Predictive Glucose Monitoring for Type 1 Diabetes Using Latent Variable-based Multivariate Statistical Analysis

Chunhui Zhao¹, Eyal Dassau¹, Rebecca A. Harvey¹, Dale E. Seborg¹, Francis J. Doyle III¹

Department of Chemical Engineering, University of California, Santa Barbara, CA 93106-5080 USA (Tel: 805-893-8133; e-mail: doyle@engineering.ucsb.edu).

Abstract: Accurate prediction of future glucose concentration for type 1 diabetes mellitus is needed to improve glycemic control, which can produce early and proactive glycemia management before glucose concentrations drift to undesirable levels. This paper assesses the feasibility of data-driven latent variable (LV) based statistical analysis methods to characterize the glycemic variability and serve as the forecasting engine. It illustrates the capability of LV-based multivariate statistical analysis to (1) model the correlation relationship among glucose and exogenous inputs, (2) make reliable predictions of future glucose concentrations, and (3) capture the underlying dynamics of prediction errors for on-line reliability evaluation. The new approach provides an automatic predictive monitoring strategy for glucose management in type 1 diabetes. Its feasibility is successfully assessed using data collected from the Food and Drug Administration (FDA)-accepted University of Virginia (UVA)/University of Padova metabolic simulator.

Keywords: Type 1 diabetes mellitus (T1DM), continuous glucose monitoring (CGM), latent variables (LVs), multivariate statistical analysis, glucose prediction.

1. INTRODUCTION

Type 1 diabetes mellitus (T1DM) is a disease characterized by the inability of the body to regulate glucose concentration in the blood, resulting from the autoimmune destruction of pancreatic β-cells that produce the hormone insulin (American Diabetes Association, 2003). Without proper treatment with exogenous insulin, people with T1DM cannot maintain blood glucose within normal region (80-180 mg/dL). Even with insulin replacement, subjects suffer from drastic glycemic excursions, including episodes with lower glucose than normal (hypoglycemia) and higher glucose than normal (hyperglycemia), which are both deleterious to one’s health and quality of life (Rubin and Peyrot, 1999). Recent developments in continuous glucose monitoring (CGM) devices have opened new opportunities for glycemia management of subjects with T1DM. Extensive glucose time-series data are measured and displayed in real time, which provide important information about a patient’s current glycemic state and also reveal its direction and rate of change. Empirical, or data-driven, analysis methods have been developed for glucose prediction and potentially helpful to assess glycemic variability. Many meaningful applications have been reported, indicating the potential value to patients and physicians.

In autoregressive (AR) modeling, only CGM data are analyzed and the intrinsic structure of blood glucose interactions is explored. In autoregressive with exogenous input (ARX) models, both systematic glucose dynamics and exogenous inputs (such as meal carbohydrates (CHO) and insulin administration) are investigated regarding their effects on the evolution of glucose concentrations. Eren-Oruklu and co-workers (Eren-Oruklu et al. 2009, 2010) have used time-series analysis for the development of subject-specific recursive models to predict future glucose concentrations. Finan and co-workers (Finan et al., 2007, 2008, 2009) have identified ARX models from simulated and clinical data. Relevant factors have been investigated such as threshold level, sampling frequency, and prediction horizon (Palerm et al., 2005).

By predicting future glucose concentrations, hypo/hyperglycemia alarms can be issued to alert the user at the early stage and thus allow immediate corrective action. For model-based control, the appropriate insulin dose for maintaining normoglycemia can be calculated, and the amplitudes of glucose excursion can be reduced. From this perspective, proactive glucose monitoring using data-driven statistical analysis methods is a promising approach. However, it still presents many challenges to understand the underlying glucose dynamics and their relations with exogenous factors.

In this paper, latent variable (LV)-based multivariate statistical analysis techniques are investigated for the identification of empirical models from T1DM data. In statistics, LVs (as opposed to observable variables), are not directly measured, but are inferred through a mathematical model from other directly observed variables. The information about the interdependencies and variability of the observed diabetes data can be extracted and represented by a potentially small number of LVs. Specifically, the current study attempts to address two issues. First, the underlying relationship between exogenous inputs and glucose concentration is explored from glucose-insulin data collected during conditions of changing insulin boluses and meal carbohydrate. This relationship is then used to estimate future glucose concentrations on-line. Second, the on-line prediction reliability is evaluated by comparing the predictions with CGM measurements and checking the underlying dynamics...
of prediction errors. Simulation studies with the FDA-accepted University of Virginia (UVA)/University of Padova metabolic simulator (Kovatchev et al., 2009) demonstrate the relevance of the proposed method.

2. LV-BASED MULTIVARIATE STATISTICAL MODELING AND ANALYSIS FOR GLUCOSE PREDICTION

2.1 LV-based Modeling algorithm

Partial Least Squares (PLS) is one of the most common statistical regression tools for determining a quantitative relationship between a predictor matrix $X$ and response $Y$ (Martens & Naes, 1994; Burnham et al., 1996). The underlying assumption is that the system or process under consideration can be described by a small number of LVs so that a LV model can be derived from predictor space to which the interpretation of both data spaces is directly linked.

A typical PLS algorithm is described (Lindgren et al., 1993) as below:

1. Mean-center and scale the $X$ and $Y$ matrices;
2. Set $u_1$ to be the first column of $Y$.
3. $w_j = X'u_j/(u_j'u_j$). Scale $w_j$ to be of length one.
4. $t_j = Xw_j$.
5. $q_j = Yt_j/(t_j't_j$), scale $q_j$ to be of length one.
6. $u_j = Yq_j/(q_j't_j$).
7. If convergence then (8) else (3).
8. $X$-loading: $p_j = X_j't_j/(t_j't_j$).
9. $Y_{i,j} = Y_i't_jq_j$.
10. Return to step (2) for the next LV.

Although many successful PLS applications have been reported for the specific purpose of regression analysis, one problem associated with the original PLS method should receive special attention. Its objective is to model the variations in both the $X$ and $Y$ space and maximize their covariance. But large covariance does not necessarily mean strong correlation. When the predictor space contains a considerable amount of process variations uncorrelated with response space, it is possible the PLS LV model may capture the major systematic variations in the predictor space ($X$) but have relatively weak correlation with the response ($Y$). This strategy will lead to a complex model structure and an overfitting problem. Unlike PLS, canonical correlation analysis (CCA) inherently ignores the variations in $X$ that are uncorrelated with $Y$ and directly maximizes their correlation (Anderson, 2002). However, since the measurement vectors are often high-dimensional and collinear, directly applying CCA to the raw space will lead to an ill-conditioned problem because it involves the calculation of the inverse of the covariance matrix in the model estimation. Yu and MacGregor (2004) have suggested that CCA can be used as a post-processing technique to further improve the PLS LVs (T) and relate them with $Y$. In this way, a parsimonious model with the same prediction ability as the standard PLS model can be obtained with improved model interpretability.

Based on these considerations, a PLS-CCA algorithm is employed to extract the regression relationship for the purpose of on-line prediction. Moreover, another statistical analysis method, principal component analysis (PCA) (Wold et al., 1987) is used to account for the underlying variability in the prediction errors for on-line prediction reliability evaluation.

2.2 Modeling data arrangement

The modeling process can be viewed as having one output, i.e., future glucose concentration, and two inputs, i.e., available glucose measurements, CHO intake and insulin boluses. The two inputs consist of series of impulses, corresponding to the amount of insulin boluses infused subcutaneously and the CHO content of meal, which are often administered simultaneously. To enhance the accurate identification of the effects of impulse inputs on glucose concentration, they should be filtered through transfer functions (Finan et al. 2007; Grosman et al., 2010), giving rise to new time-smoothed inputs with different “shapes”. Second-order transfer functions with different dynamics (Grosman et al., 2010) are used here, which attempt to separate the two inputs.

The transformed insulin and meal inputs up to the current time ($t$) are denoted by $u_{it}(t)$ and $u_{it}(t)$, respectively. The CGM time-series data are denoted $g(t)$, which should contain rich glycemic variability and autocorrelation information. $L$ is the number of samples included in the chosen prediction horizon (PH). For example, with 5 min as sampling frequency, $L=6$ denotes that glucose concentration 6-steps-ahead (30 min) from the current time is estimated using data available up to time ($t$). To capture the underlying predictor correlations, the original observation vectors are augmented to include time-lagged values. Assume that $D_i, D_M$ and $D_C$ are the number of samples involved in the predictor regions (i.e., predictor length, PL) of insulin, meal and glucose, respectively, and $D=\max(D_i,D_M,D_C)$. They denote the numbers of previous samples on which each prediction depends. For example, the insulin predictor matrix can be arranged as,

$$U_1(N×D_1)=\begin{bmatrix}
\text{u}_{1,1}'(1×D_i) \\
\text{u}_{1,2}'(1×D_i) \\
\vdots
\text{u}_{1,(K−D_i+1)}'(1×D_i)
\end{bmatrix}$$

where $N=K−D_1−L+1$ is the number of modeling observations and $K$ is the total number of observations. Each sampling vector $u_{it}(1×D_i)$ $(i=1,2,...,N)$ contains the insulin bolus information from time $i+D_1−D$ to $i+D_1−1$. Similarly, the other two predictors are also arranged, forming a meal data set $U_M(N×D_M)$ and a glucose data set $G(N×D_G)$ respectively. The predictor matrix is designated by including all insulin, meal and glucose information.

$$X(N×Z)=\begin{bmatrix}
U_1(N×D_1) & U_M(N×D_M) & G(N×D_G)
\end{bmatrix}$$

where $Z(D_1+D_M+D_G)$ is the number of input vectors in the final predictor matrix. Here, for the convenience of mathematical expression, the PH is assumed to be the same for all three inputs. The response vector $y(N×1)$ is arranged with the $L$-step-ahead glucose value corresponding to each row vector $x_i'(1×Z)$ in $X(N×Z)$. The number of modeling
observations (N) changes with the different choices of PH and PL as indicated by K−D−L+1. In this way, regression modeling data pairs \( \{X, y\} \) are arranged, from which their correlation is decomposed by PLS-CCA, and the predictor variations correlated to response are described by LVs.

2.3 Regression modeling and prediction

Considering the modeling data pair, \( \{X, y\} \), data normalization is first performed in each data set. The PLS-CCA algorithm is then employed to relate the underlying information of the two data matrices:

\[
\begin{align*}
X &= Xw \\
\hat{X} &= tp^T, X = tp^T + E \\
t &= Xw
\end{align*}
\]

where \( w(Z \times 1) \) is the PLS-CCA weight and \( t(N \times 1) \) is the extracted LVs, indicating the systematic variations in predictor space closely related to response. Here it should be noted that as a result of the PLS-CCA algorithm, only one LV is needed for the single response vector, \( y \). That is, all predictor variations that are correlated with response are collected by \( t(N \times 1) \). Vector \( p(Z \times 1) \) is the loading of predictors and \( q \) is the loading coefficient of the response. Based on the extracted LVs, the future glucose prediction \( \hat{y} \) is made and the predictor variations are also modeled as \( \hat{X} \).

In this way, instead of the direct use of external measurements, the underlying systematic glycemic variability that is closely related with the response is characterized by one LV and extracted from the existing glucose measurements, insulin infusion and meal intake. The other variation information that is uninformative for inferring future glucose is excluded and will not influence the prediction model.

During an on-line application, the newly available predictor vector \( x_{\text{new}}(1 \times Z) \) at each time can be composed as \( [u_{\text{new}}(1 \times D_1), u_{\text{new}}(1 \times D_2), \ldots, u_{\text{new}}(1 \times D_k)] \) by including their L-step-ahead information. The prediction \( \hat{y}_{\text{new}} \) is then made at each time:

\[
\begin{align*}
t_{\text{new}} &= x_{\text{new}}^Tw \\
\hat{y}_{\text{new}} &= t_{\text{new}}q \\
f_{\text{new}} &= y_{\text{new}} - \hat{y}_{\text{new}}
\end{align*}
\]

The predictions can be used to monitor whether the future glucose concentration will violate a safety limit after \( L \) steps.

2.4 On-line prediction reliability evaluation

When predictions indicate large glucose excursions, some corrective actions can be taken to avoid the violation and keep the glucose concentration within the confidence limits. However, inappropriate control actions based on the unreliable prediction may adversely affect the glucose excursion. Therefore, before taking any corrective action, a “reliability evaluation” should be performed based on past predictions to indicate the possibility of accepting current predictions.

Considering the process evolution in time, only the predictable glucose dynamics are captured by PLS-CCA. The prediction errors include not only measurement noise but also reflect the underlying glucose dynamics that cannot be estimated from previous measurement information. From this viewpoint, the underlying glucose variability hidden in the past prediction errors can be explored using PCA. PCA (Wold et al., 1987) can transform a number of possibly correlated variables into a smaller number of uncorrelated variables defined as principal components (PCs). Computationally, PCA is usually handled by computing eigenvectors of the covariance matrix of the data set. Often, its operation can be interpreted as revealing the internal structure of the data in a way that best explains the variance. The first few PCs account for most of the variability in the data.

The analysis is based on \( F_f(N \times D_1) \), which is constructed from time-series prediction errors \( f \) in analogy with Eq. (3). \( D_f \) is the number of included prediction error variables and \( N \neq N \times D_f + 1 \) is the number of modeling observations. PCA is then performed on the residual matrix \( (F) \) to decompose the systematic variation information \( (T) \) and the final residuals \( (F) \):

\[
F = T_fP_f^T + F_f = FP_fP_f^T + F_f
\]

where the underlying systematic variations are formulated by PCA scores \( T_f(N \times R_f) \) and separated from the final residuals \( F_f \), i.e., measurement noise. \( P_f \) is the PCA loadings, which reveal the correlation between PCA scores \( (T_f) \) and the variables \( (F_f) \).

For prediction reliability checking, two types of monitoring statistics can thus be calculated: Hotelling’s \( T^2 \)-statistic (Lowry and Montgomery, 1995) and the sum of the prediction error \( (SPE) \) statistic (Jackson and Mudholkar, 1979), describing the modeled systematic part and the residuals, respectively:

\[
T_{f,\alpha}^2 = (t_{f,\alpha} - \bar{T}_f)^T\Sigma_f^{-1}(t_{f,\alpha} - \bar{T}_f) \quad \text{SPE}_{f,\alpha} = \bar{T}_f^Tf_{f,\alpha}
\]

where \( t_{f,\alpha}(R_f \times 1) \) and \( f_{f,\alpha}(D \times 1) \) are respectively the PCA score and residual vector at the \( \alpha \)th time; \( \bar{T}_f \) denotes the corresponding mean vector and \( \Sigma_f \) is the variance-covariance matrices of \( T_f(N \times R_f) \).

A Gaussian-distribution hypothesis provides an important basis for deriving the confidence limits of monitoring statistics. The \( T^2 \) control limits can be defined by an \( F \)-distribution with \( \alpha \) significance factor (Lowry and Montgomery, 1995). In error subspace, the representative confidence limit of SPE can be approximated by a weighted Chi-squared distribution, \( g^2_{p,n} \) (Jackson and Mudholkar, 1979). In this way, the reference normal variation region is defined, and the dynamics of new prediction errors can be checked in real time.

During an on-line application, the new monitoring statistics are calculated based on the available prediction errors. If a statistic exceeds its limit, an alarm is issued,
indicating incorrect past predictions, so that the current prediction should be taken with caution.

3. SIMULATION AND DISCUSSION

As a proof-of-concept study, the simulated data are generated with a 5 min sampling interval using the FDA-accepted UVA/Padova metabolic simulator (Kovatchev et al., 2009). In order to simulate daily life, the study is performed open-loop with no adjustments to the prescribed daily routine. It includes a three-meal scenario for breakfast, lunch, and dinner taken at approximately 7am, 12 noon and 6pm with 60g, 80g and 80g CHO, respectively. To simulate realistic ambulatory conditions, the meal times and CHO amounts are varied to accommodate variations in daily life, which are set to be one hour changes (forward or backward) and 10g CHO increment or decrement in the present work. The insulin bolus is given immediately in response to CHO absorbed from a meal. As shown in Table 1, the insulin bolus amounts are artificially set to accommodate a certain variation in daily life. The relative insulin-to-carbohydrate ratio (RI:CR) is defined as the ratio of actual insulin bolus to the "optimal" subcutaneous (SQ) insulin bolus; it varies from 0.75 to 1.25. Ten days of normal data are simulated for one subject (adult #1) and used for model training with two days corresponding to each RI:CR value. The PH is set to be 30 min. For each predictor term, six samples (i.e., 30 min) are included, providing 18 predictor variables for modeling. The glucose prediction and reliability evaluation system is set up for \( D = 12 \) and then the system is used for on-line application. To check the generalization capability of the prediction model, besides the RI:CR values considered in the training data, three extra cases with different insulin boluses (RI:CR set to be 1.5, 0.5 and 0, respectively) are also taken into account as shown in Table 1. In each case, five-day data are generated for testing.

The prediction model training results are shown in Figure 1(a), where the fitted model is accurate and the glucose trends are captured. The rate error grid analysis (REGA) metric (Kovatchev et al., 2004) is used to quantify the prediction performance in terms of rate of glucose change. 82.72% of the data points are in Zone A, which is considered clinically accurate; 10.84% in Zone B, 4.07% in Zone C, 0.71% in Zone D and only 1.66% in Zone E. Based on these prediction results, the monitoring system for prediction reliability evaluation is also developed as shown in Figure 1(b) where the 99% confidence limit is shown.

![Figure 1](image1.png)

Figure 1 Model training results: (a) predicted glucose concentrations and measurements; (b) monitoring statistics for prediction reliability evaluation.

![Figure 2](image2.png)

Figure 2(a) Online glucose prediction for testing data with 1.5 RI:CR, (b) on-line prediction reliability evaluation.

Considering the three testing cases where RI:CR values are 1.5, 0.5 and 0 respectively, which are not included in the training data, the on-line prediction and checking performance is illustrated in Figures 2&3. For clarity, only a two-day glucose profile is plotted for each case. When RI:CR=1.5 in Figure 2(a), generally, the glucose trends are
captured, although around the 300th sampling instant, the prediction error is larger. Based on Figure 2(b), SPE monitoring statistics exceed their confidence limits during the time period around 300, indicating that the prediction accuracy may not be acceptable, which agrees with the results shown in Figure 2(a). Throughout the other time regions, the estimation accuracy is satisfactory since both monitoring statistics stay well below the 99% confidence limits.

For simulated Subject #1, the modeling and prediction performance based on LV with exogenous input (LVX) is summarized in Table 1. Moreover, LV results with only CGM data are also shown in Table 1. The coefficient of determination ($R^2$), a commonly used statistical metric that describes how much of the variability in the data is captured by the prediction model, is used to quantify the prediction accuracy. The larger the $R^2$ value is, the better the prediction model is. In general, the prediction performance improves when insulin and meal inputs are included in the model (i.e., LVX vs. LV). Also, prediction accuracy decreases as PH increases, as would be expected. Table 1 also indicates that the identified LV model (for RI:CR=1) is also accurate when the RI:CR ratio has changed significantly, for example, due to an inaccurate CHO estimate.

![Figure 3 (a) Online glucose prediction for testing data with 0.5 RI:CR, (b) on-line prediction reliability evaluation.](image1)

![Figure 4 (a) Online glucose prediction for testing data with 0 RI:CR, (b) on-line prediction reliability evaluation.](image2)

The new LV approach was also evaluated for ten simulated adults as shown in Table 2. A value of RI:CR=1 and the same variations in meal timings and meal amounts were considered. Ten-day data were used for modeling and five days for testing. The prediction performance based on LVX/LV method is compared with that for the AR/ARX in Table 2. The PH is six. Six samples (i.e., 30 min of data) for each predictor term were used in the prediction model, providing a total of 18 predictor variables. For the ARX a value of zero was assigned to the unknown future inputs.
The LVX method provides the best results. Different ARX model orders were evaluated for the meal and insulin inputs. However, the results are still worse than the LVX and LV results. In general, the ARX models are more accurate than the AR models.

4. CONCLUSIONS

In this paper, an LV-based statistical analysis and modeling method for glucose predictive monitoring of type 1 diabetes has been proposed. On-line predictions and prediction evaluations were made in a simulation study involving ten TIDM subjects. The models based on the LV method provided more accurate predictions than conventional AR/ARX models. In the future, clinical evaluations are needed to fully evaluate the proposed approach.

ACKNOWLEDGEMENTS

The research work is supported by the Juvenile Diabetes Research Foundation (grants 22-2009-797, 22-2009-76), and the National Institute of Health (grant DK085628-01).

REFERENCES


<table>
<thead>
<tr>
<th>Methods</th>
<th>Adult subject #</th>
<th>Mean±Std</th>
</tr>
</thead>
<tbody>
<tr>
<td>LVX</td>
<td>92</td>
<td>89±4</td>
</tr>
<tr>
<td>LV</td>
<td>92</td>
<td>89±4</td>
</tr>
<tr>
<td>ARX</td>
<td>92</td>
<td>89±4</td>
</tr>
<tr>
<td>AR</td>
<td>92</td>
<td>89±4</td>
</tr>
</tbody>
</table>

Table 1 Summary of $R^2(\%)$ prediction results for adult #1

![Table](https://example.com/table1.png)

Table 2 Summary of $R^2(\%)$ prediction results for ten adult subjects

![Table](https://example.com/table2.png)