33rd Annual International Conference of the IEEE EMBS
Boston, Massachusetts USA, August 30 - September 3, 2011

Assessment of Linear Regression Techniques for Modeling Multisensor Data for Non-Invasive Continuous Glucose Monitoring

Mattia Zanon, Michela Riz, Giovanni Sparacino, Andrea Facchinetti, Roland E. Suri, Mark S. Talary and Claudio Cobelli, *Fellow, IEEE*

***Abstract*— New scenarios in diabetes treatment have been opened in the last ten years by continuous glucose monitoring (CGM) sensors. In particular, Non-Invasive CGM sensors are particularly appealing, even though they are still at an early stage of development. Solianis Monitoring AG (Zürich, Switzerland) has proposed an approach based on a multisensor concept, embedding primarily dielectric spectroscopy and optical sensors. This concept requires a mathematical model able to reconstruct the glucose concentration from the 150 channels measured with the device. Assuming a multivariate linear regression model (valid and usable for different individuals), the aim of this paper is the assessment of some techniques usable for determining such a model, namely Ordinary Least Squares (OLS), Partial Least Squares (PLS) and Least Absolute Shrinkage and Selection Operator (LASSO). Once the model is identified on a training set, the accuracy of prospective glucose profiles estimated from "unseen" multisensor data is assessed. Preliminary results obtained from 18 in-clinic study days show that sufficiently accurate reconstruction of glucose levels can be achieved if suitable model identification techniques, such as LASSO, are considered.**

I. INTRODUCTION

DIABETES is a disease that affects 285 million people in the world and this number is expected to increase to 439 million in 2030, thus making diabetes an "epidemic" disease [1]. In healthy people, glucose levels in the blood are controlled by insulin using a negative feedback. In people with diabetes, the body does not secrete insulin (type 1 diabetes) or imbalances in both insulin secretion and action (type 2 diabetes) occur. Therapy is mainly based on insulin administration and diet, which are tuned by self-monitoring of blood glucose (SMBG) levels 3-4 times a day. Nevertheless, blood glucose concentration of the patients is often outside of the normal range of 70-180 mg/dL. While hyperglycemia mostly affects long-term complications (such as neuropathy, retinopathy, cardiovascular, and heart diseases), hypoglycemia can be very dangerous in the short-term and, in the worst-case scenario, may bring the patient into hypoglycemic coma. New scenarios in diabetes treatment were opened in the last ten years, when minimally invasive continuous glucose monitoring (CGM) sensors, able to monitor glucose concentration continuously (i.e. with a reading every 1-5 min) for several days (up to 7 consecutive days), entered clinical research. It has been suggested that the retrospective assessment of glucose profiles measured through CGM sensors might help in the optimization of metabolic control [2] in people with diabetes. On-line applications are potentially more appealing and with a greater impact in the patient daily life. Ideally these would include the "smart CGM sensor", i.e. a system able to generate alerts when glucose concentrations exceed the normal range thresholds, combined with "the artificial pancreas", i.e. a device conceived for Type 1 people with diabetes aimed at maintaining glucose concentration within safe ranges by infusing subcutaneously insulin via a pump under the control of a closed-loop algorithm (see [3] and [4] for reviews). Most of the CGM sensors are based on the glucose-oxidase principle and they are called "minimally invasive", because a thin needle must be inserted in the subcutis. Non-Invasive Continuous Glucose Monitoring (NI-CGM) technologies have been also investigated [5], [6], and their ability to monitor glucose changes in the human body has been demonstrated under highly restricted conditions [7], [8]. As soon as these conditions become less favourable, e.g. in daily-life use, several problems have been experienced due to physiological and environmental perturbations [9]. Solianis Monitoring AG (Zürich, Switzerland) recently proposed a multisensor approach for NI-CGM mainly based on dielectric and optical sensors. Such an approach has the aim of achieving a broader bio-physical characterization of skin and underlying tissues in order to account for confounding factors which can significantly deteriorate the accuracy of glucose readings. In particular, dielectric spectroscopy tracks changes in the dielectric properties of skin and underlying tissues modulated by glucose level variations and by other physiological "perturbing factors" [10]. Optical, as well as temperature, humidity, and movement sensors, embedded within the same sensor substrate, can provide useful information for the compensation of such "perturbing factors" [10]. In this multisensor approach, 150 individual measurements are collected every 20 seconds. Fig. 1 (left) shows some representative data taken from 16 channels. Fig. 1 (right) shows reference glucose concentrations measured in parallel through a standard laboratory glucose analyzer (HemoCue). A crucial point is thus to design and identify a mathematical model which can combine the information provided by the above mentioned 150 multisensor channels in order to obtain an estimate of the glucose level (Fig. 1, middle). The method to combine several time-series (i.e. the 150 multisensor channels) in a model with the aim of estimating a target variable (i.e. HemoCue glucose readings) can be described as a multivariate regression problem. In this work, we consider

Manuscript received April 15, 2011, accepted June 9, 2011.

M. Zanon, M. Riz, G. Sparacino, A. Facchinetti and C. Cobelli* are with the Department of Information Engineering, University of Padova, Via G. Gradenigo 6/B, 35131 Padova, Italy (* corresponding author, phone: +39 049 827 7803; fax: +39 049 827 7826; e-mail: cobelli@dei.unipd.it)

R. Suri and M. Talary are at Solianis Monitoring AG, Zürich, Switzerland.

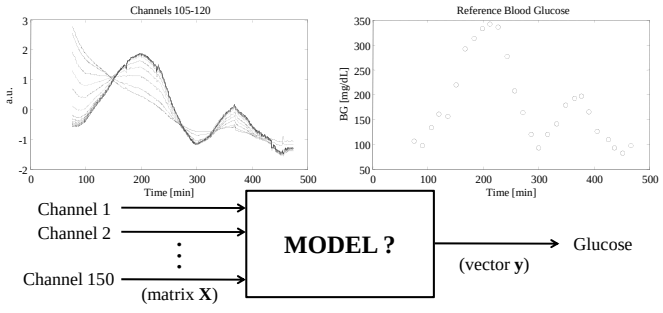

Fig. 1. Scheme of model building. Data correspondent to the 150 multisensor channels (matrix X, left) are combined through a model (to be built, middle) with the aim of inferring glucose levels measured through a gold-standard technique (vector y, right).

a static linear regression model determined with data from different subjects and thus with a global validity, and we investigate three techniques usable for parameter estimation, i.e. Ordinary Least Squares (OLS), Partial Least Squares (PLS) and Least Absolute Shrinkage and Selection Operator (LASSO). For such a scope, multisensor data and reference glucose, obtained in 18 in-clinic study days, are considered. Performance of the three parameter estimation techniques were initially assessed in terms of "internal validation" criteria over the first 9 study days. Then, the identified models undertook the so-called "external validation" phase in order to assess their usability to reconstruct glucose concentration from multisensor data that have not participated to model derivation (remaining 9 study days). Results show that the regularization performed within the LASSO method is of crucial importance when the aim is prospective estimation of glucose from unseen multisensor data.

II. DATA BASE

Data was gathered from 6 subjects with Type 1 Diabetes Mellitus (T1DM) as part of an experimental clinical study approved by the local ethical review commission and run according to the requirements of GCP. Reference glucose and multisensor data were available for each of the 18 study days considered in this present paper. The length of an experiment was 8 hours during which plasma glucose was induced to vary according to a predetermined profile either orally or by i.v. glucose administration. Multisensor data was obtained by placing the multisensor in the upper arm with a sampling time interval of 20 sec, while reference glucose values were acquired in parallel, every 10 to 20 min, using a HemoCue Glucose Analyzer. The database, consisting of roughly 70000 multisensor data points and 2000 glucose reference points, was then split in two subsets of 9 study days each of approximately the same dimension. Each Multisensor channel was centered and scaled and a causal median filter for removing outliers was then applied.

Data for Model Identification. The first dataset, subsequently called "the training set", was first used to perform a k-fold cross validation analysis for model complexity selection and, subsequently, to estimate model parameters by the techniques described in Section III.

Data for Model Test. The second data set was used to test the previously identified models over new, "unseen" data, as external validation reported in Section IV.

III. MODEL IDENTIFICATION

A. Problem Statement

Formally, the linear regression model is given by:

$$y = X\beta + v \tag{1}$$

where y is a $(N \times 1)$ vector representing the target variable (glucose), β is the $(p \times 1)$ vector containing the unknown coefficients of the global linear model (i.e. valid for all the subjects), X is the $(N \times p)$ matrix whose columns contains the time series measured by the multisensor and v the $(N \times 1)$ vector of the errors depicting the data unexplained by the model. Here p is around 150, whereas N is of the order of some thousands and typically depends on the number of study days considered. Denoting the Residual Sum of Squares as:

$$RSS(\beta) = [(y - X\beta)^T (y - X\beta)] \tag{2}$$

the model parameters can be estimated by several techniques, and in particular by means of the three methods described in the next paragraph.

B. The Three Chosen Parameter Estimation Methods

Ordinary Least Squares (OLS)

OLS is a well known method for the estimation of linear regression models. The solution $\hat{\beta}^{OLS}$ minimizes the distance between the reference values contained in y and the model predictions obtained for a certain parameter vector $\hat{\beta}$:

$$\hat{\beta}^{OLS} = \underset{\hat{\beta}}{\operatorname{argmax}} RSS(\hat{\beta}). \tag{3}$$

There exists a closed form solution for the calculation of $\hat{\beta}^{OLS}$. Since the high dimension of the measurement space and the high correlation between subset of predictors cause the X matrix to be rank-deficient (making the problem ill-conditioned) a QR decomposition of X is used for estimating the parameter vector $\hat{\beta}^{OLS}$

$$\hat{\beta}^{OLS} = (X^T X)^{-1} X^T y = R^{-1} Q^T y \tag{4}$$

where $X = QR$, with Q being a $(N \times p)$ orthogonal matrix and R a $(p \times p)$ upper-triangular matrix.

Partial Least Squares (PLS)

PLS is a technique for the estimation of linear regression models resorting to an idea also used by Principal Component Analysis. In particular, PLS tries to find $m(< p)$ new variables, the so-called latent variables, with high variance and exhibiting high correlation with the target variable y [11]. The PLS estimate of the parameter vector is given by:

$$\hat{\beta}^{PLS} = W\theta \tag{5}$$

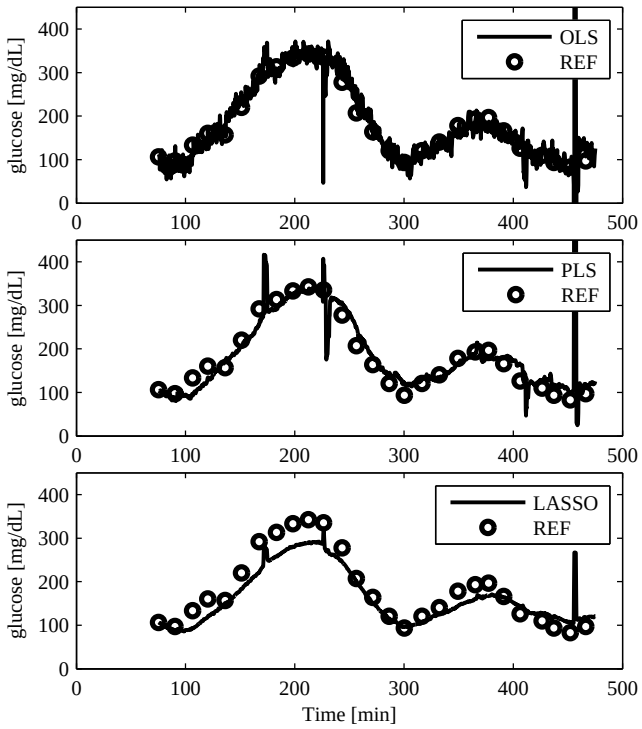

Fig. 2. Study day of representative subject # 2. Reference glucose (circles) *vs.* estimated glucose profile (lines) by means of OLS (top), PLS (middle) and LASSO (bottom) in internal validation.

where W is a $(p \times m)$ matrix representing the kernel of the transformation from the original to the new latent variables and θ is a $(m \times 1)$ parameter vector of the linear model from the latent variables to the target y. See [12] for more details. The number of new latent variables m controls model complexity, since it is strictly related to the amount of variance retained from the original variables.

Least Absolute Shrinkage and Selection Operator (LASSO)

The LASSO solution minimizes a cost function consisting of (2) plus a regularization term over the model parameter vector $\hat{\beta}$:

$$\hat{\beta}^{LASSO} = \underset{\hat{\beta}}{\operatorname{argmax}} \left(RSS(\hat{\beta}) + \lambda \sum_{j=1}^{p} |\hat{\beta}_j| \right) \qquad (6)$$

λ controls the model complexity and avoids the coefficient of the linear model to assume large absolute values, thus preventing overfitting as might happen with OLS. Thus, besides shrinking the linear model coefficients, LASSO also performs variable selection according to λ. For values of λ sufficiently small, the coefficients of some variables are exactly zero, making easier the interpretation of the results [13]. The nature of the regularization term makes the LASSO solution non linear in y and its computation a quadratic programming problem. The solution can be obtained efficiently for example with a modification of the LAR algorithm [14].

C. Assessment Criteria for Model Identification

The parameters controlling the model complexity for PLS and LASSO can be estimated by 10-fold cross validation [11]. Model parameters were then estimated with the three aforementioned techniques. Afterwards, each model was used to predict glucose profiles with the same multisensor data used for its identification. Each estimated glucose profile was first shifted to the value of the first reference glucose value and then compared with the reference HemoCue measures by means of the Root Mean Squared Error (RMSE) and the Pearson coefficient of determination (R^2).

D. Internal Validation Results

By visual inspection of Fig. 2 and observing the key indicators for internal validation (Table I), it seems that the linear model obtained with OLS outperforms both PLS and LASSO in estimating glucose profiles when data used for the estimation of the models are considered. This is obviously expected since OLS estimates the model parameters in such a way to maximize the adhesion to the training data without any constraint on the complexity. However, this indicates that OLS overfitted the training data with the risk of leading to poor generalization when trying to estimate glucose from the test set multisensor data (see Section IV).

IV. MODEL TEST

A. Assessment Criteria for Model Test

Models identified in Section III-B have been tested for validity using a multisensor test data set unseen during the model derivation stage. The relative glucose estimates undergo an initial adjustment of the baseline considering the first reference glucose value as happened in Section III-C. This means that a blood glucose finger stick measure is required in an every-day use setting to allow for the adjustment of the baseline. The performance of the three models in prospective glucose profile estimation is assessed by means of the same indexes used in internal validation.

TABLE I
KEY INDICATORS FOR INTERNAL AND EXTERNAL VALIDATION. ROOT MEAN SQUARED ERROR (RMSE) AND PEARSON COEFFICIENT OF DETERMINATION (R^2).

	RMSE [mg/dL]		R^2 (averaged over runs)	
	Internal Validation	External Validation	Internal Validation	External Validation
OLS	15	781.5	0.97	0.44
PLS	39.6	92.6	0.88	0.63
LASSO	61.3	66	0.79	0.61

B. External Validation Results

The key indicators in Table I show that the LASSO estimated model outperforms both OLS and PLS thus achieving better generalization performances in predicting "unseen" data (see Fig. 3). Prediction accuracy is improved because LASSO forced a sparse solution (with many coefficients shrank to zero) and estimated parameters with low absolute

values, trading off decreased variance for increased bias [11]. While RMSE is worse for PLS than for LASSO, the averaged correlation coefficients (R^2) over study days of the profiles for the two methods are comparable, indicating that, although PLS might give good prediction of glucose trends, it is too sensitive to noisy channels (see Fig. 3). This might be due to the fact that noisy channels containing glucose or confounding factors information might be used by PLS for building new latent variables. Instead, the regularization term in the cost function of LASSO leads to the selection of original variables that are likely to be less sensitive to noise.

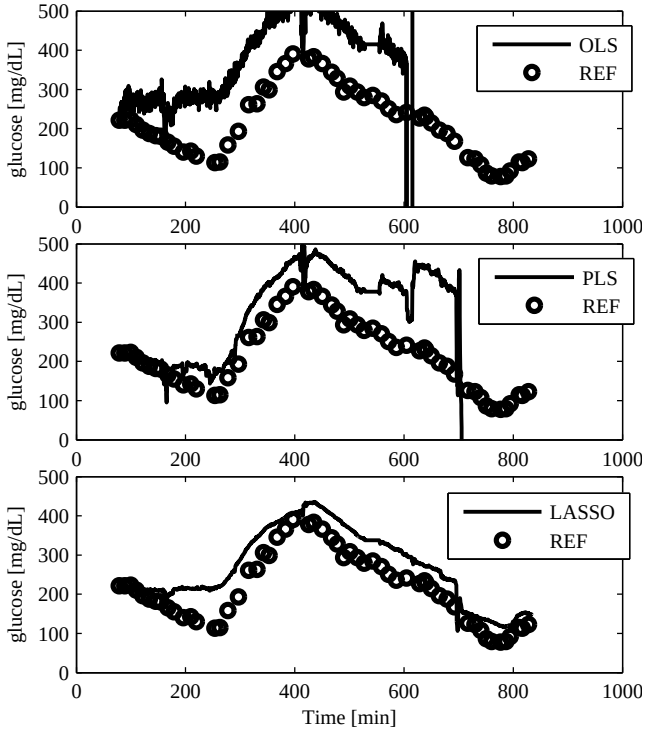

Fig. 3. Study day of representative subject # 2, different from the one showed in Fig. 2. Reference glucose (circles) *vs.* estimated glucose profile (lines) by means of OLS (top), PLS (middle) and LASSO (bottom) in external validation.

V. CONCLUSIONS

CGM sensors allow monitoring glucose concentration continuously for several days and are of great interest in both research and clinical practice. NI-CGM technologies are still at an early stage of development, but they are particularly appealing for obvious reasons related to patient's comfort. The multisensor platform for NI-CGM proposed by Solianis Monitoring AG (Zürich, Switzerland) has been considered in this work. In this multisensor approach, a mathematical model must be developed to reconstruct glucose concentration from the data of 150 measurement channels. The objective of this paper was the assessment of some techniques usable for such a purpose assuming a linear regression model. Nine study day data were considered for model identification and nine, collected from the same subjects, for model test. Results showed that both LASSO and PLS avoid the typical overfitting occurring with OLS. LASSO outperforms PLS in external validation performances since the latter seems less robust to noisy channels. Indexes in Table I show that point accuracy performance of the non-invasive sensors are not yet comparable with state of the art "gold standard" SMBG sensors (HemoCue). However, thanks to its continuous nature, the non-invasive multisensor device combined with the LASSO model may be a good candidate for being used as a complement to SMBG. Further developments of this work will be focused on the confirmation of the results over a wider data set and on the consideration of more key indicators for assessing models performance and the robustness of the model to data sets from new subjects not part of the model derivation data set. Moreover, strategies for improving the calibration of the reconstructed glucose profiles can be developed [15].

REFERENCES

[1] J. E. Shaw, R. A. Sicree, and P. Z. Zimmet, "Global estimates of the prevalence of diabetes for 2010 and 2030," *Diabetes Res. Clin. Pract.*, vol. 87, pp. 4–14, 1 2010.

[2] B. W. Bode and T. Battelino, "Continuous glucose monitoring," *Int. J. Clin. Pract. Suppl.*, vol. (166), pp. 11–15, Feb 2010.

[3] G. Sparacino, A. Facchinetti, and C. Cobelli, ""Smart" continuous glucose monitoring sensors: On-line signal processing issues," *Sensors*, vol. 10, no. 7, pp. 6751–6772, 2010.

[4] C. Cobelli, C. D. Man, G. Sparacino, L. Magni, G. D. Nicolao, and B. P. Kovatchev, "Diabetes: Models, Signals, and Control," *IEEE Rev. Biomed. Eng.*, vol. 2, pp. 54–96, Jan 1 2009.

[5] A. Tura, A. Maran, and G. Pacini, "Non-invasive glucose monitoring: assessment of technologies and devices according to quantitative criteria," *Diabetes Res. Clin. Pract.*, vol. 77, pp. 16–40, Jul 2007.

[6] A. Caduff, M. S. Talary, M. Mueller, F. Dewarrat, J. Klisic, M. Donath, L. Heinemann, and W. A. Stahel, "Non-invasive glucose monitoring in patients with type 1 diabetes: a multisensor system combining sensors for dielectric and optical characterisation of skin," *Biosens. Bioelectron.*, vol. 24, pp. 2778–2784, May 15 2009.

[7] K. V. Larin, M. S. Eledrisi, M. Motamedi, and R. O. Esenaliev, "Non-invasive blood glucose monitoring with optical coherence tomography: a pilot study in human subjects," *Diabetes care*, vol. 25, pp. 2263–2267, Dec 2002.

[8] M. A. Arnold and G. W. Small, "Noninvasive glucose sensing," *Anal. Chem.*, vol. 77, pp. 5429–5439, Sep 1 2005.

[9] A. Caduff, F. Dewarrat, M. Talary, G. Stalder, L. Heinemann, and Y. Feldman, "Non-invasive glucose monitoring in patients with diabetes: a novel system based on impedance spectroscopy," *Biosens. Bioelectron.*, vol. 22, pp. 598–604, Dec 15 2006.

[10] A. Caduff, M. S. Talary, and P. Zakharov, "Cutaneous blood perfusion as a perturbing factor for noninvasive glucose monitoring," *Diabetes Technol. Ther.*, vol. 12, pp. 1–9, Jan 2010.

[11] T. J. Hastie, R. Tibshirani, and J. H. Friedman, *The elements of statistical learning: data mining, inference, and prediction*. New York: Springer, 2009.

[12] S. de Jong, "SIMPLS: An alternative approach to partial least squares regression," *Chemometrics Intellig. Lab. Syst.*, vol. 18, pp. 251–263, 3 1993.

[13] R. Tibshirani, "Regression shrinkage and selection via the Lasso," *Journal of the Royal Statistical Society, Series B*, vol. 58, pp. 267–288, 1994.

[14] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, "Least angle regression," *The Annals of Statistics*, vol. 32, pp. pp. 407–451, Apr. 2004.

[15] S. Guerra, A. Facchinetti, G. Sparacino, G. D. Nicolao, and C. Cobelli, "Comparison of four methods for on-line calibration of cgm data," in *Diabetes Technology Meeting (DTM)*, p. A51, 2009.