Knowledge-driven dictionaries for sparse representation of continuous glucose monitoring signals

Niraj Goel, Theodora Chaspari, Bobak J. Mortazavi, Temiloluwa Prioleau,
Ashutosh Sabharwal, and Ricardo Gutierrez-Osuna

Abstract— **Continuous glucose monitoring (CGM) of patients with diabetes allows the effective management of the disease and reduces the risk of hypoglycemic or hyperglycemic episodes. Towards this goal, the development of reliable CGM models is essential for representing the corresponding signals and interpreting them with respect to factors and outcomes of interest. We propose a sparse decomposition model to approximate CGM time-series as a linear combination of a small set of exemplar atoms, appropriately designed through parametric functions to capture the main fluctuations of the CGM signal. Sparse decomposition is performed through the orthogonal matching pursuit (OMP). Results indicate that the proposed model provides up to 0.1 relative reconstruction error with 0.8 compression rate on a publicly available dataset containing 25 patients diagnosed with Type 1 diabetes. The atoms selected from the OMP procedure can be further interpreted in relation to the clinically meaningful components of the CGM signal (e.g. glucose spikes, hypoglycemic episodes, etc.).**

I. Introduction

Type 1 diabetes is a chronic condition related to the ability of the pancreas to adjust to normal insulin levels [1]. The immune system of patients suffering from this disease tends to destroy the beta cells present in part of the pancreas, leading to the inability to produce insulin and therefore resulting in abnormal glucose levels. According to recent statistics, Type 1 diabetes affects 9.4% of the US population with an overall cost of $245 billion [2]. Despite the tremendous improvements in treating methods, the mortality rate of patients suffering from this disease still remains high, ranging between three to seven times larger compared to the general population [3]. Furthermore, patients with Type 1 diabetes require a life-long therapy with multiple daily insulin injections adjusted on the basis of their glucose levels, rendering the long-term diabetes management laborious and expensive.

Continuous glucose monitoring (CGM) systems provide time-stamped series of glycemic data that are being continuously recorded in every-day life. While such systems have the potential to assist towards individualized glycemic goals [4], [5], the resulting CGM time-series contain multiple sources of noise related to sensor calibration and delay issues, as well as needle drifts. These challenges render signal processing steps essential in order to identify the important signal components and appropriately interpret the underlying information. CGM time-series depict a characteristic structure over time, since the corresponding signal increases abruptly after food intake and slowly recovers. We take advantage of this characteristic shape by representing the CGM signal through a linear combination of modular components that capture general glucose levels as well as glucose spikes.

Sparse representation techniques allow to model a signal as a linear combination of a small number of exemplar signals (called “atoms”) selected from an over-complete set of such sub-signals (called “dictionary”). These techniques can effectively model the inherent variability of biomedical time-series, such as glucose, and allow accurate and scalable representations of scientific and translational value through the use of appropriately selected dictionaries [6], [7]. We take advantage of the typical shape of CGM signals and design dictionaries that take into account smooth and abrupt variations as well as general trends. The design of the aforementioned CGM-specific dictionary promotes the interpretability of our models, since the dictionary atoms correspond to meaningful structural characteristics of the CGM signal (e.g. glucose spikes) (Section III-A). The CGM signal is further decomposed into a small set of dictionary atoms using the orthogonal matching pursuit (OMP) (Section III-B). Results on publicly available data obtained from the Diabetes Research In Children Network (DirecNet) Glucagon Study [8] (Section IV) indicate that the proposed approach can yield up to 0.1 relative reconstruction error with 0.8 compression rate for a 10 hour CGM signal (Section V). Plots of the CGM time-series and the selected dictionary atoms are presented and discussed in relation to the interpretability of the proposed framework.

II. Previous Work

Previous research has examined glucose variability metrics captured through signal statistics (e.g. interquartile range, area under the curve) to provide measures of glycemic variation and assess hypoglycemic or hyperglycemic excursions [9], [10], [11]. Other studies have proposed frequency analysis of the CGM signal through Fourier and wavelet decomposition [12], [13], [14]. Neural networks have recently been used to predict abnormal glucose levels from raw CGM data [15], [16]. Finally, a variety of model identification techniques have been proposed to model CGM time-series though a system of differential equations that

*This work was supported, in part, by funding from the National Science Foundation Engineering Research Center for Precise Advanced Technologies and Health Systems for Underserved Populations (PATHS-UP) (Award #1648451).

N. Goel, T. Chaspari, B. Mortazavi, and R. Gutierrez-Osuna are with the Texas A&M University, College Station, TX, 77843 (email: {nirajgoel;chaspari;bobakm;rgutier}@tamu.edu)

T. Prioleau and A. Sabharwal are with the Rice University, Houston, TX, 77005 (email: {to9;ashu}@rice.edu).

captures the dynamics between glucose levels and insulin intake [17], [18]. Despite their encouraging results, some of the limitations of the previous studies include their lack of interpretability [15], [16] or their limited ability to represent fine-grain signal fluctuations that might be meaningful in clinical settings [14], [18]. Given these limitations, we propose a knowledge-driven framework for representing the CGM time-series through sparse approximation techniques and the design of interpretable dictionaries, able to reliably reconstruct the signal and provide meaningful information about the glucose fluctuations.

III. Sparse Representation of Continuous Glucose Monitoring (CGM)

We present the proposed sparse representation framework for modeling a CGM signal.

A. Dictionary Design

The dictionary consists of two types of parameterized CGM-specific atoms that capture the main components of CGM signals in a knowledge-driven way. The first type represents the general glucose levels through straight lines, while the second captures the glucose fluctuations (e.g. glucose spikes) expressed as Bateman functions (Fig. 1a). Straight lines are expressed as $\phi_1(t) = \Delta_0 + \Delta t$, where Δ_0 and Δ are the offset and the slope, respectively, while Bateman functions are written as:

$$\phi_2(t) = \left(e^{-b(st-t_0)} - e^{-a(st-t_0)}\right) u(t-t_0), a < b \tag{1}$$

where $u(t)$ is the step function centered at t_0, $a \in \{.2, .4. \ldots, 2\}$ and $b \in \{.4, .8. \ldots, 2\}$ are parameters related to the steepness of recovery and onset of a CGM fluctuation, $s \in \{.6, .12. \ldots, .60\}$ is the time scale, and $t_0 \in \{0, 2, 4, \ldots, N\}$ captures the time-shift of an atom within an analysis frame of length M. We further included time-reversed Bateman atoms, i.e. $\phi_3 = \phi_2(M-t)$ to capture multiple shapes of CGM fluctuations (Fig. 1b). The parametric nature of the dictionaries yields from the use of a specific functional form (e.g. Bateman) for representing the dictionary atoms. By varying the parameters of this function, we are able to obtain high variability in the shape of the corresponding atoms, which promotes the ability of our model to capture variable CGM shapes.

B. Signal Approximation

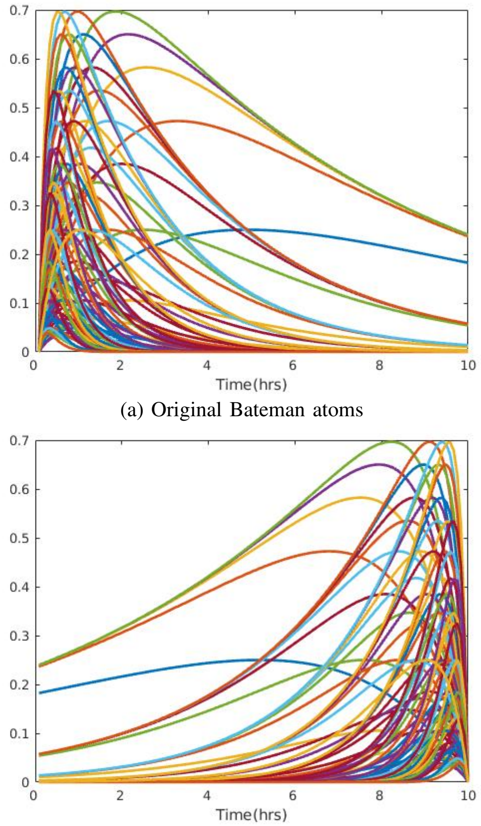
We approximate an input signal $f(t)$ as a weighted sum of finite number of atoms selected from the aforementioned CGM-specific dictionary, as follows:

$$f(t) \approx \hat{f_N}(t) := \sum_{n=1}^{N} a_n g_n(t) \tag{2}$$

where g_n represents the selected atoms and a_n the corresponding scalar coefficients. The problem of selecting a small number of atoms from an overcomplete dictionary is NP-hard and hence several approximation algorithms can be used to achieve a suboptimal solution. For our purposes, we used the Orthogonal Matching Pursuit (OMP) because of its low computational cost and theoretical guarantees of correctness [19]. OMP is an iterative greedy approximation algorithm. At each iteration n, OMP selects the atom which has the maximum correlation with the current signal residual $R^n\mathbf{f}$. The maximum absolute correlation value is not used as a criterion, because this would result in negative atom coefficients negatively impacting the interpretability of the model. The residual gets updated by applying the following orthogonal projection operator which includes all the atoms that have been selected until the current iteration n:

$$R^n\mathbf{f} = \mathbf{f} - \mathbf{f_n} \;\; , \;\; \mathbf{f_n} = \mathbf{G_n}(\mathbf{G_n}^T * \mathbf{G_n})^{-1}\mathbf{G_n}^T\mathbf{f} \tag{3}$$

where $\mathbf{G_n}$ is the matrix of atoms selected until the n^{th} iteration, $\mathbf{f}$ is the discretized version of the continuous input signal $f(t)$, and $\mathbf{f_n}$ is the approximation of the input signal after n iterations of the OMP.



(a) Original Bateman atoms

(b) Time-reversed Bateman atoms

Fig. 1. Example of Bateman atoms used to capture continuous glucose monitoring (CGM) fluctuactions.

C. Evaluation criteria

We evaluate the ability of the proposed model to reliably represent the signals of interest. First we report the relative RMS error, defined as:

$$RelErr = \frac{1}{K}\sum_{k=1}^{K}\sqrt{\frac{\|\mathbf{f}^{(k)} - \mathbf{f}_{\mathbf{N}}^{(k)}\|_2}{\|\mathbf{f}^{(k)}\|_2}} \tag{4}$$

where $\mathbf{f}^{(k)}$ denotes the k^{th} signal frame, $\mathbf{f}_{\mathbf{N}}^{(k)}$ denotes the approximation of the k^{th} frame using N OMP iterations, and $\| \cdot \|$ is the $l2$-norm. Low relative RMS error values reflect high-quality signal representation.

Assuming prior knowledge of the parametric functions of the dictionary (Section III-A), each atom can be represented through an 6-dimensional vector of parameters (each encoded using 32-bits), where the first element includes the atom coefficient, while the remaining five capture the parameters of the selected atom. If a straight line is selected, these include the offset Δ_0, slope Δ, and three zero-elements, whereas if a Bateman atom is selected, these include the time shift t_0, time scale s, steepness of recovery and onset a, b, and whether the corresponding Bateman atom is reversed in time. Distinction between the type of selected atom (i.e. straight line or Bateman) is performed based on whether the last three elements of the corresponding vector are zero, noting that this can never occur for the Bateman atoms based on the way they were designed. Taking these into account, the compression rate can be calculated as follows:

$$CR = \frac{(M - 6 \cdot N)}{M} \tag{5}$$

where M is the analysis frame length in samples and N is the number of selected atoms. High CR values reflect the model's ability to efficiently encode the underlying data.

IV. Data Description and Pre-Processing

Our data come from the publicly available Glucagon Study of the DirecNet repository [8]. The dataset contains 25 participants (8-19 years old) with Type 1 diabetes wearing a continuous glucose monitor for 6-7 days at a sampling rate of $F_s = 0.0033$Hz. At the end of this period, patients re-visited the lab, where they performed a hypoglycemic clamp test in order to assess plasma glucagon responses to hypoglycemia. Data from each participant were segmented per test day, resulting in 328 total segments. The missing samples from the beginning and the end of the recordings were removed, while the rest were interpolated using linear interpolation.

V. Results

Our experiments are performed with various analysis window lengths, i.e. 4, 6, 8, 10, 12 hours (hrs), corresponding to 48, 72, 96, 120, 144 samples, respectively. The window shift was empirically set to one third of the analysis frame. As expected, reliable reconstruction is achieved using short analysis windows and a large number of OMP iterations, while the opposite holds for compression rate (Fig. 2, Table I). The error curve depicts a steep decrease in the first four iterations and becomes smoother afterwards (Fig. 2), suggesting that including a very large number of atoms might be not be as beneficial: 4 iterations can represent 10 hrs of CGM data with .8 compression rate and 0.1 reconstruction error. Selection of the appropriate window size and number of OMP iterations is tightly associated with the specific application, however, our results indicate that reliable representations can be achieved using 8-10 hrs of CGM signal with 4-6 OMP iterations.

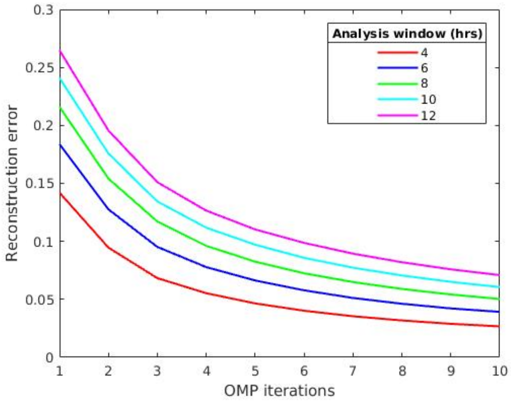


Fig. 2. Relative reconstruction error with varying number of orthogonal matching pursuit (OMP) iterations.

TABLE I

Signal representation results for different analysis window lengths and orthogonal matching pursuit (OMP) iterations.

window size(hrs)	# OMP Iterations (N)	Reconstruction error ($RelErr$)	Compression rate (CR)
4	2	.0946	.750
4	4	.0553	.500
6	2	.1276	.833
6	4	.0778	.666
8	4	.0961	.750
8	6	.0725	.625
10	4	.1118	.800
10	6	.0858	.700
10	8	.0706	.600
12	4	.1265	.833
12	6	.0987	.750
12	8	.0820	.666

Examples of CGM time-series along with the selected Bateman atoms and the reconstructed signals that result from the linear combination of these atoms are presented in Fig. 3. Through visual inspection we understand that Bateman atoms of different parameters and amplitudes can be used to represent abrupt CGM spikes (Fig. 3a), as well as smoother glucose fluctuations (Fig. 3b). By knowing a priori the parameters of the selected atoms, we are able to infer the shape of the corresponding fluctuation, which has been related to several factors of interest, such as nutrient intake [20]. We further note that the proposed models can also capture decreasing glucose trends, such as the ones observed after a hypoglycemic clamp test (Fig. 3c). Despite these encouraging results, our models fail capturing fluctuations that occur right after an abrupt signal drop, which is mostly due to the fact that negative coefficients are not allowed during the OMP decomposition (Section III-B).

VI. Conclusions and Future work

We proposed a sparse representation framework for modeling CGM time-series through the use of appropriately designed dictionaries, that capture general levels and fluctuations of glucose signals. Results obtained through signal reconstruction and compression rate indicate that the proposed model achieves reliable signal representation, while visual inspection highlights the interpretability of the selected atoms in relation to the underlying signal components.

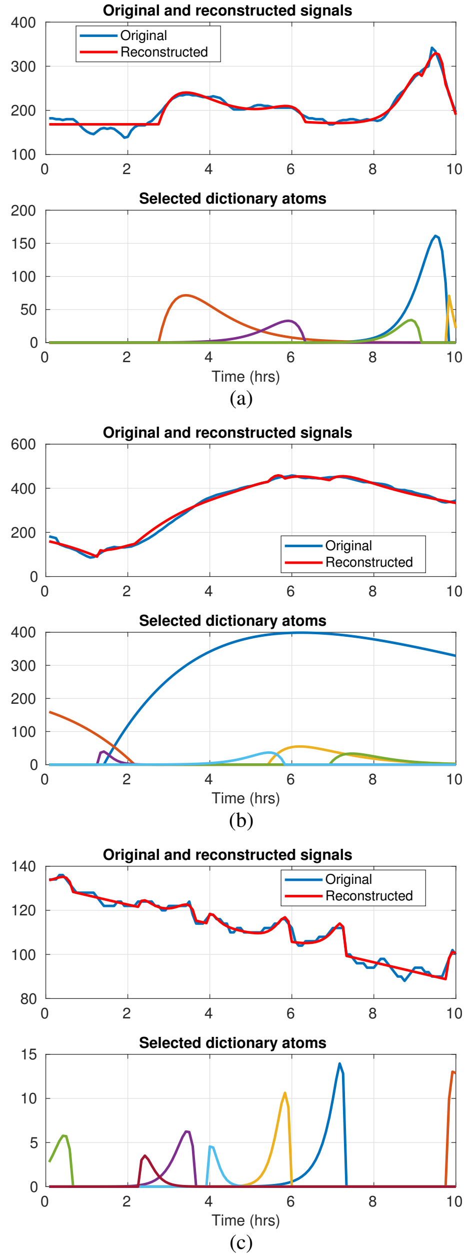


Fig. 3. Example of original and reconstructed continuous glucose monitoring (CGM) time-series and the selected Bateman atoms.

As part of our future work, we plan to quantitatively assess the interpretability of our approach through clinically-relevant factors and outcomes (e.g. prediction of abnormal glucose levels, meal composition). We further plan to examine the feasibility of our framework with different types of dictionary atoms, as well as to explore automated dictionary learning techniques.

References

[1] Bart O Roep, "The role of T-cells in the pathogenesis of Type 1 diabetes: from cause to cure," *Diabetologia*, 46(3):305–321, 2003.

[2] Centers for Disease Control, Prevention, et al., "National diabetes statistics report, 2017," *Atlanta, GA: Centers for Disease Control and Prevention, US Department of Health and Human Services*, 2017.

[3] Aaron M Secrest, Dorothy J Becker, Sheryl F Kelsey, Ronald E LaPorte, and Trevor J Orchard, "All-cause mortality trends in a large population-based cohort with long-standing childhood-onset Type 1 diabetes: the Allegheny County Type 1 diabetes registry," *Diabetes care*, 33(12):2573–2579, 2010.

[4] Janet Silverstein, Georgeanna Klingensmith, Kenneth Copeland, Leslie Plotnick, Francine Kaufman, Lori Laffel, Larry Deeb, Margaret Grey, Barbara Anderson, Lea Ann Holzmeister, et al., "Care of children and adolescents with Type 1 diabetes: a statement of the American Diabetes Association," *Diabetes care*, 28(1):186–212, 2005.

[5] Jan Šoupal, Lenka Petruželková, Milan Flekač, Tomáš Pelcl, Martin Matoulek, Martina Daňková, Jan Škrha, Štěpán Svačina, and Martin Prázný̀, "Comparison of different treatment modalities for Type 1 diabetes, including sensor-augmented insulin regimens, in 52 weeks of follow-up: a COMISAIR study," *Diabetes technology & therapeutics*, 18(9):532–538, 2016.

[6] Theodora Chaspari, Andreas Tsiartas, Leah I Stein, Sharon A Cermak, and Shrikanth S Narayanan, "Sparse representation of electrodermal activity with knowledge-driven dictionaries," *IEEE Transactions on Biomedical Engineering*, 62(3):960–971, 2015.

[7] Alberto Greco, Gaetano Valenza, Antonio Lanata, Enzo Pasquale Scilingo, and Luca Citi, "cvxEDA: A convex optimization approach to electrodermal activity processing," *IEEE Transactions on Biomedical Engineering*, 63(4):797–804, 2016.

[8] Jennifer Sherr, Dongyuan Xing, Katrina J Ruedy, Roy W Beck, Craig Kollman, Bruce Buckingham, Neil H White, Larry Fox, Eva Tsalikian, Stuart Weinzimer, et al., "Lack of association between residual insulin production and glucagon response to hypoglycemia in youth with short duration of Type 1 diabetes," *Diabetes Care*, 36(6):1470–1476, 2013.

[9] David Rodbard, "Interpretation of continuous glucose monitoring data: glycemic variability and quality of glycemic control," *Diabetes technology & therapeutics*, 11(S1):S55–67, 2009.

[10] Renata A Rawlings, Hang Shi, Lo-Hua Yuan, William Brehm, Rodica Pop-Busui, and Patrick W Nelson, "Translating glucose variability metrics into the clinic via continuous glucose monitoring: A graphical user interface for diabetes evaluation (CGM-GUIDE©)," *Diabetes technology & therapeutics*, 13(12):1241–1248, 2011.

[11] William Clarke and Boris Kovatchev, "Statistical tools to analyze continuous glucose monitor data," *Diabetes technology & therapeutics*, 11(S1)S45-54, 2009.

[12] Michael Miller and Poul Strange, "Use of Fourier models for analysis and interpretation of continuous glucose monitoring glucose profiles," 2007.

[13] Edward Aboufadel, Robert Castellano, and Derek Olson, "Quantification of the variability of continuous glucose monitoring data," *Algorithms*, 4(1):16–27, 2011.

[14] Giuseppe Fico, Liss Hernández, Jorge Cancela, Miguel María Isabel, Andrea Facchinetti, Chiara Fabris, Rafael Gabriel, Claudio Cobelli, and María Teresa Arredondo Waldmeyer, "Exploring the frequency domain of continuous glucose monitoring signals to improve characterization of glucose variability and of diabetic profiles," *Journal of diabetes science and technology*, 11(4):773–779, 2017.

[15] Carmen Pérez-Gandía, A Facchinetti, G Sparacino, C Cobelli, EJ Gómez, M Rigla, Alberto de Leiva, and ME Hernando, "Artificial neural network algorithm for online glucose prediction from continuous glucose monitoring," *Diabetes technology & therapeutics*, 12(1):81–88, 2010.

[16] Scott M Pappada, Brent D Cameron, Paul M Rosman, Raymond E Bourey, Thomas J Papadimos, William Olorunto, and Marilyn J Borst, "Neural network-based real-time prediction of glucose in patients with insulin-dependent diabetes," *Diabetes technology & therapeutics*, 13(2):135–141, 2011.

[17] Dimitri Boiroux, Morten Hagdrup, Zeinab Mahmoudi, Niels Kjølstad Poulsen, Henrik Madsen, and John Bagterp Jørgensen, "Model identification using continuous glucose monitoring data for Type 1 diabetes," *IFAC-PapersOnLine*, 49(7):759–764, 2016.

[18] Sasikarn Sakulrang, Elvin J Moore, Surattana Sungnul, and Andrea de Gaetano, "A fractional differential equation model for continuous glucose monitoring data," *Advances in Difference Equations*, 2017:150, 2017.

[19] Y.C. Pati, R. Ramin, and P.S. Krishnaprasad, "Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition," *Conference on Signals, Systems and Computers*, 1993.

[20] Mary C Gannon, Frank Q Nuttall, Brian J Neil, and Sydney A Westphal, "The insulin and glucose responses to meals of glucose plus various proteins in Type II diabetic subjects," *Metabolism-Clinical and Experimental*, 37(11):1081–1088, 1988.