# Sparse Principal Component Analysis for the parsimonious description of glucose variability in diabetes

C. Fabris, A. Facchinetti, G. Sparacino, and C. Cobelli

*Abstract*— **Abnormal glucose variability (GV) is considered to be a risk factor for the development of diabetes complications. For its quantification from continuous glucose monitoring (CGM) data, tens of different indices have been proposed in the literature, but the information carried by them is highly redundant. In the present work, the Sparse Principal Component Analysis (SPCA) technique is used to select, from a wide pool of GV metrics, a smaller subset of indices that preserves the majority of the total original variance, providing a parsimonious but still comprehensive description of GV. In detail, SPCA is applied to a set of 25 literature GV indices evaluated on CGM time-series collected in 17 type 1 (T1D) and 13 type 2 (T2D) diabetic subjects. Results show that the 10 GV indices selected by SPCA preserve more than the 75% of the variance of the original set of 25 indices, both in T1D and T2D. Moreover, 6 indices of the parsimonious set are shared by T1D and T2D.**

## I. Introduction

Diabetes is a chronic disease that occurs when the pancreas is no longer able to produce insulin (type 1 diabetes, T1D), or when the body cannot use it effectively to inhibit glucose production and stimulate glucose utilization (type 2 diabetes, T2D) [1]. Prolonged raised blood glucose levels and anomalous glucose variability (GV) induced by diabetes are both considered to be risk factors for the development of long-term complications from diabetes. GV, in particular, has become the focus of considerable attention in the last decades, and several efforts have been devoted to design indices able to quantify it from either sparse self-monitoring blood glucose measurements or continuous glucose monitoring (CGM) profiles [2,3]. Popular metrics proposed in the literature include mean, standard deviation and coefficient of variation of glucose readings, number of readings within, above and below the euglycemic range, indicators measuring the amplitude of glucose excursions, parameters derived from nonlinear transformations of glucose values and quantifying the overall quality of glycemic control [4-10].

Despite the large number of available GV indices, a "gold-standard" metric to quantify GV has not been identified yet, and a combination of indices is very likely to be needed in order to extensively characterize GV from glucose profiles. However, because some indices have very similar mathematical formulations or measure almost the same physiological entity, considering all available GV metrics may provide highly redundant information, and, actually, some indices could be of limited added value in the characterization of GV within a diabetic population. Thus, a method that allows reducing redundancy in the characterization of GV, selecting a subset of indices that provides a parsimonious but still comprehensive GV description, could be desirable and necessary as well.

In this context, the Sparse Principal Component Analysis (SPCA) technique is here proposed as a method to select a small subset of GV indices from a wider pool of metrics, preserving the majority of the total original variance of the data. SPCA is applied to a pool of established literature GV indices evaluated on CGM time-series datasets from both T1D and T2D, and results are compared to assess if there are selected GV indices shared by the two types of diabetes.

C. Fabris, A. Facchinetti, G. Sparacino, and C. Cobelli are with the Department of Information Engineering, University of Padova, Padova, 35131 Italy (e-mails: chiara.fabris@dei.unipd.it, facchine@dei.unipd.it, gianni@dei.unipd.it, cobelli@dei.unipd.it).

## II. Materials and Methods

### A. Database

Two CGM datasets of T1D and T2D subjects, respectively, were considered in the analysis.

The T1D dataset was collected within the EU FP7 project "Diadvisor" (2008-2012) using the Dexcom SEVEN Plus CGM System and is made up of 4-day CGM time-series acquired from 17 T1D males.

The T2D dataset was made available during the EU FP7 project "Mosaic" and consists of CGM time-series measured by the Medtronic Guardian REAL-Time CGM System in 13 T2D males, for an average period of 6 days.

### B. The original pool of 25 GV indices

Twenty-five well-established literature indices for GV quantification were considered.

In detail, this pool of metrics includes mean and sample standard deviation (SD) of all glucose readings, percentage coefficient of variation (%CV), mean of daily SDs ($SD_w$), SD of daily means ($SD_{dm}$), J-index, percentages of values below, within and above the euglycemic target range (70-180 mg/dl), 50$^{th}$ percentile, Inter-Quartile Range (IQR), range of glucose readings, and the Mean Amplitude of Glycemic Excursions (MAGE) index.

Moreover, measures derived from nonlinear transformations of glucose values and quantifying the risk associated to a glucose profile were also evaluated. In particular, we considered the Low and High Blood Glucose Indices (LBGI, HBGI), the Average Daily Risk Range (ADRR), and the Blood Glucose Risk Index (BGRI); the Hyperglycemic and Hypoglycemic Indices, and the Index of Glycemic Control (IGC); the Glycemic Risk Assessment Diabetes Equation (GRADE) score, and the three contributions due to the different glycemic states, i.e., %GRADEhypo, %GRADEeu, %GRADEhyper; and, finally, the $M_{100}$ index. For all these indices, hypo- and hyperglycemic thresholds, when involved in the calculation, were set at 70 and 180 mg/dl, respectively. We refer the

978-1-4244-7929-0/14/$26.00 ©2014 IEEE

reader to [4-10] for recent literature review providing detailed mathematical descriptions of the considered indices.

The 25 GV indices were evaluated on the CGM time-series of the dataset under analysis. Then, before entering SPCA, each GV metric was mean centered and scaled (i.e., divided by its SD) in order to avoid any bias in SPCA results and facilitate the comparison of different datasets.

### C. SPCA

SPCA is a two-step data processing technique introduced by Zou et al. in [10]. In our implementation (see also [11] for further details), the two steps consist in:

*(a)* apply PCA to a matrix containing the normalized GV indices to retrieve the principal components (PCs) and select a small number of them (typically, 1 or 2), saving the majority of the original variance of the data;

*(b)* since each PC is a linear combination (regression) of all GV indices and has no direct physiological meaning, apply to each of the selected PCs the Least Absolute Shrinkage and Selection Operator (LASSO) constraint, which allows obtaining sparse estimates of regression coefficients, maintaining in the PC regressor a reduced number of GV indices and still saving a high percentage of the total original variance.

#### *(a) PCA and formulation of the regression problem*

Let $\boldsymbol{X}$ be the $n$ x $m$ data matrix, where $n$ is the number of observations (i.e., subjects of the considered dataset) and $m$ is the number of variables (i.e., calculated GV indices). Let the covariance matrix of the data $\boldsymbol{X}$ be a general full matrix (i.e., the $m$ variables are correlated among them). The aim of PCA is to find a linear transformation

$$\boldsymbol{Y} = \boldsymbol{XS} \tag{1}$$

that allows reducing the correlation of the original data. In particular, we are searching for an orthogonal matrix $\boldsymbol{S}$, such that the transformed data $\boldsymbol{Y}$ have a diagonal covariance matrix. Singular Value Decomposition (SVD) of the data matrix, used to compute PCA, allows expressing $\boldsymbol{X}$ as

$$\boldsymbol{X} = \boldsymbol{U}\Sigma\boldsymbol{V}^{\boldsymbol{T}} \tag{2}$$

where $\boldsymbol{U}$ is an $n$ x $n$ orthogonal matrix collecting the vectors $\boldsymbol{u}_i = \boldsymbol{X}\boldsymbol{v}_i / \boldsymbol{\sigma}_i$, $\boldsymbol{\Sigma}$ in an $n$ x $m$ rectangular diagonal matrix with nonnegative diagonal entries $\boldsymbol{\sigma}_i = \sqrt{\boldsymbol{\lambda}_i}$ known as singular values and $\boldsymbol{V}$ is an $m$ x $m$ orthogonal matrix whose columns are the eigenvectors of the matrix $\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X}$, with associated eigenvalues $\boldsymbol{\lambda}_i$. Since $\boldsymbol{V}$ is orthogonal, Eq. 2 can be easily rewritten as

$$\boldsymbol{XV} = \boldsymbol{U}\Sigma \tag{3}$$

and from this statement, it is straightforward to see that the transformation matrix $\boldsymbol{S}$ we are looking for is exactly the matrix $\boldsymbol{V}$ collecting the eigenvectors of $\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X}$. As requested by the algorithm, in fact, $\boldsymbol{V}$ is orthogonal and the new data expressed as $\boldsymbol{U}\Sigma$ have a diagonal covariance matrix (since $\boldsymbol{U}$ is orthogonal and $\boldsymbol{\Sigma}$ is diagonal). Thus, being $\boldsymbol{S}$=$\boldsymbol{V}$, the transformed data collected in the $n$ x $m$ matrix $\boldsymbol{Y}$ can be written as

$$\boldsymbol{Y} = \boldsymbol{XV} \tag{4}$$

whose columns are the so-called PCs of the original data $\boldsymbol{X}$.

The main advantages of data decomposition through PCA is that PCs capture the maximum variability among the columns of $\boldsymbol{X}$ [10] and are sorted in decreasing order in terms of explained variance of the original data (i.e., the first greatest variance is explained by the first PC, the second greatest variance by the second PC, and so on); furthermore, PCs are uncorrelated and, thus, can be considered separately one from another. Typically, through PCA, $p$ PCs (usually $p$=1,2) are selected and data dimensionality can be reduced from $m$ (in our case $m$=25, the number of GV indices) to $p$<<$m$, with minimum loss of information.

Regarding specifically the PCs, they are linear combinations of all the $m$ variables with nonzero coefficients collected along the columns of the matrix $\boldsymbol{V}$ (coefficients are hereafter called loadings in agreement with [10]). As it appears from Eq. 4, in fact, the $i$-th column of $\boldsymbol{V}$ contains the loadings of the $m$ variables allowing to obtain the $i$-th PC. Denoting now $\boldsymbol{V} = \boldsymbol{B}$, the problem of searching for the (unknown) loading matrix $\boldsymbol{B}$ can be seen as a regression problem. In particular, the $i$-th column of $\boldsymbol{Y}$ can be written as

$$\boldsymbol{y}_i = \boldsymbol{X}\boldsymbol{\beta} \tag{5}$$

where $\boldsymbol{\beta}$ is a column vector of $\boldsymbol{B}$ collecting the loadings that allow transforming the original data into the $i$-th PC. Thus, for each selected PC, we are facing the problem of the optimum estimation of vector $\boldsymbol{\beta}$ from data $\boldsymbol{y}_i$.

#### *(b) LASSO estimation of sparse loadings*

PCA allows reducing data dimensionality through selection of uncorrelated PCs. As shown above, however, each PC is a linear combination of all the original variables with coefficients that are typically nonzero. The aim of the LASSO estimation is to reduce also the number of explicitly used variables, through the calculation of sparse loadings. Defining $\hat{\boldsymbol{\beta}}^{\boldsymbol{LASSO}}$ the estimated vector of sparse coefficients, it is obtained from the solution of the following optimization problem

$$\hat{\boldsymbol{\beta}}^{\boldsymbol{LASSO}} = \left( \left\| \boldsymbol{y}_i - \sum_{j=1}^{m} \boldsymbol{X}_j \boldsymbol{\beta}_j \right\|^2 + \boldsymbol{\lambda} \sum_{j=1}^{m} \left| \boldsymbol{\beta}_j \right| \right) \tag{6}$$

where $\boldsymbol{X}_j$ is the $j$-th column of the data matrix $\boldsymbol{X}$ collecting the $n$ observations of the same variable and $\boldsymbol{\lambda} \geq 0$ is a complexity parameter related to the number of loadings that will be shrunk to zero. As one can see from Eq. 6, the cost function is made up of two different terms: the first one is the residual sum of squares that, alone, would lead to a full estimated loading vector; the second one is the sum of coefficients absolute values, which is aimed to shrink some loadings exactly to zero, if $\boldsymbol{\lambda}$ is large enough. The LASSO estimation, thus, allows us to select a small number of variables from a larger pool; the ensemble of the selected variables is able to preserve a great part of the variance originally explained by the whole set of measures.

## III. RESULTS

### A. *Parsimonious set of GV indices in T1D*

After the application of PCA to the data matrix, the number of selected PCs was determined so that at least the 85% of the total original variance was saved. This constraint led us to select the first 2 (out of 25) PCs. In fact, as can be seen from the plot of the percentage explained variance as a function of the number of selected PCs reported in the left panel of Fig. 1, this number of PCs allowed going beyond the defined threshold (red horizontal line in the figure) and saving the 87% of the total variance originally present in the data.

Then, since PCs were obtained as linear combinations of all GV indices, the LASSO estimation of sparse loadings was computed for each selected PC, to reduce the number of variables contributing to the regression. In particular, the maximum number of nonzero coefficients was set to 5 and the GV indices selected by SPCA were those corresponding to nonzero loadings. The number of nonzero coefficients, and thus of selected GV metrics for each PC, was determined as the smallest number of variables that allowed explaining at least the 75% of the initial variance in both datasets. From the inspection of right panels of Figs. 1 and 2, that show the percentage explained variance plotted against the number of selected GV indices per PC for T1D and T2D, respectively, it can be seen that the minimum number of metrics that allowed preserving at least the 75% of the original variance (red horizontal lines in the figures) in both cases was equal to 5. Thus, 5 was the chosen number of nonzero loadings per PC.

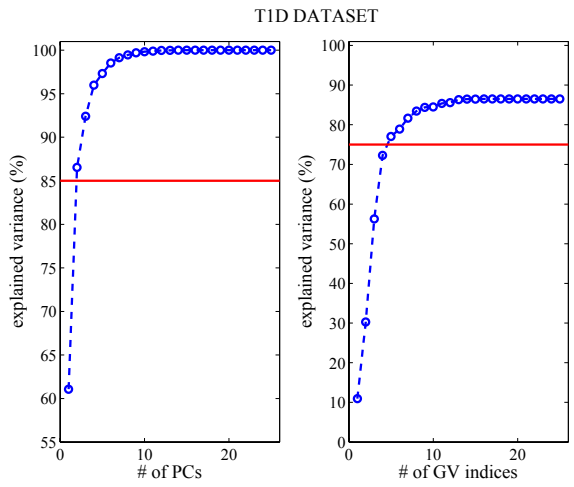


Figure 1. Dependence of SPCA results from the parameter values in T1D. In particular, the explained variance (%) is plotted as a function of the number of selected PCs (left) and, for the chosen number of PCs (i.e., 2), as a function of the number of selected GV metrics (right). Thresholds are shown in red and are set to 85% for the left panel and to 75% for right one.

For the T1D dataset, with 5 (out of 25) GV indices for each of the 2 (out of 25) PCs, SPCA finally allowed explaining the 77% of the variance originally explained by the whole pool of the considered indices. The selected metrics, summarized in Table I for each PC, are thus sufficient for a parsimonious but still comprehensive characterization of GV in the considered population of 17 T1D subjects.

TABLE I. RESULTS FROM SPCA – T1D

| **SELECTED GV INDICES** (EXPLAINED VARIANCE: 77%) | |
|---|---|
| **PC #1** | **PC #2** |
| J-index | %CV |
| MAGE | range |
| ADRR | LBGI |
| IGC | ADRR |
| %GRADEeu | Hypoglycemic Index |

### B. *Parsimonious set of GV indices in T2D*

The choice of SPCA parameters was developed as described for the T1D dataset.

In particular, the number of considered PCs resulted equal to 2 also in this case, saving the 88% of the original variance (Fig. 2, left panel), and 5 GV indices were selected for each PC.

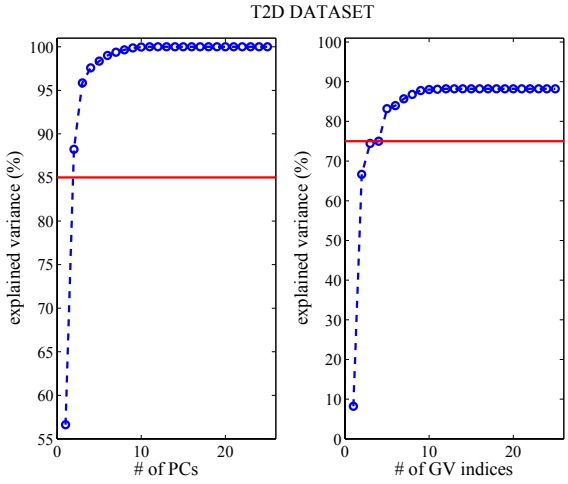


Figure 2. Dependence of SPCA results from the parameter values in T2D. In particular, the explained variance (%) is plotted as a function of the number of selected PCs (left) and, for the chosen number of PCs (i.e., 2), as a function of the number of selected GV metrics (right). Thresholds are shown in red and are set to 85% for the left panel and to 75% for right one.

For the T2D dataset, with 5 (out of 25) selected GV indices for each of the 2 (out of 25) PCs, SPCA allowed explaining the 83% of the variance originally explained by all the considered metrics. Results obtained from the T2D dataset are summarized in Table II.

TABLE II. RESULTS FROM SPCA – T2D

| **SELECTED GV INDICES** (EXPLAINED VARIANCE: 83%) | |
|---|---|
| **PC #1** | **PC #2** |
| J-index | %CV |
| ADRR | 50$^{th}$ percentile |
| BGRI | %values below target |
| %GRADEeu | LBGI |
| %GRADEhyper | Hypoglycemic Index |

Remarkably, it can be seen from the comparison of Tables I and II that 6 out of 10 selected GV indices, i.e., J-

index, ADRR, %GRADEeu, %CV, LBGI and Hypoglycemic Index, are shared by T1D and T2D.

## IV. Conclusion

GV is a risk factor for the development of diabetes complications, and the indices available for its quantification are redundant in terms of conveyed information. In this work, we tested the possibility of using SPCA as a technique to determine, from a wide pool of GV indices, a smaller subset of metrics able to explain a large part of the total original variance present in the data. In particular, SPCA was applied to a pool of 25 well-established literature GV indices evaluated on 17 T1D and 13 T2D CGM time-series. Results show that SPCA allowed selecting 10 (out of 25) GV indices, with more than the 75% of the original variance saved in both datasets, thus coming out to be a valuable tool to provide a parsimonious but still comprehensive description of GV in diabetes. From our results, it can be also seen that 6 indices of the parsimonious set were shared by T1D and T2D, seeming to be independent from the specific dataset. Though interesting, this second result is still preliminary and the shared pool of selected indices cannot be interpreted as a "gold standard" combination of metrics to condense GV information. Larger datasets and a higher number of metrics for GV quantification have to be considered to strengthen and validate the results, and to robustly identify a subset of indices well-representative of GV in diabetes, regardless from the specific dataset.

## Acknowledgment

CGM data have been kindly provided by Alberto Maran, (University of Padova, Padova, Italy) for T1D, and by Alejandra Guillén (Medtronic Iberica, Madrid, Spain) and Giuseppe Fico (Technical University of Madrid, Madrid, Spain) for T2D.

Part of the present research has been developed under the aegis of "MOSAIC" (http://www.mosaicproject.eu/), a project funded by EU within the 7th Framework Program (grant agreement FP7 - 600914).

## References

[1] www.idf.org
[2] D. Rodbard, "Glycemic variability: measurement and utility in clinical medicine and research - One viewpoint", *Diabetes Technol Ther*, vol. 13, pp. 1077-1080, 2011.
[3] J.H. DeVries, "Glucose variability: where it is important and how to measure it," *Diabetes*, vol. 62, pp. 1405-1408, 2013.
[4] B. Kovatchev, D. Cox, L. Gonder-Frederick, D. Young-Hyman, D. Schlundt, W. Clarke, "Assessment of risk of severe hypoglycemia among adults with IDDM – Validation of the low blood glucose index," *Diabetes Care*, vol. 21, pp. 1870-1875, 1998.
[5] N.R. Hill, P.C. Hindmarsh, R.J. Stevens, I.M. Stratton, J.C. Levy, D.R. Matthews, "A method for assessing quality of control from glucose profiles," *Diabet Med*, vol. 24, pp. 753-758, 2007.
[6] B. Kovatchev, M. Straume, D. Cox, R.S. Farhy, "Risk analysis of blood glucose data: A quantitative approach to optimizing the control of insulin dependent diabetes," *J Theor Med*, vol. 3, pp. 1-10, 2000.
[7] W. Clarke, B. Kovatchev, "Statistical tools to analyze continuous glucose monitor data," *Diabetes Technol Ther*, vol. 11, pp. S45-S54, 2009.
[8] D. Rodbard, "New and improved methods to characterize glycemic variability using continuous glucose monitoring," *Diabetes Technol Ther*, vol. 11, pp. 551-565, 2009.
[9] D. Rodbard, "Interpretation of continuous glucose monitoring data: glycemic variability and quality of glycemic control," *Diabetes Technol Ther*, vol. 11, pp. S55-S76, 2009.
[10] D. Rodbard, T. Bailey, L. Jovanovic, H. Zisser, R. Kaplan, S.K. Garg, "Improved quality of glycemic control and reduced glycemic variability with use of continuous glucose monitoring," *Diabetes Technol Ther*, vol. 11, pp. 717-723, 2009.
[11] H. Zou, T. Hastie, R. Tibshirani, "Sparse Principal Component Analysis," *J Comput Graph Stat*, vol. 15, pp. 265-286, 2006.
[12] C. Fabris, A. Facchinetti, G. Sparacino, M. Zanon, S. Guerra, A. Maran, and C. Cobelli, "Glucose variability indices in type 1 diabetes: parsimonious set of indices revealed by Sparse Principal Component Analysis," *Diabetes Technol Ther*, vol. 16, 2014.