Postprint of: Barman I., Dingari C., Singh G., Soares J., Dasari R., Smulko J., Investigation of Noise-Induced Instabilities in Quantitative Biological Spectroscopy and Its Implications for Noninvasive Glucose Monitoring, Analytical Chemistry, Vol. 84, Iss. 19 (2012), pp. 8149–8156, DOI: 10.1021/ac301200n

Investigation of noise-induced instabilities in quantitative biological spectroscopy and its implications for non-invasive glucose monitoring

Ishan Barman[1], Narahara Chari Dingari[1], Gajendra Pratap Singh[1,3], Jaqueline S. Soares[1], Ramachandra R. Dasari[1], Janusz M. Smulko[1,2,*]

[1]*Laser Biomedical Research Center, G. R. Harrison Spectroscopy Laboratory, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA*

[2]*Gdansk University of Technology, Faculty of Electronics, Telecommunications and Informatics, G. Narutowicza 11/12, 80-233 Gdansk, Poland*

[3]*Current Address: Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA*

* To whom correspondence should be addressed. E-mail: jsmulko@eti.pg.gda.pl

Abstract: Over the past decade, optical spectroscopy has been employed in combination with multivariate chemometric models to investigate a wide variety of diseases and pathological conditions, primarily due to its excellent chemical specificity and lack of sample preparation requirements. Despite promising results in several proof-of-concept studies, its translation to the clinical setting has often been hindered by inadequate accuracy of the conventional spectroscopic models. To address this issue and the possibility of curved (non-linear) effects in the relationship between the concentrations of the analyte of interest and the mixture spectra (due to fluctuations in sample and environmental conditions), support vector machine-based least squares non-linear regression (LS-SVR) has been recently proposed. In this paper, we investigate the robustness of this methodology to noise-induced instabilities and present an analytical formula for estimating modeling precision as a function of measurement noise and model parameters. This formalism can be readily used to evaluate uncertainty in information extracted from spectroscopic measurements, particularly important for rapid-acquisition biomedical applications. Subsequently, using field data (Raman spectra) acquired from glucose clamping study on an animal model subject, we perform the first systematic investigation of the relative effect of additive interference components (namely, noise in prediction spectra, calibration spectra and calibration concentrations) on the prediction error of non-linear spectroscopic models. Our results show that LS-SVR method gives more accurate results and is substantially more robust to additive noise when compared with conventional regression methods such as partial least-squares regression (PLS), when careful selection of the LS-SVR model parameters are performed. We anticipate that these results will be useful for uncertainty estimation in similar biomedical applications where the precision of measurements and its response to noise in the dataset is as important, if not more so, than the generic accuracy level.

1. Introduction

The wealth of information present in optical spectra acquired from biological tissues can be used to determine the concentration of important bio-markers and clinically relevant constituents for the purpose of real-time diagnosis. Even without the addition of exogenous imaging agents or dyes, one can probe the intrinsic information present in the form of absorbers, chromophores and scatterers (elastic and inelastic) to understand the properties of the tissue sample. Indeed, over the past decade, researchers in biomedical spectroscopy have tried to employ this basic underlying principle to diagnose diseases including cancer, atherosclerosis, malaria and diabetes (Bodanese et al., 2010; Haka et al., 2005; Lin et al., 2012; Peres et al., 2011; Kang et al., 2011; Dingari et al., 2012). A promising example of such application is spectroscopy-based non-invasive blood glucose monitoring, which has the potential of making substantive social and economic impact by alleviating the pain and inconvenience associated with conventional finger-stick measurements. Notably, blood glucose detection is critical to the management and therapeutics of nearly 366 million diabetics around the globe (as per the 2011 estimates of the International Diabetes Federation) - 26 million in the US alone [CDC, 2011]. Consequently, blood glucose sensors form nearly 85% of the world market for all biosensors, which amounts to more than $6 billion in annual sales (Vashist, 2012). To this end, several non-invasive detection methodologies have been proposed and implemented with varying degrees of success (Khalil et al., 2004; Roe et al., 1998). In particular, considerable contemporary attention has been focused on using infrared (NIR) Raman spectroscopy, which combines the substantial penetration depth of NIR light in biological tissue with the excellent chemical specificity of Raman spectroscopy (Chaiken et al., 2010).

Despite promising results in whole blood samples and physical tissue models, clinical translation of this promising technology (as well as that of other similar approaches *e.g.* NIR absorption spectroscopy) has been largely impeded by the variations in the blood-tissue matrix (including skin-layer thickness, hydration state and blood perfusion) (Barman et al., 2009; Dingari et al., 2011a). In order to overcome such fluctuations in sample and surrounding conditions, which may introduce non-linear effects in the spectra-concentration relationship, several investigators, including our own laboratory, have recently employed support vector machine regression (SVR), a relatively new class of multivariate methods that can handle ill-posed problems and lead to unique global models (Barman et al., 2010b; Barman et al., 2011b). While several studies have addressed the issue of improvement in prediction (or classification) accuracy arising from use of SVR in relation to conventional linear methods (e.g. principal component regression (PCR) and partial least squares (PLS)), the prediction uncertainty (or its response to noise in calibration data or/and prediction spectra) has not been systematically investigated.

Such an investigation is imperative for successful clinical implementation as the confidence in the measurement of a specific diagnostic parameter can alter the course of disease management, with ramifications to the health of the patient. In fact, the response (*i.e.* sensitivity) of the prediction algorithm to noise sources in the dataset could critically affect the diagnostic value of the data. In this article, we provide a significant extension to previous reports of SVR application for spectroscopic datasets by assessing the prediction response of such an algorithm to three sources of error: noise in the calibration spectra, noise in the prediction spectra and noise in the (reference) calibration concentration measurement. Using transcutaneous Raman spectra acquired from an animal model in conjunction with blood glucose concentrations, we first demonstrate that application of least-squares SVR (LS-SVR) enables a significant improvement in prediction accuracy over PLS. Importantly, by addition of white noise component

simulating potential noise incorporation under real experimental conditions, we observe that the LS-SVR method exhibits lower sensitivity of the analyte prediction error to noise of the aforementioned input quantities when compared with PLS, which makes the non-linear LS-SVR method particularly attractive.

In addition, we derive analytical expressions for the limiting uncertainty for LS-SVR models that are introduced in the concentration predictions due to presence of noise in the prediction spectra. Here, limiting uncertainty is defined as the uncertainty in concentration determination in the special case the calibration model is assumed to be completely accurate and noise free and the sole source of noise emanates from measurement noise in the prediction sample. This represents a particularly important case as the constraints on measuring tissue constituents under clinical conditions can cause noise in the prediction spectra to be substantially larger than the calibration noise. Our formalism provides a direct method to compare the LS-SVR limiting uncertainty with that of linear regression models estimated by previous investigators (originally by Lorber and Kowalski (Lorber and Kowalski 1988) and later by Feld and co-workers (Berger and Feld, 1997; Scepanovic et al., 2007)). We believe that this formalism can be readily used to evaluate uncertainty in information extracted from spectroscopic measurements, particularly for rapid-acquisition biomedical applications.

Taken together, we anticipate that our demonstration of LS-SVR robustness alongside its well-established accuracy and ability to construct a unique global model will firmly establish a foundation for its usage in applications where insensitivity to noise elements is critical. In particular, the incorporation of an accurate and noise-insensitive modeling technique is likely to provide significant impetus to the construction of a truly non-invasive glucose monitoring device. Similarly, the impact of noise due to interference from matrix elements in several other biosensing applications (Song et al., 2005) may be substantively reduced by employing LS-SVR as long as the model parameters are carefully optimized for the specific dataset (especially to avoid selection of singular points). Notably, although this work employs Raman spectroscopy as a specific example to demonstrate the robustness of LS-SVR, it should be stated that the approach used in our work is broad and general enough to be applicable to similar spectral datasets (such as NIR absorption) acquired in clinical diagnostics as well as in process monitoring applications in pharmaceutical and food industries.

2. Materials and methods

2.1. Experimental studies on animal model

For our study of robustness of algorithm performance in response to incorporation of noise in specific input elements, a representative dataset acquired from an animal model (beagle dog) during a glucose clamping study was used. The detailed experimental protocol and the Raman system used for these experiments were originally detailed in one of our laboratory's previous publications (Dingari et al., 2011a). For orientation, a brief description of the experimental study is provided in the Supplementary Information (Sec.-S1) (the interested reader is directed to the relevant references in the literature for a more comprehensive understanding (Shih et al., 2007; Enejder et al., 2002; Enejder et al., 2005).

2.2. Data analysis

In the aforementioned glucose clamping study, the total measured blood glucose concentrations and spectra were averaged in blocks of 5 resulting in a net of 666 distinct data points. In this paper, PLS and LS-SVR (Cristianini et al., 2000; Suykens et al., 2002) are used to study the performance of linear and

non-linear methods and their respective response to noise in the input datasets, as detailed in Sec.-2.3. Notably, we have employed LS-SVR as this algorithm provides a powerful tool for small sampling and high dimensional problems and also has greater generalization power with the ability to avoid over fitting (Lewis et al., 2006). Further details of the data analysis are given in the Supplementary Information (Sec.-S2).

2.3 Numerical simulations

As mentioned previously, the animal model study was undertaken in pursuance of systematic characterization of indigenous noise in non-invasive glucose measurements arising from reference glucose concentrations, Raman spectra used for regression model and Raman spectra used for prediction. To simulate this situation, white noise components of different intensities were separately added to the three independent measurements (calibration spectra, calibration (reference) glucose concentrations and prediction spectra). The simulations address the issue of how the additive corrupting white noise reduces accuracy of blood glucose predictions.

Intensity of the corrupting noise was selected by approximating the acquired tissue spectrum by polynomial curve fitting and establishing standard deviation σ_o of the difference between the spectra and its polynomial approximation averaged over a set of all recorded 666 spectra. A representative case, with the original spectrum and the noise-added spectrum of white noise component having variance equal to $0.1\sigma_o^2$ (spatially offset for sake of clarity), is shown in Fig. 1(a). In a case of glucose concentration, the additive noise component was selected as 5%, 10% or 15% of the averaged glucose level measured by the reference method (*ca.* 137 mg/dL).

Figure 1(a) appears here

All computations were repeated 500 times by adding random sequences of additive noise having assumed intensity and zero mean value. The number of simulations and the number of spectra points were high enough to guarantee that the estimated root-mean-square error of cross-validation (RMSECV) did not change more than a percent when the same simulations were repeated.

3. Formulation of limiting uncertainty for non-linear LS-SVR models

For the determination of prediction uncertainty in the non-linear LS-SVR models, we use a framework similar to that used to describe the error propagation for conventional linear multivariate prediction algorithms (such as PCR and PLS) (Lorber et al., 1998; Faber, 2000). Specifically, we focus on the important case of limiting uncertainty where noise in the prediction dataset (spectra) is the dominant source of error (Berger and Feld, 1997; Scepanovic et al., 2007). This is especially relevant in biomedical and clinical applications where the noise in the prediction spectra is substantially higher than that in the calibration set and the errors in the reference concentrations, due to the inherent acquisition time constraints in prospective (field) samples. While the limiting uncertainty has been extensively addressed for linear chemometric methods, it has received scant attention for non-linear methods. Here, for the first time, we establish an analytical formula for the precision (or reproducibility) of the LS-SVR method in response to noise in the prediction spectra and outline the presence of instability (or lack thereof) in the output LS-SVR predictions.

For the linear models, uncertainty of glucose concentration can be determined by using the standard additive noise model [Faber and Kowalski, 1997; Scepanovic et al., 2007]:

$$s = Pc + w. \tag{1}$$

The vector s is the measured spectrum vector of size equal to number M of the recorded spectra wavelengths, the matrix P contains the model constituent vectors ($M \times N$), the vector c contains the coefficients of the model constituent ($N \times 1$) of N components and the vector w represents noise component of the recorded spectra. Based on Eq. (1), the model's limiting uncertainty in analyte extraction (Δc) can be expressed as (Lorber and Kowalski, 1988; Berger and Feld, 1997):

$$\Delta c = \sqrt{\sigma_n^2 \cdot B^2} \tag{2}$$

where the elements of the vector σ_n are the standard deviations, at each wavelength, of the Raman intensity measurements and correspond to the RMS spectral noise amplitudes in the spectroscopic signals; B is the so-called regression spectrum and the contravariant vector to the spectrum of the predicted analyte (glucose). Calibration methods arrive at the approximation to the true regression spectrum via different methods and accordingly the above formalism can be specialized for a particular linear regression scheme such as PLS and ordinary least squares (OLS). Regardless of the determination method of regression spectrum, Eq. (2) clearly shows the linear dependence between the spectroscopic noise (in the prediction sample) and the uncertainty in the predicted concentration.

Here, we consider how the noise inherent in the measurement of the Raman spectrum (Δs) relates to the uncertainty of glucose concentration prediction (Δc) when LS-SVR method is used to connect the concentration and spectra blocks. Similar to the above case, the developed calibration model itself is assumed to be accurate, i.e. devoid of noise and physiological lag-related errors [Barman et al., 2010a]. Also, in the following we assume that the Gaussian RBF (Eq. (S.2)) is used for the regression modeling, due to its extensive usage for such applications. From Eq. (S.1) and (S.2) in the Supplementary Information (Sec.-S2), we obtain:

$$c + \Delta c = \sum_{i=1}^{M} \alpha_i \exp\left(-\frac{\left(s_i - s - \Delta s\right)^2}{\sigma^2} \right) + b \tag{3}$$

We can re-arrange the above equation to arrive at a simplified form. Alternately, using a partial derivative to relate the two terms of interest (Δc and Δs) (and neglecting higher order terms) from Eq. (S.1), we can derive the following relation:

$$\Delta c \approx \frac{\delta c}{\delta s} \Delta s = \frac{\Delta s}{\sigma^2} \sum_{i=1}^{M} \alpha_i \left(s_i - s\right) \exp\left(-\frac{\left(s_i - s\right)^2}{2\sigma^2} \right) \tag{4}$$

Also, the Lagrange multipliers α_i depends on the regularization parameter γ and can be written as (Thissen et al., 2004):

$$\alpha_i = \left(s_i^T s_i + \frac{1}{2\gamma} \right)^{-1} \tag{5}$$

Combining Eq. (4) and (5), we obtain the relation for the uncertainty in concentration prediction:

$$\Delta c = \frac{\Delta s}{\sigma^2}\sum_{i=1}^{M}\frac{(c_i - b)}{\left(s_i^T s_i + \frac{1}{2\gamma}\right)}(s_i - s)\exp\left(-\frac{(s_i - s)^2}{2\sigma^2}\right) \quad (6)$$

Eq. (6) predicts a complex and nonlinear dependence of Δc on the kernel parameter σ^2 but a more easily interpretable relation with the regularization parameter, γ. In general, a large value of γ indicates that a relatively large emphasis is assigned to obtaining low prediction errors while retaining possibly high weight coefficients (thereby running the risk of overfitting). In this context, Eq. (6) once again highlights the necessity for lower value of γ. On the other hand, the effect of selection of σ^2 on Δc is more difficult to predict in a general sense (*i.e.* without the context of a particular spectroscopic dataset) due to the two competing terms: the exponential term of the RBF kernel (which is computed separately for each of the calibration spectra and summed over those points) and the $1/\sigma^2$ term. The former term exhibits lower values at smaller values of σ and rises with increasing σ to a maximum value of 1. In direct contrast, $1/\sigma^2$ term has maximum value at $\sigma = 0$ and decreases to 0 as $\sigma \rightarrow \infty$. Thus, the presence and position of local maxima depends on the specific dataset.

The above formulation provides a general insight into the prediction uncertainty obtained using a support-vector machine based non-linear regression approach. Expectedly, one clear point of contrast with Eq. (2) is the presence of non-linear terms that are a function of the calibration and prediction data. It is also instructive to note that while the selection of the model parameters is not based on the above equation, it signifies the necessity for having a high degree of model generalizability (low value of γ). Using this formulation, we can determine the uncertainty (which is intrinsically linked to limit of detection) using a single spectrum rather than having to evaluate the uncertainty by repeating the measurement many times and analyzing the standard deviation of parameters extracted from each of these multiple measurements. We anticipate that this method can be directly used to guide improvements in data modeling, as well as in further selection and optimization of the spectroscopic hardware.

4. Results and Discussion

4.1 Calibration model performance in response to noise addition to prediction spectra

The Raman spectra collected during the experiment had high SNR values due to relatively long acquisition time (Fig. 1(a)), to ensure accurate calibration model preparation for both types of regression methods: linear PLS and non-linear RBF-kernel based LS-SVR. Figure 1(b) shows representative profiles of the reference glucose concentrations and the PLS leave-one-out cross-validation-based predictions using 7 loading vectors when no noise is added to any of the spectra. It is worth mentioning that since the spectra are acquired in an animal model over a finite time frame, they have an intrinsic noise component stemming largely from the shot noise in the measurements – which in turn gives rise to the error in prediction even when no additional noise is incorporated. Thus, we observe that the root mean squared error of cross-validation (RMSEV) is observed to be 27.9 mg/dL.

Figure 1(b) appears here

To simulate the real world situation where substantially larger noise levels may be present in the acquired spectra due to limited acquisition times as well as other measurement errors in spectral and

reference concentration determination, corrupting noise component was added to investigate how such interferences influence the overall performance of glucose prediction models. For the first case, the noise was added to only the prediction sample (analogous to the situation mentioned in Sec.-3). Based on leave-one-out cross-validation protocol where the sample left out of the calibration model was mixed with white noise, we were able to determine the prediction errors for PLS and LS-SVR for different levels of additive noise (Fig. 2). First, it is evident that LS-SVR models give significantly better accuracy as indicated by the nearly 2 times lower RMSEV of glucose predictions for LS-SVR in relation to PLS. Here, we consider additive noise levels over two orders of magnitude from the practical case of glucose Raman signal being higher than the noise floor to the case where the noise floor significantly exceeds the glucose Raman signal (thereby rendering impossible any efforts towards realistic prediction). Specifically, for both PLS (Fig. 2(a)) and LS-SVR (Fig. 2(b)), we observe that at the intermediate-higher end of additive noise levels, the models exhibit little or no predictive capability as the RMSEV appears to reach or even cross 100 mg/dL. Importantly, however, in the more interesting region of the additive noise (lower-intermediate range) the PLS method shows greater sensitivity to additive noise in prediction spectra when compared with LS-SVMR as the RMSEV begins to increase at a much lower noise level (*i.e.* has a smaller value of the crossing point σ_n/σ_0 where the RMSEV spikes up rapidly above the no-noise case). This underlines the more robust nature of LS-SVR models to the added noise component in the prediction spectra.

Figure 2 appears here

Further, we observe that PLS shows approximately linear dependence versus intensity of noise corrupting spectra (Fig. 2a) as suggested by Eq. (2). A similar form is shown in Fig. 2(b) as well (and would be expected based on Eq. (6)). The deviations from the linear dependence in both cases can be attributed to the fact that the calibration models are evidently not noise-free and accurate which violates the assumptions used to build the theoretical framework of Eq. (2) and (6). Additionally, there appears to be a non-linear component at higher additive noise levels for LS-SVR models in Fig. 2(b) which we suspect appears from the importance of the higher-order terms that were neglected in developing the linearized version of Eq. (6). Interestingly, the observed dependence of the RMSEV on the noise in the prediction spectra does not change in form when another set of the parameters γ and σ^2 is selected, although a change in intensity is noted. This manifests itself in a difference in position of the crossing point where the RMSEV shows a substantive rise in relation to the no-noise case.

4.2. Calibration model performance in response to noise addition to calibration and prediction spectra

In contrast to the aforementioned case, in many spectroscopic (Raman or otherwise) applications, the dominant source of noise may be independent of the signal intensity (homoscedastic) and may comprise of detector noise, contributions from the background and laser intensity fluctuations. This is especially true when high signal levels can be achieved and exposure times can be increased without any significant downside as stated elsewhere in the literature (Bell et al., 1998; Barman et al., 2011a; O'Grady et al., 2001). In this section, we investigate the case where both calibration and prediction spectra are affected by noise.

Figure 3 appears here

Figure 3 shows the plots of the RMSEV as a function of the noise in prediction spectra and the noise in the calibration spectra. Specifically, the Y-axis corresponds to the RMSEV values and the X-axis corresponds to the mean level of noise added to the prediction spectra. The three curves represent the obtained values for three distinct levels of noise added in the calibration spectra (black squares, red circles and green diamonds denote σ_n/σ_0 = 5x10^{-4}, 4x10^{-3} and 8x10^{-3}, respectively). The overall results further validate the accuracy of LS-SVR models in relation to comparable PLS ones as well as highlight the robustness of the former to added noise in the spectral dataset. Expectedly, Fig. 3(b) shows that at lower levels of noise in prediction spectra, the noise in calibration spectra has a substantive (adverse) impact on determining the RMSEV value – but as prediction noise increases the two effects can be deemed comparable.

Surprisingly, though, in Fig. 3(a) we find the situation to be reverse, i.e. the higher noise levels in the calibration dataset appears to be consistent with a superior predictive performance. This is rather unexpected and we are not completely certain of its underlying mechanism. We attribute this surprising finding to the fact that when noise is increased at the calibration stage it causes the automated algorithm to limit the number of loading vectors. In our case the number of loading vectors was 6 for the presented data when more intense corrupting noise was added to the calibration spectra (green diamonds in Fig. 3(a)). In contrast, the number of loading vectors for the data without additional noise (Fig. 2(a)) or with moderate noise in calibration spectra (black squares and red circles in Fig. 3(a)) was observed to equal to 7. However, having a larger number of loading vectors actually may induce spurious correlations (with noise elements or in a temporal sense with system drift) thereby adversely affecting the algorithm performance when a smaller amount of noise is present in the calibration data. Nevertheless, we note that this may be an issue specific to our spectroscopic dataset rather than a generic phenomenon. Further investigations in this direction are currently ongoing in our laboratory across a wide variety of datasets measured across a number of spectroscopic units and will be reported at a later date.

4.3 Calibration model performance in response to noise addition in reference glucose concentration and prediction spectra

Finally, influence of accuracy of the reference glucose concentrations was also considered by adding white noise component to the reference glucose data measured by the Analox analyzer. As noted in the literature (Brereton, 2007), often the more significant source of error in inverse spectroscopic calibration (where the intention is to predict concentration from spectra and not vice versa) is the sample preparation such as extraction and concentration measurement, rather than spectroscopic instrument reproducibility. In other words, the measurement of a concentration (even when determining a calibration model) is likely to be less certain than measurement of spectral intensity as the pace of improvement of spectroscopic systems has not been matched by a concomitant improvement in sample extraction or in the quality of measurement cylinders, pipettes etc. Thus, it is imperative to consider noise in reference concentration measurements. For example, the nationally accepted standard error of glucometers is ±20%, which implies a calibration model based on finger-prick measurements (rather than clinical laboratory determination) are likely to face up to 20% variations in their reference values (Mann et al., 2007).

Based on variations in level of noise added to the concentration dataset, we created distinct calibration models and used a leave-one-out cross-validation protocol to estimate the dependency of RMSEV. Figure 4 plots the RMSEV values as a function of the additive noise in the prediction spectra as well as

the noise in the reference concentration for PLS (Fig. 4(a)) and LS-SVR (Fig. 4(b)) models. In particular, four sets of RMSEV values corresponding to 0%, 5%, 10% and 15% noise in concentration levels are shown. (The noise levels added in the concentration dataset are consistent with that typically observed in point-of-care glucose measurement devices.)

Figure 4 appears here

As we could expect the corrupting noise decreases accuracy of glucose concentration prediction but the observed deterioration of accuracy is relatively low when the interference noise does not exceed a few percent (Fig. 4). Much more significant changes are evident when the additive noise intensity reaches 10%. Interestingly, the relative change of RMSEV between the successive curves (corresponding to different levels of noise in the reference concentrations) is lower for PLS models (Fig. 4(a)) than for LS-SVR (Fig. 4(b)) models at the lowest level of noise added to the prediction spectra. At higher levels of noise in prediction spectra, the differences due to noise levels in reference concentration are largely masked and the curves seem to coalesce both for PLS and LS-SVR models. Nevertheless, for all values of noise in reference concentration and prediction spectra, the LS-SVR retains substantially higher accuracy than the optimal PLS models.

In summary, we have proposed, for the first time, an analytical method for quantitative determination of prediction uncertainty in non-linear support vector regression models. This formulation presents a powerful methodology for transformation of known noise or error levels in input datasets to a precision value of the output property of interest. It has substantive implications for translation of a laboratory-based spectroscopic methodology to the clinic as diagnostic confidence in the measurement of an analyte can radically impact the potential avenues for disease management. Moreover, in addition to investigation the limiting uncertainty, we have performed numerical computations on a clinical dataset collected from an animal model to individually investigate the specific effects of noise/error. Based on our observations, it is evident that the LS-SVR technique not only provides a high degree of accuracy in concentration prediction especially for datasets where considerable non-linearity can be expected (here, due to the effects of turbidity (Barman et al., 2010b)) but also provides a reasonable amount of robustness to noise sources. Specifically, it is found that additive noise components in the spectral data have less of an adverse impact on the LS-SVR models than the PLS models.

Nevertheless, it is worth noting that the presented results of LS-SVR required the establishment of RBF parameters, which can be a time-consuming process. Careful selection of parameters is particularly necessary because in addition to obtaining a low prediction error one must search for a smooth sub area (i.e. the set of γ and σ^2 should not be a singular point) otherwise the model is liable to be rigid. Optimal and fast selection of these parameters will be one of the focal points of our future work. In addition, for our future investigations, we plan to incorporate noise reduction schemes (such as Minimum Noise Fraction Transform (Reddy and Bhargava, 2010), where data are forward transformed, components that correspond mostly to signal are selected and used in inverse transformation) to improve the lower signal-to-noise ratios in the spectroscopic data acquired from clinical samples. We anticipate that the ultimate result will be a customized calibration tool that is based on a hybrid of different chemometric methods. Finally, while we have demonstrated the potential accuracy and robustness of LS-SVR models with respect to glucose monitoring using Raman spectroscopy, the underlying concepts here can be easily extended to analogous spectroscopic calibration and classification problems frequently encountered in

clinical diagnostics (Sattlecker et al., 2010, Saha et al., 2011), bio-reactor monitoring (Lee et al., 2004) and quality control in the food and drug industry (Myakalwar et al., 2011).

5. Concluding Remarks

In this article, we have investigated the noise-induced perturbations in prediction performance of chemometric models typically used for tissue spectroscopy. Here, we have described a relatively straightforward and direct method of computing prediction (limiting) uncertainty of spectroscopic measurements for non-linear support vector regression models, which relates to the limit of detection (approximately 3x the limiting uncertainty). Our investigations also reveal the relative impact of specific noise sources (namely in the spectral and concentration datasets) on the linear and non-linear regression models. In particular, we have determined that the non-linear support vector regression models not only provide higher prediction accuracy (potentially due to the inherent non-linearity in the concentration-spectra relationship) but also greater robustness against noise incorporate in the spectral data. Nevertheless, the parameters of the non-linear regression model must be selected carefully, especially the kernel parameter (σ^2 for radial basis function kernel), which largely determines its resistance to the noise components. In the future, the robustness of the algorithm(s) can also be further enhanced by the extraction of the most informative regression features (Dingari et al., 2011b). One evident avenue for such enhancement is the employment of suitable feature (*i.e.* wavelength) selection methodologies, *e.g.* genetic algorithms and simulated annealing.

Importantly, we are currently investigating the possibility of employing noise (and noise-based fluctuations in concentration prediction or diagnostic classification) as a property of interest to characterize the system under investigation. Drawing inspiration from varied examples in nature and in the scientific literature (for example, the computation of Brownian fluctuations of an optically trapped particle to measure torque (Volpe and Petrov, 2006)), our studies will focus on elucidating latent correlation between analytes of interest (glucose for diabetes monitoring or protein concentration in cellular mapping) and the fluctuations in the spectroscopic signals.

ACKNOWLEDGEMENTS

The authors wish to thank the NIH National Center for Research Resources for their grant P41-RR02594, at the MIT Laser Biomedical Research Center. One of the authors (J.S.S.) would like to acknowledge the support of CNPq fellowship.

References

Barman, I., Singh, G.P., Dasari, R.R., Feld, M.S., 2009. Anal. Chem. 81, 4233–4240.

Barman, I., Kong, C.R., Singh, G.P., Dasari, R.R., Feld, M.S., 2010a. Anal. Chem. 82, 6104–6114.

Barman, I., Kong, C., Dingari, N.C., Dasari, R.R., Feld, M.S., 2010b. Anal. Chem. 82, 9719–9726.

Barman, I., Kong, C.R., Singh, G.P., Dasari, R.R., 2011a. J. Biomed. Opt. 16, 011004.

Barman, I., Dingari, N.C., Rajaram, N., Tunnell, J.W., Dasari R.R., 2011b. Biomed. Opt. Express. 2, 592–599.

Bell, S.E.J., Bourguignon, E., Dennis, A., 1998. Analyst 123, 1729–1734.

Berger, A.J., Feld, M.S., 1997. Appl. Spectrosc. 51, 725−732.

Bodanese, B., Silveira, L. Jr., Albertini, R., Zangaro, R.A., Pacheco, M.T., 2010. Photomed. Laser Surg. 1, S119-27.

Brereton, R.G., 2007. "Applied Chemometrics for Scientists", John Wiley and Sons Ltd., Chichester, West Sussex, England.

Centers for Disease Control and Prevention. National diabetes fact sheet: national estimates and general information on diabetes and prediabetes in the United States, 2011. Atlanta, GA: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, 2011.

Chaiken, J., Deng, B., Bussjager, R.J., Shaheen, G., Rice, D., Stehlik, D., Fayos, J., 2010. Rev. Sci. Instrum. 81, 034301.

Cristianini, N., Shawe-Taylor, J., 2000. "An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods", Cambridge University Press, New York.

Dingari, N.C., Barman, I., Singh, G.P., Kang, J.W., Dasari, R.R., Feld, M.S., 2011a. Anal. Bioanal. Chem. 400, 2871–2880.

Dingari, N.C., Barman, I., Kang, J.W., Kong, C.R., Dasari, R.R., Feld, M.S., 2011b. J. Biomed. Opt. 16, 087009.

Dingari, N.C., Horowitz, G.L., Kang, J.W., Dasari, R.R., Barman, I., 2012. PLoS ONE 7, e32406.

Enejder, A.M.K., Koo, T.W., Oh, J., Hunter, M., Sasic, S., Feld, M.S., Horowitz, G.L. 2002 Opt Lett 27, 2004–2006.

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

Faber, K., Kowalski, B.R., 1997. J. Chemometrics, 11, 181.

Faber, N.M., 2000. Anal. Chem. 72, 4675-4676.

Haka, A.S., Shafer-Peltier, K.E., Fitzmaurice, M., Crowe, J., Dasari, R.R., Feld, M.S., 2005. Proc. Natl. Acad. Sci. U. S. A. 102, 12371–12376.

Kang, J.W., Lue, N., Kong, C.R., Barman, I., Dingari, N.C., Goldfless, S.J., Niles, J.C., Dasari, R.R., Feld, M.S., 2011. Biomed. Opt. Exp. 2, 2484–2492.

Khalil, O.S., 2004. Diabetes Technol. Ther. 6, 660–697.

Lee, H.L.T., Boccazzi, P., Gorret, N., Ram, R.J., Sinskey, A.J., 2004. Vibrational Spectroscopy 35, 131–137.

Lewis, D.P., Jebara, T., Noble, W.S., 2006. Bioinformatics 22, 2753–2760.

Lin, K., Lau, D., Cheng, P., Huang, Z., 2012. Biosensors and Bioelectronics, ISSN 0956-5663, 10.1016/j.bios.2012.02.050.

Lorber, A., Kowalski, B., 1988. J. Chemom. 2, 93–109.

Mann, E.A., Pidcoke, H.F., Salinas. J., Wade, C.E., Holcomb. J.B., Wolf. S.E., 2007. Am. J. Crit. Care. 16, 531–532.

Myakalwar, A.K., Sreedhar, S., Barman, I., Dingari, N.C., Venugopal Rao, S., Prem Kiran, P., Tewari, S. P., Manoj Kumar, G., 2011. Talanta 87, 53–59.

Noble, W.S., 2004. "Support vector machine applications in computational biology," In: Schoelkopf B, Tsuda K, Vert JP, ed. Kernel Methods in Computational Biology. MIT Press, 71–92.

O'Grady, A., Dennis, A.C., Denvir, D., McGarvey, J.J., Bell, S.E., 2001. Anal. Chem. 73, 2058–2065.

Peres, M.B., Silveira, L. Jr., Zangaro, R.A., Pacheco, M.T., Pasqualucci, C.A., 2011. Lasers Med. Sci. 26, 645–655.

Reddy, R.K., Bhargava, R., 2010. Analyst 135, 2818–2825.

Roe, J.N., Smoller, B.R., 1998. Crit. Rev. Ther. Drug. 15, 199–241.

Saha, A., Barman, I., Dingari, N.C., McGee, S., Volynskaya, Z., Galindo, L.H., Liu, W., Plecha, D., Klein, N., Dasari, R.R., Fitzmaurice, M., 2011. Biomed. Opt. Exp. 2, 2792–2803.

Sattlecker, M., Bessant, C., Smith, J., Stone, N., 2010. Analyst 135, 895–901.

Scepanovic, O.R., Bechtel, K.L., Haka, A.S., Shih, W.C., Koo, T.W., Berger, A.J., Feld, M.S., 2007. J. Biomed. Opt. 12, 064012-1.

Shih, W.C., 2007. "Quantitative biological Raman spectroscopy for non-invasive blood analysis", Massachusetts Institute of Technology, Dept. of Mechanical Engineering.

Song, J.M., Culha, M., Kasili, P.M., Griffin, G.D., Vo-Dinh, T., 2005. Biosens. Bioelectron. 20, 2203–2209.

Suykens, J.A.K., Van Gestel, T., De Brabanter, D., De Moor, B., Vandewalle, J., 2002. "Least Squares Support Vector Machines", World Scientific, Singapore.

Thissen, U., Ustun, B., Melssen, W. J., Buydens, L.M.C., 2004. Anal. Chem. 76, 3099–3105.

Vashist, S.K., 2012. Anal. Chim. Acta. ISSN 0003-2670, 10.1016/j.aca.2012.03.043.

Volpe, G., Petrov, D., 2006. Phys. Rev. Lett. 97, 210603.

Figure Captions

Fig. 1. Blood glucose measurements using Raman spectroscopy: (a) acquired spectrum and spectrum with additive white noise having variance $0.1\sigma_o^2$ from a representative time point during the multi-level glucose clamping study, (b) reference and PLS predictions using seven loading vectors, when no additive noise component was included in the spectra and concentration datasets.

Fig. 2. Plot of root mean squared error of validation (RMSEV) (mg/dL) as a function of additive noise in prediction spectra: (a) PLS application (the crossing point is determined to be for RMSEV = 29 mg/dL at $\sigma_n/\sigma_0 = 0.0085$), (b) LS-SVR application with RBF kernel having parameters $\gamma = 160$, $\sigma^2 = 80$ (the crossing point is determined to be for RMSEV = 17.5 mg/dL at $\sigma_n/\sigma_0 = 0.031$). Here crossing point is defined as the intersection between the initial horizontal line (denoting the RMSEV at $\sigma_n/\sigma_0 = 0.001$) and the best fit line for RMSEV at higher values of σ_n/σ_0.

Fig. 3. Plot of root mean squared error of validation (RMSEV) (mg/dL) as a function of additive noise in calibration and prediction spectra for: (a) PLS application, (b) LS-SVR application with RBF kernel ($\gamma = 160$, $\sigma^2 = 80$). The X-axis represents normalized standard deviation σ_n/σ_0 of white noise added to prediction spectra. Three sets of data points with black squares, red circles and green diamonds are represented corresponding to noise in calibration spectra $\sigma_n/\sigma_0 = 5\text{x}10^{-4}$, $4\text{x}10^{-3}$ and $8\text{x}10^{-3}$, respectively.

Fig. 4. Plot of root mean squared error of validation (RMSEV) (mg/dL) as a function of noise in reference concentrations and additive noise in prediction spectra for: (a) PLS application, (b) LS-SVR application with RBF kernel ($\gamma = 160$, $\sigma^2 = 80$). The X-axis represents normalized standard deviation σ_n/σ_0 of white noise added to prediction spectra. Four sets of data points with black squares, red circles, green diamonds and blue triangles are represented corresponding to noise in reference concentrations equal to 0%, 5%, 10% and 15%, respectively.

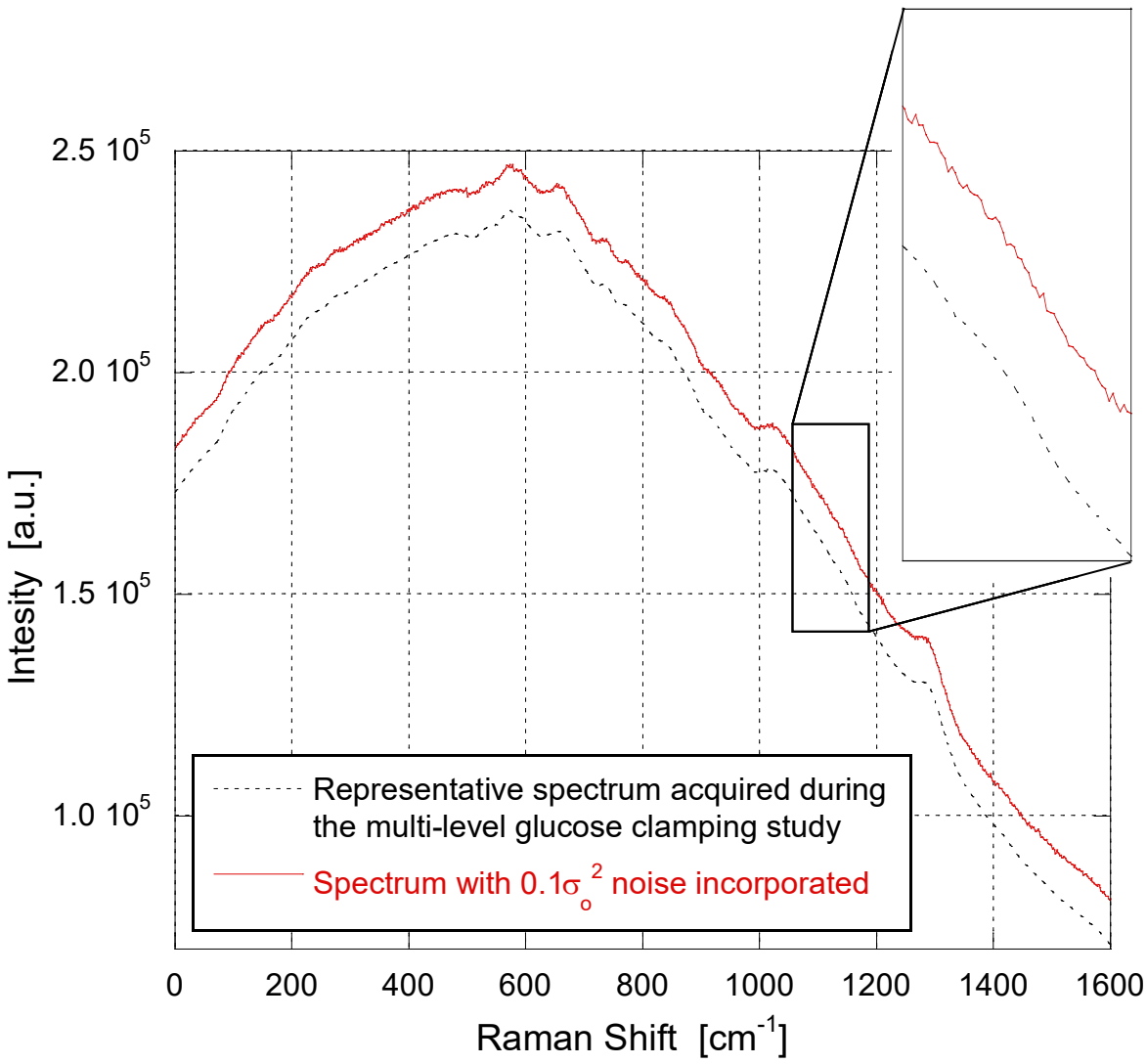

Figure 1 (a)

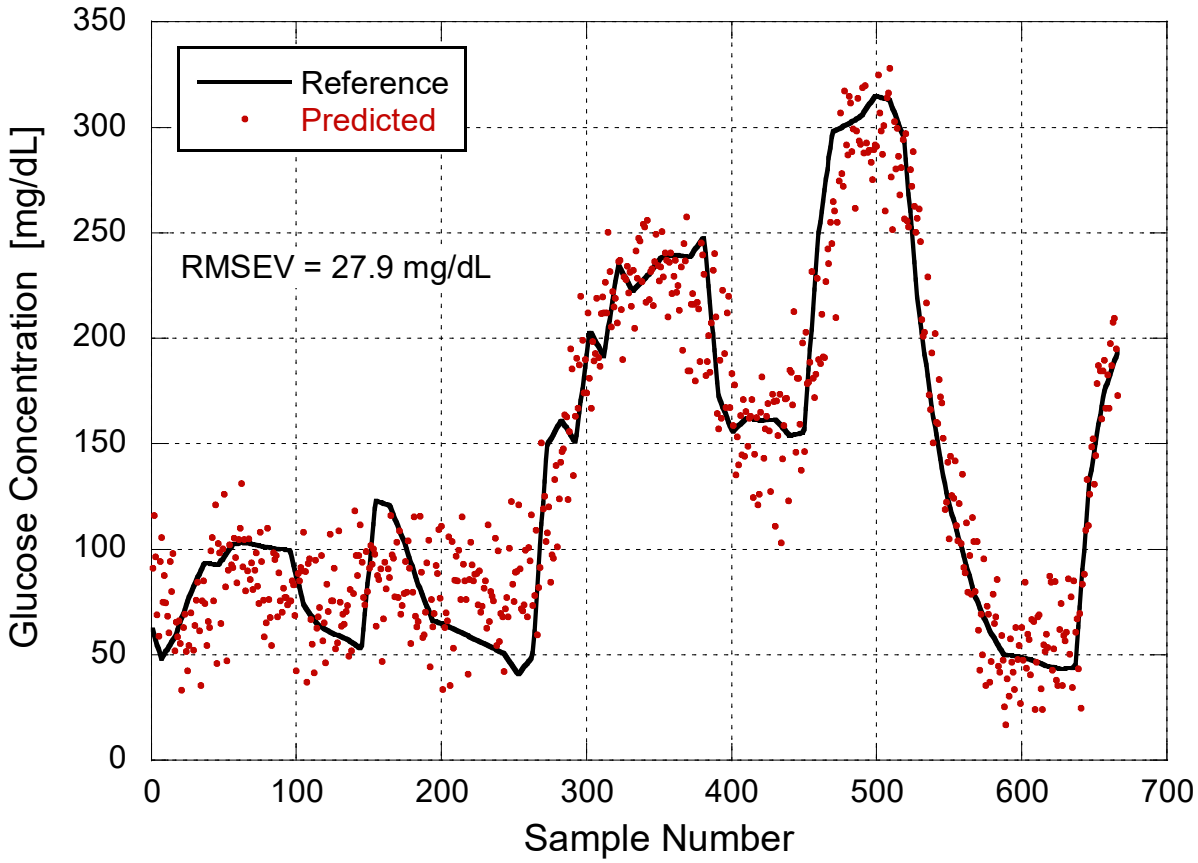

Figure 1 (b)

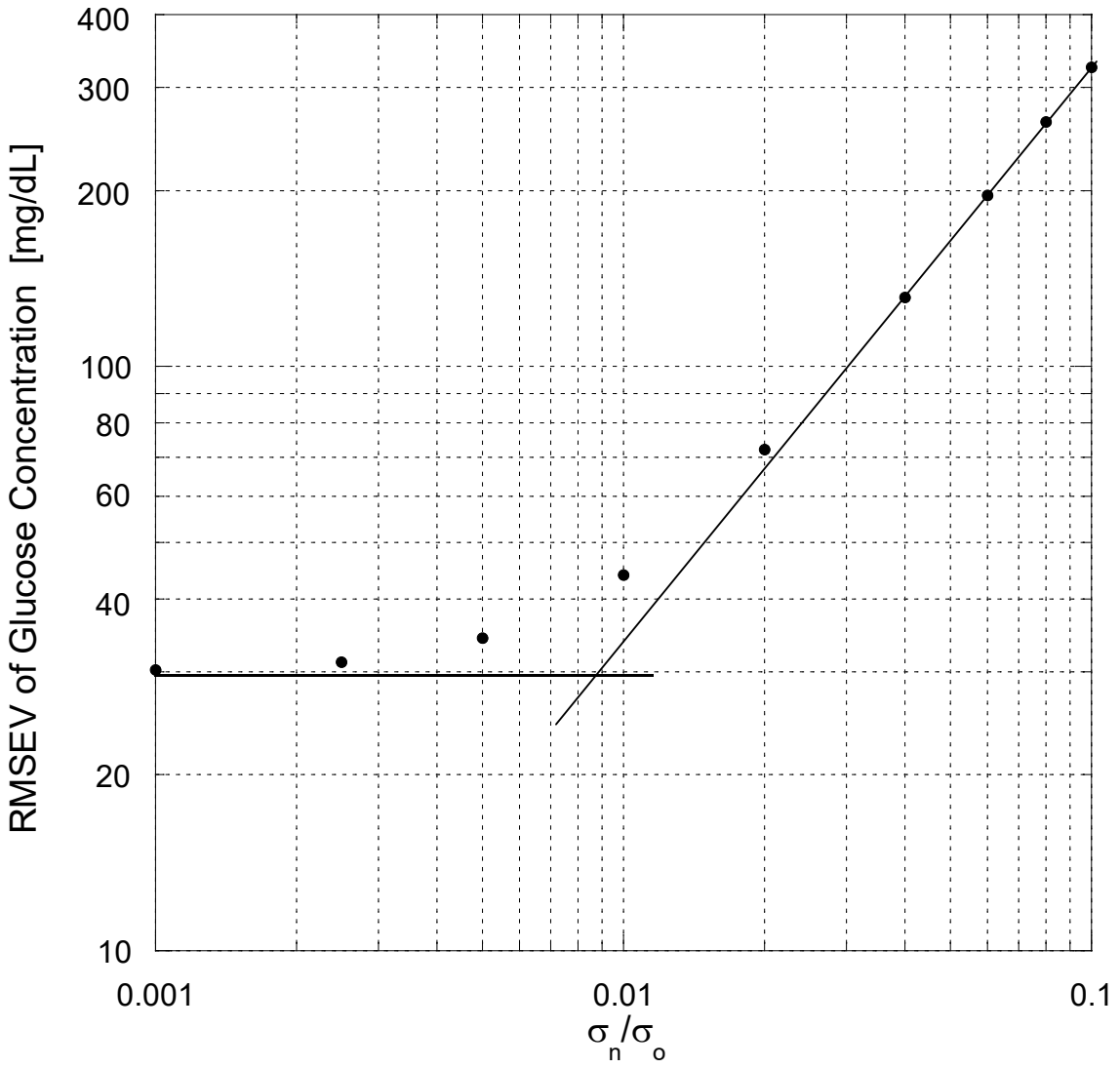

Figure 2 (a)

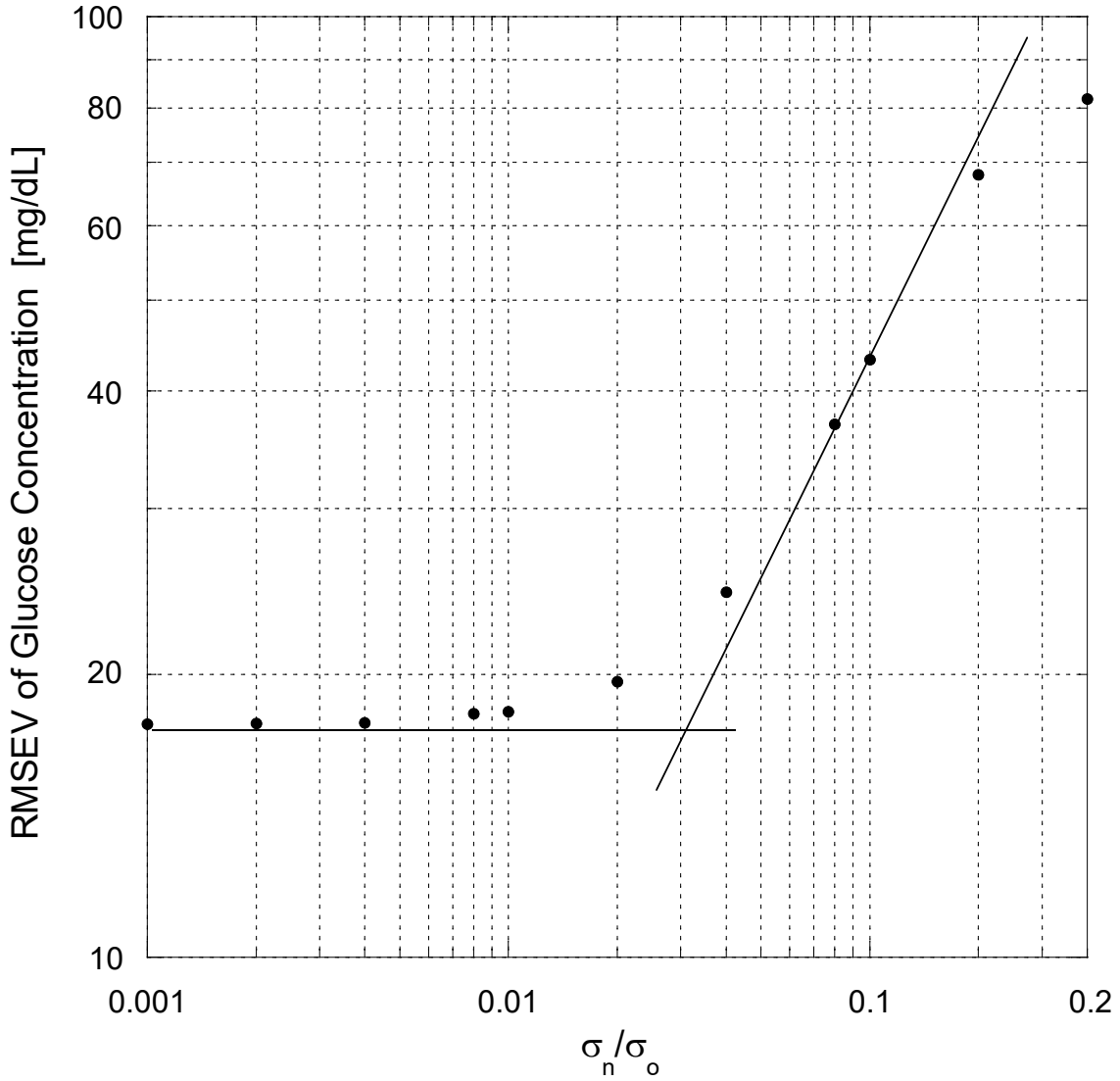

Figure 2 (b)

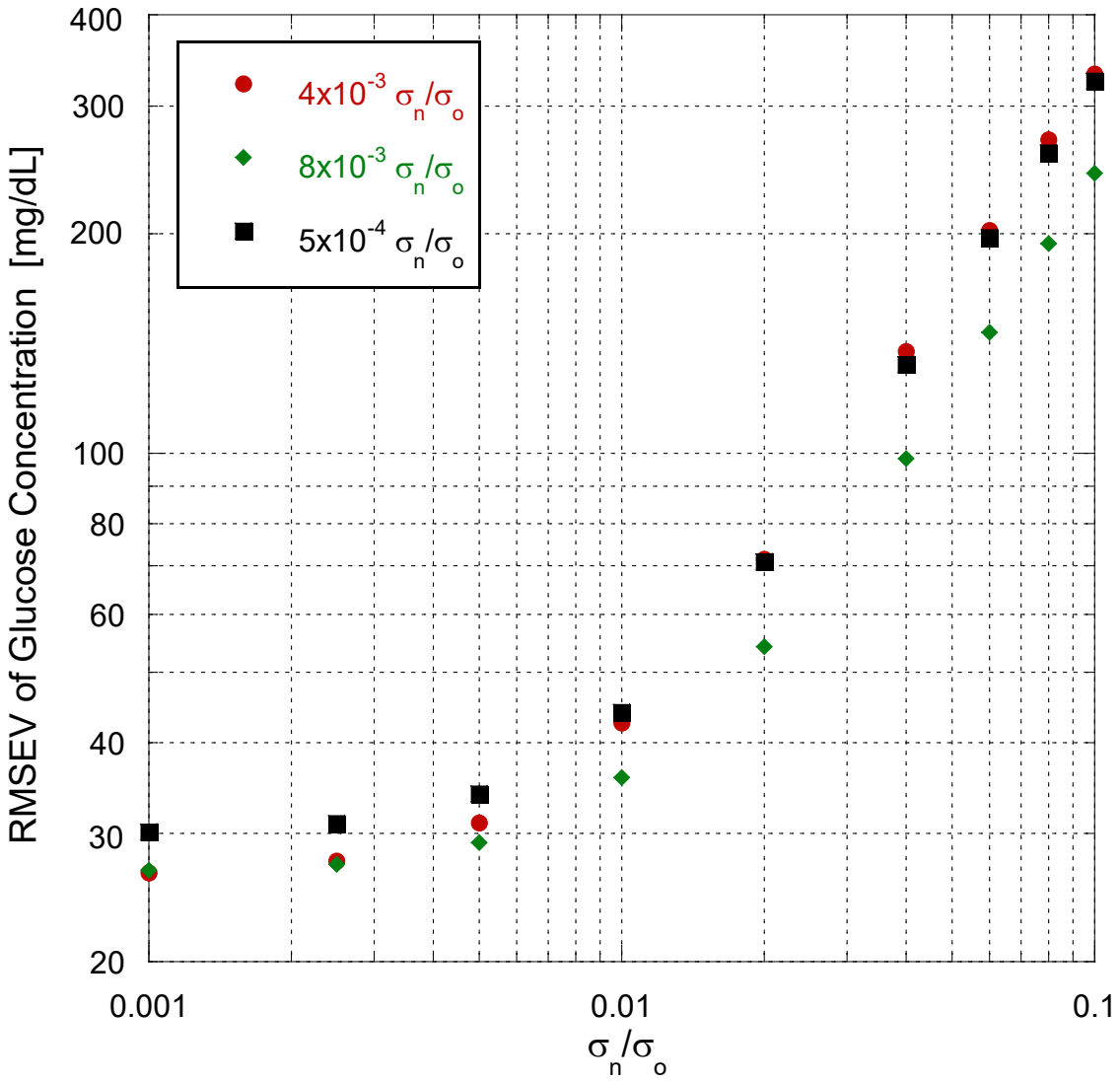

Figure 3 (a)

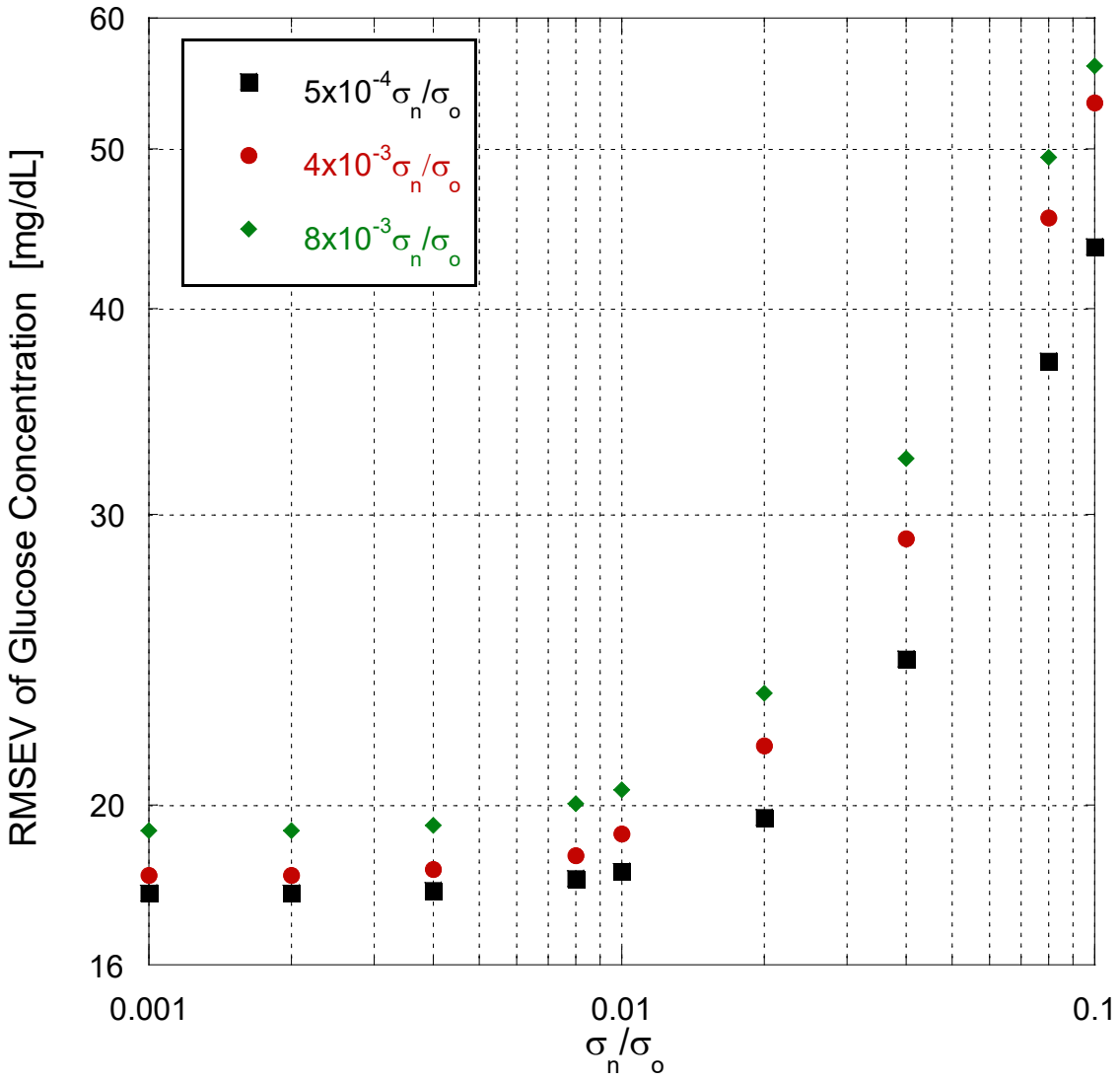

Figure 3 (b)

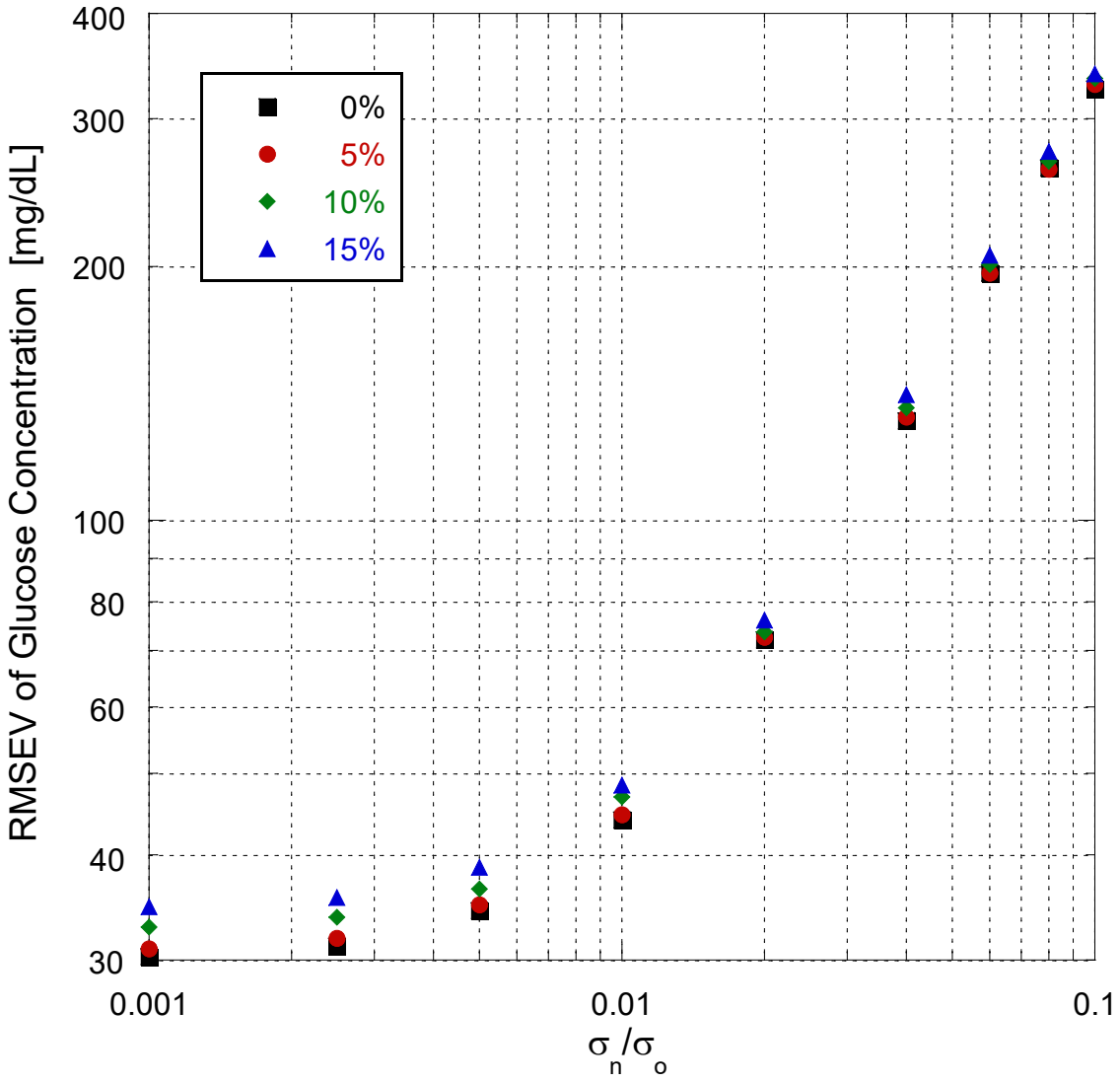

Figure 4 (a)

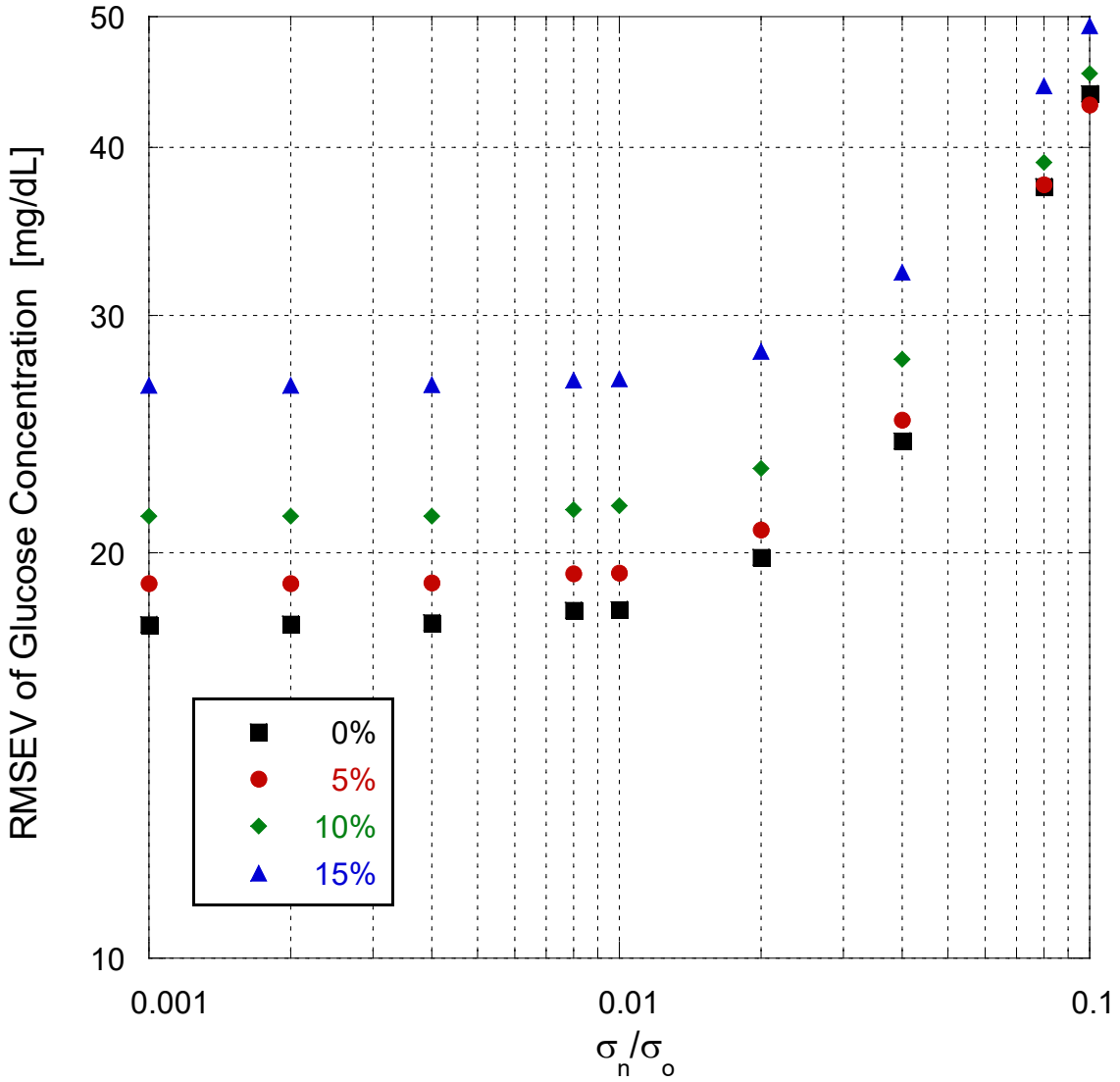

Figure 4 (b)