the product of the concentration of the analyte (in each corresponding sample)

and its pure spectrum. PCA is, then, performed on the subtracted spectra to obtain the principal components of the background spectra (spectra with the features of the analyte of interest removed). To this set of PCs is added the spectrum of the analyte of interest and this complete set is used like the matrix of pure spectra of the chemical constituents, as in OLS. However, due to the reliance on direct spectral subtraction, the intensity and profile of the pure spectrum is of extreme importance in HLA. In other words, any inaccuracy in the pure spectrum can set off major errors propagating through the algorithm. This causes instability in the system and reduces the robustness of the algorithm.

To overcome this lack of robustness in HLA, our laboratory has recently developed the novel methodology of constrained regularization (CR)[7]. Since the least squares problem to obtain the regression vector from the inverse calibration model (Eqn (5)) cannot be solved unless the number of samples is greater than or equal to the number of wavelengths sampled, we regularize the problem such that the inversion operation can be successfully undertaken. In other words, Eqn (5) represents an underdetermined problem, with greater number of unknowns (λ) than there are equations (j) – thus, the problem needs to be regularized for the matrix **S**, which does not have full rank, can be inverted. Regularization pushes the eigenvalues of the $\mathbf{S}^T\mathbf{S}$ matrix away from zero – thereby making it invertible. This new regularized problem statement can be written as:

$$\Phi(\Lambda) = \|\mathbf{Sb} - \mathbf{c}\|^2 + \Lambda\|\mathbf{b}\|^2 \qquad [11]$$

where Φ is the function that needs to be minimized and Λ is the regularization parameter.

It is pertinent to note that the addition of the regularization term has turned the linear least squares problem (which is the minimization of the first term in the RHS of Eqn. (11)) to a two squares minimization problem. However, the introduction of the second term in the current form implies we want to minimize the norm of the regression vector, which is not completely true. We would like to minimize not the norm of the regression vector but the norm of the regression vector minus the pure glucose spectrum. This redefinition of the constraint ensures that the regression vector should tend towards the glucose spectrum. The full CR problem statement can now be expressed as:

$$\Phi(\Lambda, \mathbf{b}_0) = \|\mathbf{Sb} - \mathbf{c}\|^2 + \Lambda\|\mathbf{b} - \mathbf{b}_0\|^2 \qquad [12]$$

where $b_0$ gives the spectral constraint for the regression vector.

While it has been shown that CR significantly outperforms PLS for correlated samples[7], the comparison of performance metrics for these two algorithms in more general situations, where sample constituents are uncorrelated, has not been investigated. It is expected that the performance metrics of the two algorithms will not be very different for the uncorrelated case. This motivates the application of sample and wavelength selection concepts to CR.

We propose to develop confidence maximization principles which can be used as a natural extension to the existing theory of CR in further enhancing prediction accuracy. Confidence maximization, as defined earlier, represents the assignment of weights to samples and wavelengths in the calibration procedure. The amount of weight assigned, to a sample or wavelength, is proportional to its probability of providing reliable and accurate constituent-specific information.

This step is crucial for the improvement of CR, in particular, because the application of the $\mathbf{b}_0$ spectral constraint must be weighted to favor those wavelengths where glucose shows significant Raman features. At other wavelengths, the contributions from noise might be adversely affecting rather than enhancing the capability of multivariate calibration. In essence, therefore, one can think of the wavelength selection principle as incorporation of prior information.

Sample selection is based on the type of study being performed. If an *in vitro* phantom study is performed, more weight is given to those samples in which the glucose concentration is higher. This is in line with the argument provided for the PLS results in the glucose clamping study on the dog, i.e. at higher concentrations of the analyte of interest the role of spectral interferents reduces considerably. Moreover, at lower concentrations, the noise in the signal might have a more adverse role. In the case of an *in vivo* glucose clamp study, larger weights are assigned to the measurements made when the glucose is stable at a particular concentration rather than when it is rising or falling rapidly. It is evident that measurements made at these stable levels are much more reliable and reproducible than those made between two stable levels, where complications arising from the lag time between the interstitial fluid glucose and plasma glucose concentrations become significant.

In order to incorporate confidence maximization into the problem statement of Eqn (11), we need to introduce wavelength ($\mathbf{W}_\lambda$) and sample ($\mathbf{W}_S$) weighting matrices. The assignment of suitable elements to populate these two matrices is based on the physical criteria stated above. One such implementation is shown in Sec-4, where the analysis of the *in vitro* tissue phantom study is undertaken.

The new problem statement for obtaining the regression vector using constrained regularization with confidence maximization is given by:

$$\Phi(\Lambda, \mathbf{b}_0) = \left\|\mathbf{Sb} - \mathbf{c}\right\|_{W_S}^{\ 2} + \Lambda \left\|\mathbf{b} - \mathbf{b}_0\right\|_{W_\lambda}^{\ 2} \qquad [13]$$

where the first and second terms are evaluated with respect to the weighting matrices $\mathbf{W}_S$ and $\mathbf{W}_\lambda$ respectively. This can be re-written in the following manner:

$$\begin{bmatrix} S^T & I \end{bmatrix} \begin{bmatrix} W_S & 0 \\ 0 & \Lambda W_\lambda \end{bmatrix} \begin{bmatrix} S \\ I \end{bmatrix} \hat{b}_\Lambda = \begin{bmatrix} S^T & I \end{bmatrix} \begin{bmatrix} W_S & 0 \\ 0 & \Lambda W_\lambda \end{bmatrix} \begin{bmatrix} c \\ b_0 \end{bmatrix} \qquad [14]$$

Eqn. (14), in turn, yields the following expression for the 'best' estimate of $b_\Lambda$:

$$\hat{b}_\Lambda = \left(S^T W_S S + \Lambda W_\lambda\right)^{-1} \left(S^T W_S c + \Lambda W_\lambda b_0\right) \qquad [15]$$

Eqn. (15) gives the final expression for the least squares regularized estimate of the regression vector subject to confidence maximization matrices. The estimated regression vector can now be used in conjunction with the spectra of the future samples to predict the unknown concentrations of the analyte of interest, using Eqn. (6).

## 3. MATERIALS AND METHODS

An 830-nm external cavity diode laser was used as the Raman excitation source for our experimental studies. The laser beam was passed through a laser line filter and focused onto the sample. A photodiode was placed along the excitation path to monitor the intensity variations of the laser source such that the variations in source intensity can be correctly accounted for. The back-scattered light was collected and directed by the paraboloidal mirror towards a holographic notch filter to reduce the

Rayleigh peak intensity. The light exiting the notch filter was input to f/1.4 spectrometer. A liquid nitrogen cooled CCD detector, having high quantum efficiency and low shot noise, was used to capture the spectra. All the optical elements were NIR anti-reflection coated to enhance the throughput of the system.

An *in vitro* tissue phantom study was undertaken to compare and contrast the prediction accuracy of CR and PLS. Although a previous study had shown the effectiveness of CR in correlated samples, no such study for uncorrelated samples in turbid media has been reported. It has to be emphasized that uncorrelated samples in turbid media provide the most realistic test bed for *in vivo* applications. The biggest motivation, however, was to determine the reduction of prediction error, if any, that could be expected if CR was extended to include confidence maximization principles.

The tissue phantoms used in this study were prepared using a mixture of glucose, creatinine, intralipid, and ink in water. This set of constituents enables the phantom to have similar scattering and absorption properties (turbidity) as human tissue. 50 such tissue phantoms were prepared with completely randomized concentration profiles, i.e. there was little or no correlation between the concentrations of any two constituents across all the phantoms. However, as the tissue phantoms were artificially created, all the concentrations were known *a priori*, so any further independent measurement of concentration was not necessary.

The data from the 50 tissue phantoms were separated into calibration (36 samples) and prediction (14 samples) sets. Each method - PLS, CR, CR with sample selection, and CR with sample and wavelength selection – was applied first on the calibration set to create a calibration algorithm. To this end, the acquired spectra (after having been corrected for the presence of any cosmic rays and smoothed using a Savitzky-Golay algorithm[11]) were used in conjunction with the known concentrations of glucose in the calibration set. Since the number of constituents in each sample was known for this study, no optimization was performed on the calibration set to determine the number of loading vectors in PLS. After the calibration algorithm was established (the b-vector was obtained), it was used to predict the concentrations of glucose in the future samples in the prediction set. The predicted concentration was compared to the actual concentration of glucose in the samples and a prediction error – root mean square error of prediction (RMSEP) – was calculated using Eqn. (16).

$$RMSEP = \sqrt{\frac{\sum_{i=1}^{n}\left(c_{pred,i} - c_{ref,i}\right)^2}{n}} \qquad [16]$$

where n is the number of samples in the prediction set.

The splitting into calibration and prediction sets were iterated 500 times such that the mean RMSEP calculated gave an accurate estimate of the prediction error that can be expected using each calibration strategy.

## 4. RESULTS

The spectra acquired from the 50 tissue phantoms are shown in Fig. 4. Clearly, the spectra contain not only Raman features of the constituents but also a broad fluorescence background, which greatly impedes the detection of analytes using Raman spectroscopy. Moreover, although the concentrations of the constituents were randomized, one cannot observe with the naked eye distinct differences in

characteristic features from one spectrum to another. Nevertheless, there does exist, on a finer scale, local intensity variations at different Raman shifts because of which the multivariate calibration techniques are able to achieve reasonable prediction accuracy.

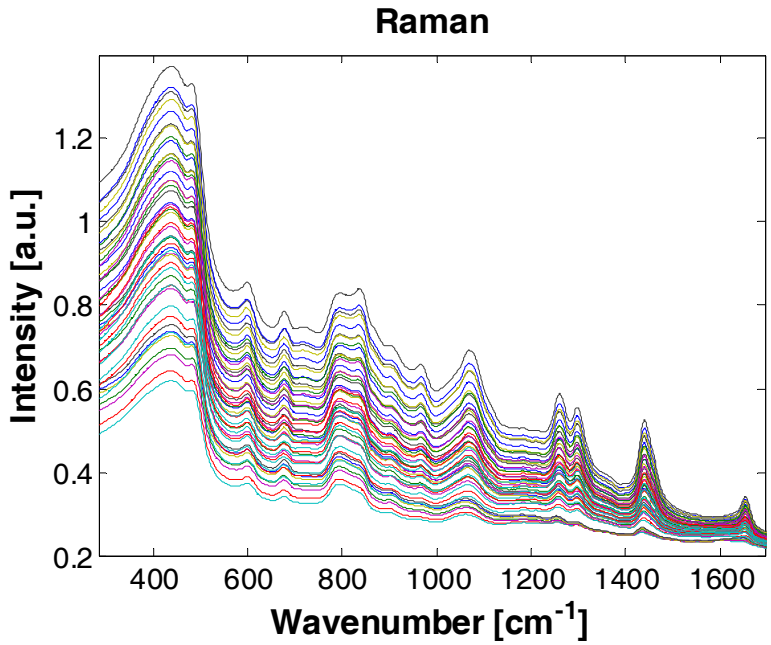


Fig. 4: Acquired Raman spectra from the 50 tissue phantoms created *in vitro*.

The spectra were first pre-processed as detailed in Sec-3. PLS and CR were then applied to the datasets (spectra and concentrations) of the phantoms to determine the RMSEP values in each case. The results are shown in Fig. 5.

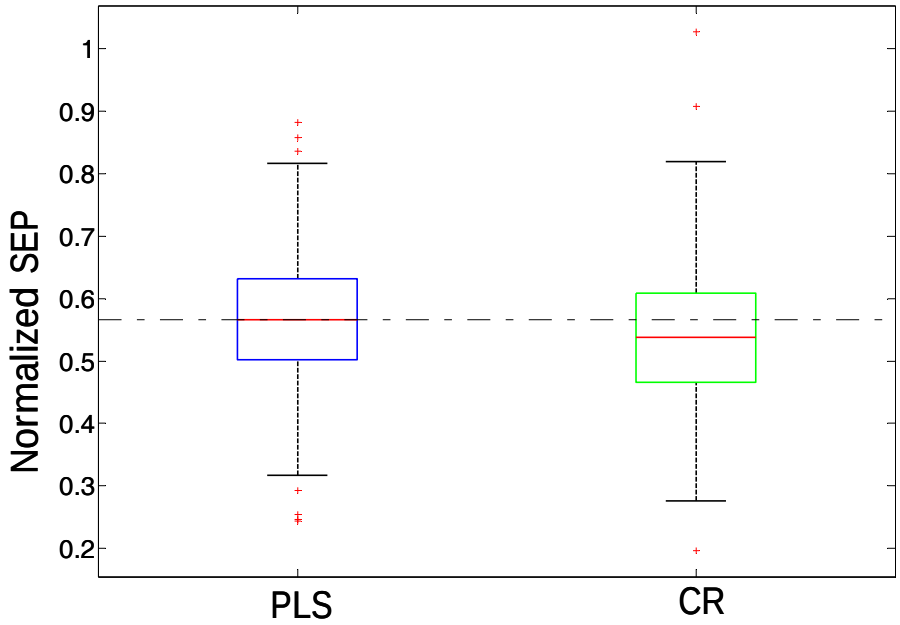


Fig. 5: Box-plot of prediction error for PLS and CR in uncorrelated samples. Values were derived from 500 random splittings of 50 samples into 36 calibration and 14 prediction samples.

It can be observed that there is a reduction in the prediction error when CR is applied rather than PLS. This reduction can be calculated to be around 4.65%. Although this is a statistically significant reduction, as can be understood by the fact that it holds over 500 iterations, the value pales in comparison to that achieved when both these techniques were applied under similar conditions to correlated samples[7]. For the sake of comparison, Fig. 6 reproduces the results stated in the aforementioned article, where the glucose and creatinine concentrations in 50 samples had a $R^2$ value of 0.48. When correlated samples were samples, the reduction in error was found to be more than 20%.

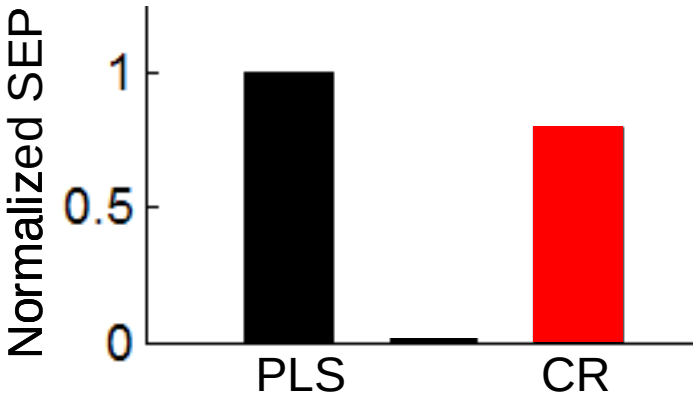


Fig. 6: Plot of the prediction errors of PLS and CR in correlated samples (adapted from Ref [7])

As stated earlier, PLS performs much better when the samples are uncorrelated and thus in a real life situation, where constituents are not likely to have a significant degree of correlation between themselves and are expected to be measured in a turbid medium, the two methods might perform equivalently. The major advantage of CR in such a situation is that there is no need to designate the number of significant components required to create the calibration model. Although this might seem like a minor issue at first glance, the validation of the calibration model created by PLS and particularly of the number of components required to create the best model requires the investment of a huge amount of effort and time. Moreover, there is no one best way of validation of the number of components with various researchers choosing between cross-validation, bootstrap and autoprediction to make their cases.

The results of Fig. 5 and 6 provide us with an excellent motivation to incorporate confidence maximization principles as one would like to work with CR (because of the reasons stated above) but with improved accuracy. To test the effectiveness of the sample and wavelength selection approaches, both CR with sample selection and CR with sample and wavelength selection were applied on the same datasets as used before (the same set of 500 splittings into 36 calibration and 14 prediction samples).

For appropriate sample selection, the assigned weight of each sample in the calibration set was set equal to the glucose concentration in that sample. The rationale behind this is that at higher levels of glucose concentration, the role of spectral interferents and noise is significantly more limited. Wavelength selection was performed by choosing those wavelength ranges (Raman shift ranges, to be precise) where glucose shows important features, namely 400-550 $cm^{-1}$ and 800-1480 $cm^{-1}$. These wavelength ranges were given unit weight while the other wavelengths were

assigned zero weights. Evidently, these weights could be assigned more systematically but the lack of a suitable validation algorithm for these assignments remains the bottleneck in the development of a more systematic weight assignment scheme.

The results of application of CR with sample selection and CR with sample and wavelength selection is shown in Fig. 7, alongside the previous results obtained for PLS and CR. It can be observed that both sample selection and wavelength selection by themselves provide ample benefits in regard to the reduction in the value of RMSEP. In particular, CR with sample and wavelength selection is shown to give an astounding **44% reduction in SEP** value from that given by the existing CR algorithm.

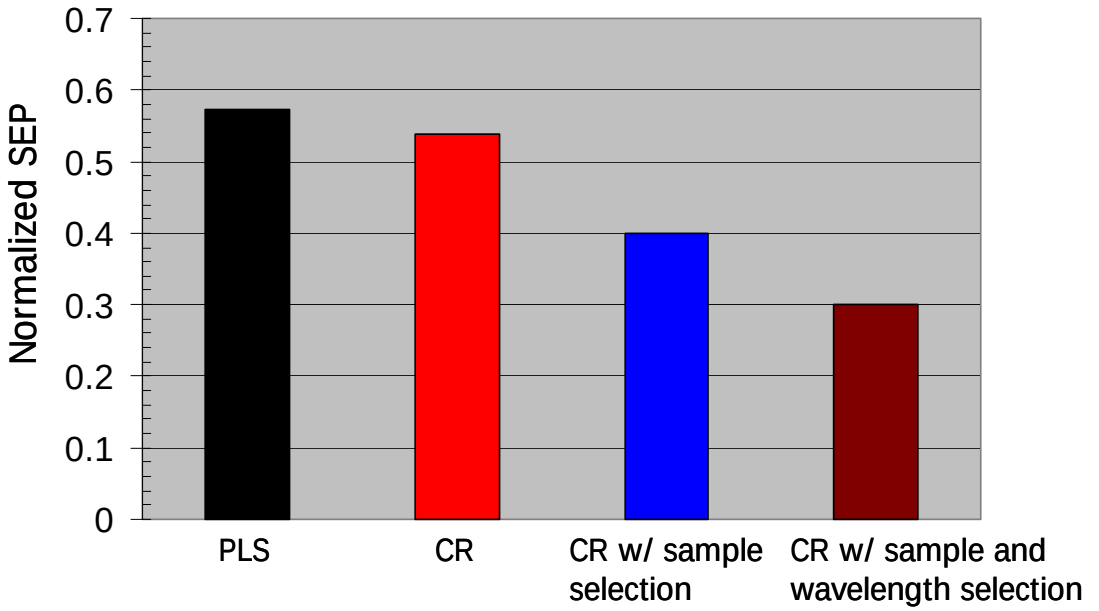


Fig. 7: Plot of the prediction errors for PLS, CR, CR with sample selection and CR with sample and wavelength selection.

CR with confidence maximization, thus retains the robustness and flexibility of CR, but tremendously improves its accuracy. Importantly, no additional knowledge beyond that which is ordinarily available for all *in vitro* and *in vivo* studies is used. As in CR, there is no need to specify the precise number of significant components which should be used in model building. The significant enhancement of prediction capability even in uncorrelated samples means that CR with confidence maximization can be used for all applications, irrespective of correlation between constituents and presence or absence of turbidity in the medium.

## 5. CONCLUSION

In this article, we have introduced confidence maximization as a means of significant enhancement of prediction accuracy for the existing constrained regularization methodology. Based on the minimization of the weighted two squares problem, it retains all the advantages of the original constrained regularization scheme by maintaining the robustness associated with the flexibility in the choice of the regularization parameter.

Our investigations revealed that although CR enjoys a substantial edge in prediction accuracy over PLS for samples where certain constituents have a reasonable degree of correlation between themselves, the difference is not as significant for uncorrelated samples. Both of the confidence maximization approaches, namely sample selection and wavelength selection, are shown to provide distinct benefits - nearly a factor of two reduction in prediction error is observed.

In future work, one would like to develop a set of schemes that can assign optimal weights both for sample selection and wavelength selection. This can provide tremendous benefits in the processing time necessary for a set of spectra. Furthermore, this could be tested across more samples with different constituents to validate the basic principles of confidence maximization. Finally, it is necessary to work towards an explicit formula for the optimal regularization parameter, which has to be initiated by characterizing the noise in the system measurements.

## REFERENCES


[1] J. N. Roe and B. R. Smoller, Critical Reviews in Therapeutic Drug Carrier Systems **15,** 199-241 (1998).

[2] L. M. Tierney, S. J. McPhee, and M. A. Papadakis, *Current Medical Diagnosis & Treatment* (Lange Medical Books/McGraw-Hill 2002).

[3] A. J. Berger, I. Itzkan, and M. S. Feld, Spectrochimica Acta Part a-Molecular and Biomolecular Spectroscopy **53,** 287-292 (1997).

[4] A. M. K. Enejder, T. G. Scecina, J. Oh, M. Hunter, W.-C. Shih, S. Sasic, G. Horowitz, and M. S. Feld, Journal of Biomedical Optics **10,** 031114 (2005).

[5] R. G. Brereton, *Applied Chemometrics for Scientists* (John Wiley & Sons Ltd,, Chichester, West Sussex, England, 2007).

[6] A. J. Berger, T. W. Koo, I. Itzkan, and M. S. Feld, Analytical Chemistry **70,** 623-627 (1998).

[7] W.-C. Shih, K. L. Bechtel, and M. S. Feld, Analytical Chemistry **79,** 234-239 (2007).

[8] R. G. Brereton, Analyst **125,** 2125–2154 (2000).

[9] M. A. Arnold, J. J. Burmeister, and G. W. Small, Analytical Chemistry **70,** 1773-1781 (1998).

[10] P. D. Wentzell, D. T. Andrews, and B. R. Kowalski, Analytical Chemistry **69,** 13 (1997).

[11] A. Savitzky and M. J. E. Golay, Analytical Chemistry **36,** 13 (1964).