Proceedings of the 6th INFORMS Workshop on Data Mining and Health Informatics (DM-HI 2011)
P. Qian, Y. Zhou, C. Rudin, eds.

Using Gaussian Processes to Monitor Diabetes Development

Si-Chi Chin
Information Science
University of Iowa
Iowa City, IA
si-chi-chin@uiowa.edu

W. Nick Street
Management Sciences Department
University of Iowa
Iowa City, IA
nick-street@uiowa.edu

Abstract

This paper uses Gaussian process techniques to model time series data of HbA1c level, a common measure to monitor or screen diabetes. The HbA1c level estimates how well blood sugar is under control. To facilitate the control of diabetes, we develop a patient-level model to individually predict the development of the disease for each patient. Gaussian processes represent a successful machine learning technique known for their flexible modeling abilities and high predictive performances. This approach allows multi-dimensional inputs and assigns a confidence score to the predictions, accounting for temporal uncertainty of time series data. The purpose of this paper is to discuss the use of the Gaussian process technique, previously unseen in diabetes research, to monitor the development of the disease.

Keywords: HbA1c Monitoring, Gaussian Processes Regression, Time Series Analysis

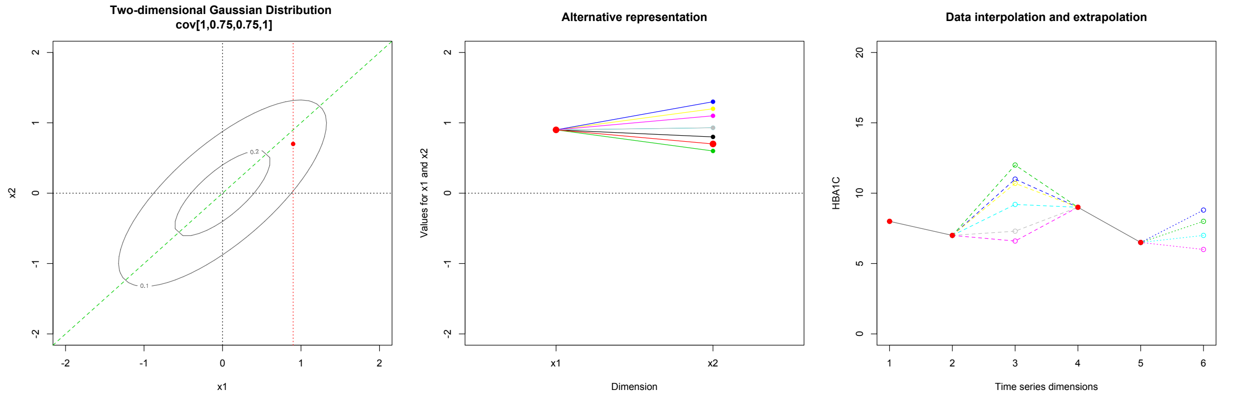
1 Introduction

The hemoglobin A1c test (HbA1c) is an important blood test used to determine how well diabetes is being controlled for patients. The test periodically measures the average amount of sugar in the blood (approximately every 90 days in general). Monitoring the development of HbA1c test results help physicians and patients to manage diabetes and keep the blood sugar level within the target range. Temporal analysis on the trajectory of HbA1c provides insights into how a patient, compared to other patients, controls his or her disease. The temporal analysis also contributes to the study of what causes the HbA1c values rise and fall. In the long term, predictive models such as the ones described here can be used to examine the effects of different treatments on patient sub-populations using retrospective data. Therefore, the paper is inspired by the question: *Given the information of patients and the observed pattern of HbA1c test history, how can we predict the future development of HbA1c?*

In this paper, we propose a Gaussian process regression approach to model the time series for the development of HbA1c. Scholarly work by MacKay [6] and Rasmussen [7] provided detailed discussions on Gaussian process theory. For completeness and convenience, we introduces the basics of Gaussian Process Regression and its application on time course data in Section 2. We describe the experimental design and the dataset in Section 3. Section 4 presents the experimental results and Section 5 concludes the paper.

2 Gaussian Process Regression

A Gaussian process (GP) is a non-parametric stochastic process. Gaussian processes generalize the Gaussian probability distribution, extending multivariate Gaussian distributions to infinite dimensionality (variables) [7]. We can loosely view a vector of infinite dimensionality as a function. A Gaussian process defines a distribution over these functions and the inference takes place directly in function space. As a Gaussian distribution is specified by a mean vector, μ, and covariance matrix Σ, a Gaussian process is specified by a mean function $m(x)$ and covariance function $k(x, x')$. If we denote a Gaussian distribution as: $f \sim \mathcal{N}(\mu, \Sigma)$, a Gaussian process would be $f(x) \sim \mathcal{GP}(m(x), k(x, x'))$. The covariance function (or kernel), $k(x, x')$ controls how one observation relates to another. We chose an RBF kernel for the covariance function.



(a) Two-dimensional Gaussian distribution

(b) Alternative representation

(c) Multi-dimensional representation

Figure 1: Multi-dimensional Gaussian distributions and an alternative representation

In the example of two-dimensional (bivariate) Gaussian distribution shown in Figure 1a, we observed $x_1 = 0.9$. Given the available two-dimensional covariance matrix and the observed x_1, we obtain a distribution for x_2 values (represented in Figure 1b). We can also compute the probability mass for an observation $(x_1, x_2) = (0.9, 0.8)$. It is natural to consider extending the setting to a higher-dimensional space, such as a time series shown in Figure 1c, where y-axis shows the readings of HbA1c and x-axis shows the dimensions in temporal order (e.g. age, sequence of visit, or number of months since the first visit).

Time series, well exemplified by stock prices, are a sequence of data points ordered by uniform successive time intervals. Time series analysis aims to extract meaningful temporal characteristics of the data and to predict data points before they are measured. The HbA1c monitoring data has a natural temporal ordering, providing an appropriate example for time series analysis. It is reasonable to believe that, in the case of HbA1c monitoring, observations close together in time will be more closely related than observations further apart. Likewise, patients who are similar to each other are more likely to develop equivalent patterns in time than patients who are dissimilar. In this paper, we consider the HbA1c test results as being some function over time and use regression models to capture patterns of changes in HbA1c. As shown in Figure 2, the trajectories of HbA1c test vary widely between different patients. In this paper, we used Gaussian process regression to discern patterns from the noisy time series.

Gaussian process regression is motivated by the problem: *given some noisy observations of a dependent variable for some independent variable x, what is the best estimate of the dependent variable at a new value, x^**. The method assumes an underlying process which generates "clean" data. The goal of the method is to estimate the unknown underlying process from a finite number of noisy observed data. Research has demonstrated the value of Gaussian process regression on respiration signal [1], on gene relevance networks [4], and on temporal gene expression data to estimate the expression trajectory [3, 8] and to estimate time shifts [5]. The advantage of Gaussian processes, compared to other learning models, is the ability to capture the relationship between variables with a multi-dimensional covariance matrix. Applying Gaussian processes in clinical studies allows us to understand how the patient demographics relate to their medical conditions, their prescribed medications or procedures, and their lab test results.

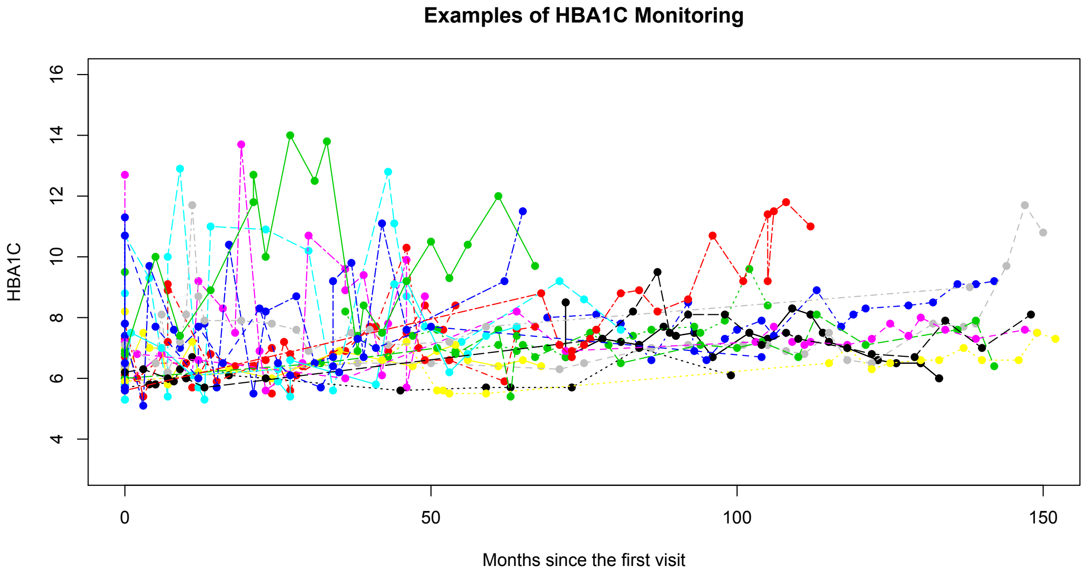


Figure 2: Examples of HbA1c monitoring for more than 10 years. Each line represents the HbA1c trajectory for a patient. The time interval for each observed HbA1c reading, as shown in the graph, varies widely between patients and even for one patient.

3 Experimental Design

We applied Gaussian process (GP) regression and Support Vector Machine (SVM) regression for the analysis of time-series on an obfuscated[1] dataset coming from a cohort of adult outpatients having diagnosis of Diabetes Type I and Type II. The dataset captures the HbA1c readings for more than ten years and minimum personal information for 8,565 patients. To examine the effect of sparse data, we created a subset of five years (60 months) of readings, containing 1,559 patients. The format of the data presents as:

{`age,race,sex,hba1c_t0,hba1c_t1,hba1c_t2,...,hba1c_t60`}.

We used the Weka[2] implementation for both Gaussian Process regression and Support Vector Machine regression for the experiments. The choice of kernel is RBF for both models, using the default parameter settings from Weka. To evaluate the two algorithms, we trained the models using the limited patient demographics and time series from t_1 to t_{x-1} and tested on time t_x, where t_1 is always the time of the patient's first measurement. We performed five-fold cross-validation for the evaluation and used the Mean Absolute Error (MAE) for each time (t_1 ... t_x) and for each patient. To evaluate the performance for each time, we averaged the MAE for all the actual observations tested at the time t_x. To evaluation the performance for each patient, we averaged the MAE from all observed times t_1 to t_x for a given patient.

We also created an interpolated dataset for the subset, in hopes of improving the similarity measure between patients that have HbA1c measurements at different time points. Missing months were filled in using linear interpolation between readings, and with the last reading for times beyond that patient's last visit. We used only the actual observed data to compute the MAE for the evaluation.

[1] We obfuscated the data by randomly swapping values in gender and race attributes and adding a zero-mean Gaussian number to each HbA1c reading.

4 Experimental Results

The experiments compared Gaussian process regression to Support Vector Machine regression. The goal of the experiments is to examine the predictive capability of the suggested approach and to observe the trends of the predictions.

Table 1 and Figure 3 compare the experimental results between GP regression and SVM regression. The GP regression model outperforms SVM regression model for the 5-year interpolated data. Although GP regression has lower performance on the 5-year non-interpolated data, the difference is small. In the setting of more sparse data such as the 10-year data, the GP regression model still outperforms SVM regression model.

Table 1: Comparison between GP and SVM regression, using paired T-test.

Experiments	GP	SMO	P-value
Month view (interp 5 yrs)	1.2156*	1.3334	7.698e-09
Patient view (interp 5 yrs)	1.2175*	1.3332	<2.2e-16
Month view (non-interp 5 yrs)	1.0250	1.0087*	0.01858
Month view (non-interp 10 yrs)	1.0274*	1.0343	0.02121

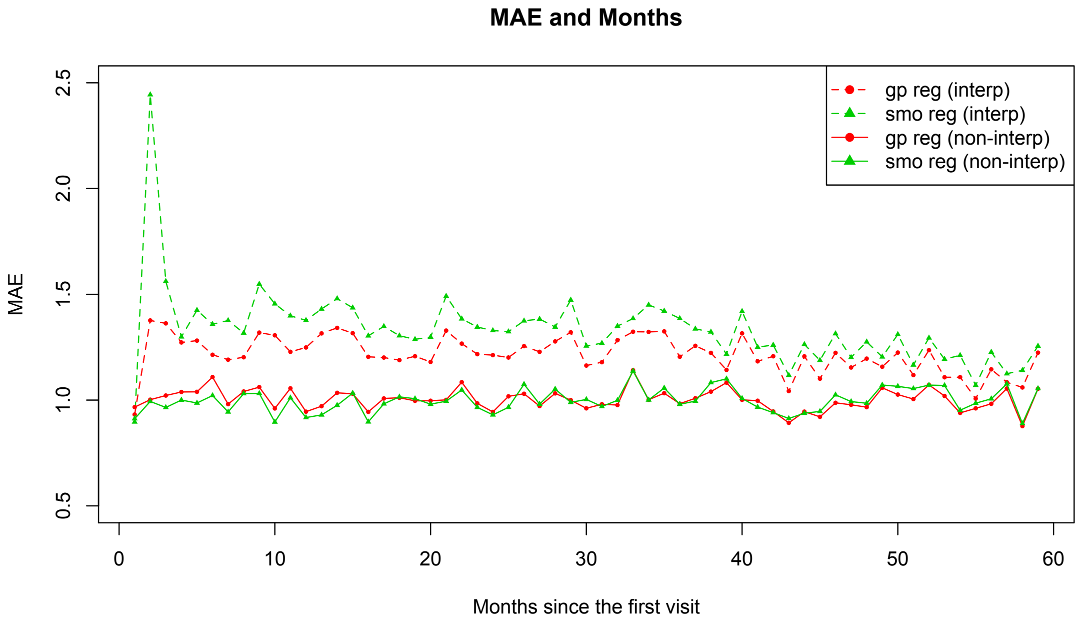
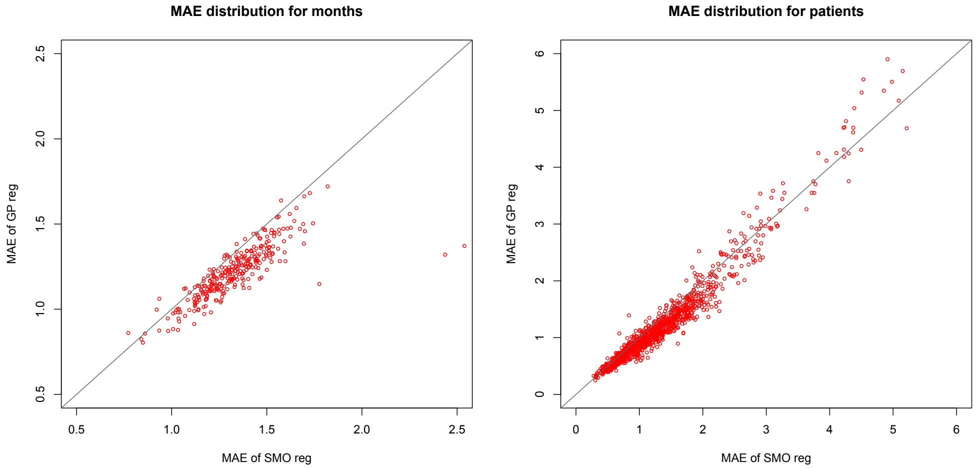


Figure 3: MAE comparison chart at the level of month

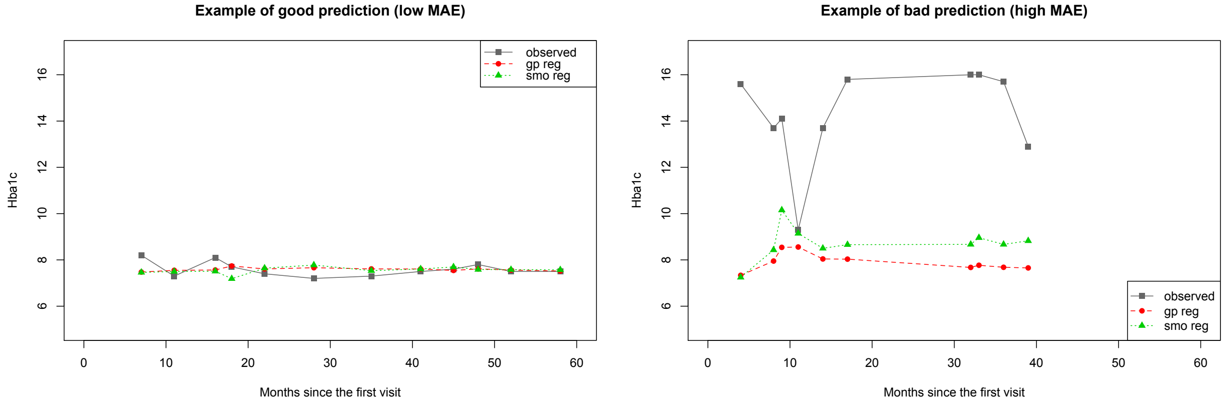
Figure 4 visualizes the performance for the two regression models at the level of months/time (Figure 4a) and at the level of patients (Figure 4b). In both subgraphs, points appearing in the bottom right indicate that the GP regression achieved lower MAE than the SVM regression. For both Figure 4a and Figure 4b, the majority of points fall to the bottom right panel, indicating that GP regression outperforms SVM regression.

Figure 5 presents two examples of predictions at the level of patients. Both models are strong in stable patients but weak for patients with wide fluctuation for their HbA1c tests.



(a) MAE on months (b) MAE on patients

Figure 4: Comparison of GP and SMO regression.



(a) Good prediction (b) Bad prediction

Figure 5: Examples of a good and a bad prediction for two patients.

5 Conclusion

In this paper, we applied Gaussian process regression to model and predict the time series of HbA1c readings for diabetes patients. Gaussian processes, a successful machine learning technique known for their flexible modeling abilities and high predictive performances, captures relationships among multi-dimensional inputs and accounts for the temporal uncertainty of time series data. Our experimental results hold promises for using Gaussian process regression to model HbA1c time series.

Future work will test the approach on a real enhanced dataset, including important additional information such as patient life styles, medication records, comorbidities, and other lab tests. By separating patients by disease type and lining them up so that t_1 represents the time of diagnosis, we can more clearly discover the underlying trends of the noisy observed data to predict the development of the disease. By including treatment information, we can observe - and therefore predict - the effect of different medications on the course of the disease, and thereby examine the expected effect of treatments regimens on individual patients, as well as demographic subgroups.

Acknowledgments

This publication was made possible by Grant Number UL1RR024979 from the National Center for Research Resources (NCRR), a part of the National Institutes of Health (NIH). Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the CTSA or NIH.

References

[1] Sofiane Brahim-Belhouari and Amine Bermak. Gaussian process for nonstationary time series prediction. *Computational Statistics & Data Analysis*, 47(4):705–712, November 2004.

[2] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. The WEKA data mining software: an update. *ACM SIGKDD Explorations Newsletter*, 11:1018, November 2009. ACM ID: 1656278.

[3] Alfredo Kalaitzis and Neil Lawrence. A simple approach to ranking differentially expressed gene expression time courses through gaussian process regression. *BMC Bioinformatics*, 12(1):180, 2011.

[4] Paul D. W. Kirk and Michael P. H. Stumpf. Gaussian process regression bootstrapping: exploring the effects of uncertainty in time course data. *Bioinformatics*, 25(10):1300 –1306, May 2009.

[5] Qiang Liu, Kevin K. Lin, Bogi Andersen, Padhraic Smyth, and Alexander Ihler. Estimating replicate time shifts using gaussian process regression. *Bioinformatics*, 26(6):770 –776, March 2010.

[6] D.J.C. MacKay. Introduction to gaussian processes. In C.M. Bishop, editor, *Neural Networks and Machine Learning, NATO ASI Series*. Springer, Berlin.

[7] Carl Edward Rasmussen. Gaussian processes in machine learning. In Olivier Bousquet, Ulrike Luxburg, and Gunnar Rtsch, editors, *Advanced Lectures on Machine Learning*, volume 3176, pages 63–71. Springer Berlin Heidelberg, Berlin, Heidelberg, 2004.

[8] Ming Yuan. Flexible temporal expression profile modelling using the gaussian process. *Computational Statistics & Data Analysis*, 51(3):1754–1764, December 2006.