ESTIMATION AND SVM CLASSIFICATION OF GLUCOSE-INSULIN MODEL PARAMETERS FROM OGTT DATA. AN AID FOR DIABETES DIAGNOSTICS

MIGUEL ANGEL MORELES, JOAQUIN PEÑA, PAOLA VARGAS, ADRIANA MONROY, AND SILVESTRE ALAVEZ

CONTENTS

ABSTRACT. In the Oral Glucose Tolerance Test (OGTT), a patient, after an overnight fast ingests a load of glucose. Then measurements of glucose concentration are taken every 30 minutes during two hours. The test is used to aid diagnosis of diabetes, namely, type 2 diabetes mellitus and glucose intolerance. Several mathematical models have been introduced to describe the glucose-insulin system during an OGTT. Models consist on systems of differential equations where most parameters are unknown. Estimation of these parameters is an aim of this work. In a minimal model, two of such parameters are proposed for classification by means of a SVM technique. Consequently, a case is made for this classification as an aid for diagnosis.

1. INTRODUCTION

According to the World Health Organization (WHO) global report on diabetes, people living with the disease has increased dramatically during the last four decades. Research on prevention, treatment, diagnostics, etc. is an urgent need. Our modest goal is to delve into the Oral Glucose Tolerance Test (OGTT) as an aid for diagnostics.

1

More precisely, we are concerned with mathematical models of the glucose-insulin regulatory system during the OGTT. These models are in terms of Ordinary Differential Equations (ODE), and are short term in the sense that they describe the dynamics after an external source excites the system. In the OGTT, the external source is a load of glucose administered to the patient after an overnight fast. Then glucose concentration is measured every 30 minutes.

In practice, an ODE model is chosen and OGTT data is used to estimate parameters in the model. This approach is of great interest, given the possibility to use the parameters of the models as indicators of diabetes development.

Research along these lines is vast. A recent review of mathematical modeling of the glucose-insulin system is presented in Palumbo et al [6]. For a more computational, parameter estimation oriented review, see Rathee & Nilam [8]. These works provide an extensive bibliography, as well as the underlying biological mechanisms of the glucose-insulin system. We shall be concise on the latter.

A natural application of parameter estimation is to validate a model by data fitting. Our motivation is to take an step further and explore good fitting parameters for classification. In the context of the glucose-insulin system we start with actual data gathered at the Mexico General Hospital. In total, 80 female patients underwent the OGTT. 51 healthy, 4 with Impaired Fasting Glycaemia (IFG), 15 with Impaired Glucose Tolerance (IGT), 7 with both alterations and 3 with Diabetes Mellitus Type 2 (T2DM).

Models for the glucose-insulin system can be quite complex, involving several parameters in a system of ordinary differential equations. Consequently, second order minimal models, e.g. Caumo, Bergman & Cobelli [2], are preferred.

For a first test of the methodology to be presented, we use the minimal model of Ackerman et al [1]. Hereafter referred as the basic model. This model is probably the simplest. We find the obtained results highly satisfactory, which have led us to report our findings.

The basic model can be reduced to a single differential equation corresponding to a harmonic oscillator. The state variable, is the the deviation from the stable fasting level of the glucose concentration. To solve the parameter estimation problem, we follow the bayesian approach. The MAP estimator is chosen from the posterior. The handling of noisy data and assessment of the quality of point estimators, is straightforward in bayesian estimation. These features have made bayesian estimation a natural choice. A strong case is made in Pillonetto, Sparacino & Cobelli [7].

For each patient, four parameters are estimated. Amplitude and damping coefficient are used for classification. A Support Vector Machine technique is applied to the 80 pairs of parameters. The classification is correct for 85% of the cases, and more importantly, the separating line appears as a transition line from healthy to mildly ill to diabetic. The application in mind is early detection of patients at risk.

Let us describe the content of this work. In Section 2 we discuss minimal models, introduce the basic model and its solution in the homogeneous damped oscillator case. Then the parameter identification problem of interest is posed. Main results are presented in Section 3. A sample of fitting curves for patients in various conditions and a SVM classification plot are shown. It is argued that one may build on this SVM plot to develop an aid for early detection of pre diabetic patients. We leave for Section 4 the bayesian estimation. We focus on the damping coefficient, which was regarded unreliable in the past. In contrast, marginal posterior densities, show that the MAP estimator of this parameter is somewhat robust. We close with a section with conclusions and future work.

2. The parameter estimation problem given OGTT data

2.1. Minimal glucose-insulin models during an Oral Glucose Tolerance Test.
The OGTT starts with an overnight fast. On arrival to the lab, a glucose load of $75g$ is orally administered to the patient. Then glucose concentrations are measures at time 0, 30, 60, 90, and 120 minutes.

Let $G(t)$ be the glucose concentration in the blood and $H(t)$ be the net concentration of a variety of hormones that influence the blood glucose levels. For OGTT conditions, insulin is considered predominant and $H(t)$ is essentially its concentration.

Minimal models consist of second order ODE systems describing the kinetics of glucose concentration and insulin action, namely

(1)
$$\begin{aligned}
\frac{dG}{dt} &= P(G, H, \mathbf{u}) + J(t) \\
\frac{dH}{dt} &= Q(G, H, \mathbf{u}),
\end{aligned}$$

Here $\mathbf{u}$ is the vector of parameters in the model. The glucose load is regarded as a function $J(t)$

2.2. The simplest model.
Following Ackerman et al (1969), it is assumed that after fasting, the patient's concentrations have stabilized to G_0 and H_0. Consequently, we have that

$$P(G_0, H_0, \mathbf{u}) = 0, \quad Q(G_0, H_0, \mathbf{u}) = 0.$$

We study the small deviations

$$g(t) = G(t) - G_0, \quad h(t) = h(t) - H_0.$$

Neglecting second order terms in Taylor's formula we are led to

$$\begin{aligned}
\dot{g} &= -m_1 g - m_2 h + J \\
\dot{h} &= -m_3 h + m_4 g
\end{aligned}$$

where m_1, m_2, m_3, m_4, are nonnegative constants.

After some time $J(t) \equiv 0$. Thus, the system reduces to the following second order differential equation,

$$\ddot{g} + 2\alpha\dot{g} + \omega_0^2 g = 0$$

A viable interpretation of the glucose-insulin system, is that of a damped harmonic oscillator. Hence we assume

$$\alpha^2 - \omega_0^2 < 0.$$

It is readily seen that the general solution is

(2) $$g(t) = Ae^{-\alpha t}\cos(\omega t - \delta).$$

for some constants A, α, ω, δ.

2.3. **The parameter estimation problem.** Let us denote the OGTT data for patient j by G_0^j, G_{30}^j, G_{60}^j, G_{90}^j, G_{120}^j.

Glucose concentrations for 80 female patients have been provided with the following conditions:

- 51 Healthy (H). $j = 1, \ldots, 51$
- 4 Impaired Fasting Glycaemia (IFG). $j = 52, \ldots, 55$
- 15 Impaired Glucose Tolerance (IGT). $j = 56, \ldots, 70$
- 7 with both alterations (IFG-IFT). $j = 71, \ldots, 77$
- 3 with Type II Diabetes Mellitus (T2DM). $j = 78, \ldots, 80$

Let us assume that $J(t) \equiv 0$ for $t > 0$ and define $g_t^j = G_t^j - G_0^j$ for $t = 30, 60, 90, 120$.

The problem of concern is: Given OGTT data for patient j: g_{30}^j, g_{60}^j, g_{90}^j, g_{120}^j, estimate the parameters

$$\mathbf{u}_j = (A_j, \alpha_j, \omega_j, \delta_j)^t,$$

constrained to

$$g^j(t) = A_j e^{-\alpha_j t}\cos(\omega_j t - \delta_j).$$

Remark. (i) The problem is commonly posed as a constrained optimization problem. If $\mathbf{y}$ is the data vector, and $\mathcal{G}$ is the so called observation operator, we pose

$$min \ |y - \mathcal{G}(\mathbf{u})|^2$$

constrained to

$$\frac{dG}{dt} = P(G, H, \mathbf{u}) + J(t)$$

$$\frac{dH}{dt} = Q(G, H, \mathbf{u}),$$

In most cases, there is no analytical solution of this ODE system and a numerical method is required.

(ii) We shall see below that the methodology to be introduced applies in general, and (i) is a particular case.

3. DATA FITTING AND SVM CLASSIFICATION

3.1. **Curve fitting.** There are some well known methodologies for fitting a curve through data by means of point estimates. Our aim is to provide additionally a gauge of the robustness of the point estimators. By formulating the problem as one of bayesian estimation, this is accomplished through the posterior probability density function, determined for each parameter. Details below.

In this paragraph we just illustrate curve fitting using the MAP estimator for some patients in different diabetic conditions. Healthy patients in Figure 1, whereas ill patients in Figure 2.
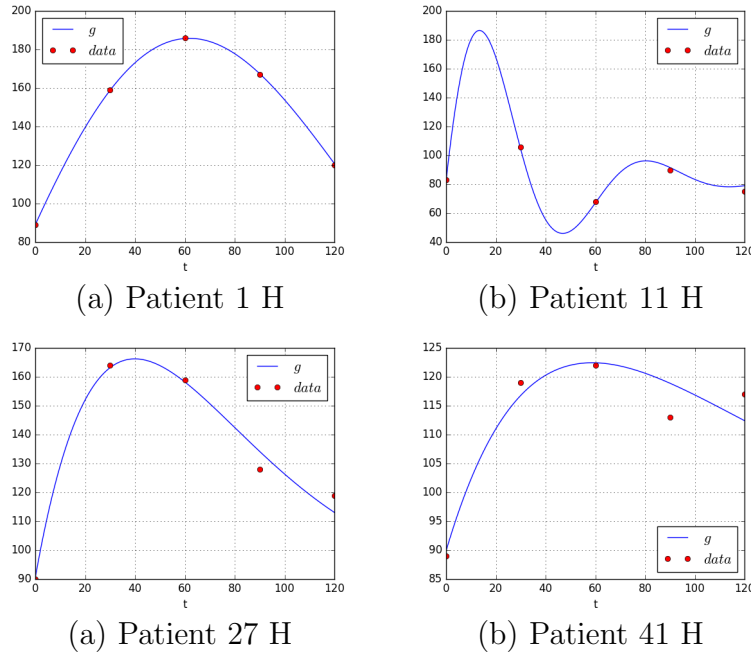

(a) Patient 1 H   (b) Patient 11 H

(a) Patient 27 H   (b) Patient 41 H

Figure 1. Fitting curves for some healthy patients

3.2. **Parameters proposal for classification.** Here we use the estimated parameters to classify a patient's condition. The most physically meaningful parameters are A, α

- A is related to patient's maximum increase of glucose concentration in response to glucose load.
- α is related to the patient's ability to attenuate the effect of the glucose load.

(a) Patient 53 IFG   (b) Patient 60 IGT

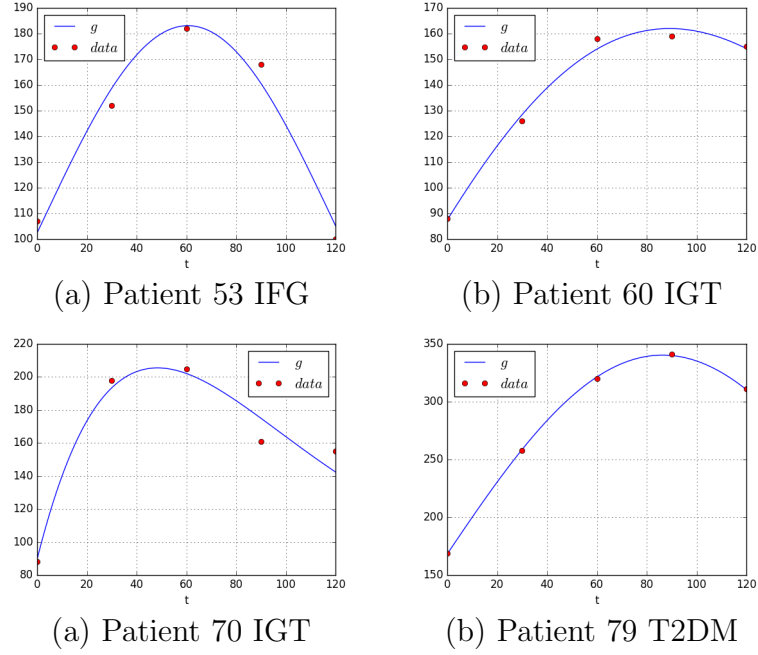(a) Patient 70 IGT   (b) Patient 79 T2DM

Figure 2. Fitting curves for some ill patients

We regard these as patient's indices to be used for classification of diabetic condition.

As a first step, we split patients in two groups, healthy patients and patients with a diabetic condition. In the plane $A - \alpha$ we perform a SVM linear separation. Here we follow the basic theory, see James et al [4].

The classification is successful for 85% of patients, results in Figure 3.
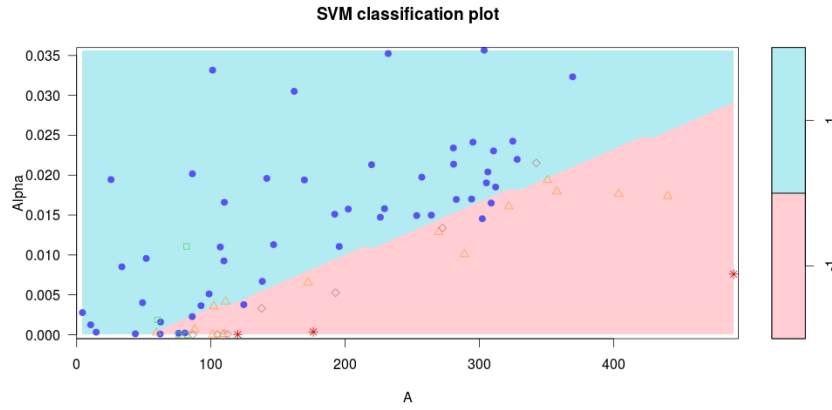

Figure 3.

Healthy (●), IFG (○), IGT (△), IFG-IGT (◇).

Several remarks are in order.

- At first sight, one may question the quality of classification in the neighborhood of the separating line. Nevertheless, there is an apparent *clockwise* transition from healthy to T2DM. The latter might have an important implication, namely, early detection of pre-diabetic patients. In such a case, a patient can be controlled with diet and exercise henceforth postponing medication.
- These results are obtained using the simplest of models.
- This may be regarded a double blind process. Data was gathered independently of the proposed methodology. The latter was tested with the data as provided.
- More data and planned sampling are needed to confirm our findings.

4. Bayesian Parameter Estimation

In this section our exposition is deliberatively terse, an excellent first read is Kaipio & Sommersalo [5]. For an insightful presentation see Stuart [9].

In bayesian estimation all variables are random, thus we consider the model

$$\mathbf{y} = \mathcal{G}(\mathbf{u}) + \eta$$

where $\mathbf{u}$ is the parameter to estimate, $\mathbf{y}$, η the noise and $\mathcal{G}$ the observation operator.

An important feature of bayesian estimation is to propose a *prior* probability density function, π_0, for the parameter $\mathbf{u}$. This prior encompasses all we know about $\mathbf{u}$. In essence it is a modelling problem.

Noise is supposed to be known with density ρ and independent of $\mathbf{u}$. Consequently, the conditional density $\pi^{\mathbf{y}}(\mathbf{u}) \equiv \pi(\mathbf{u}|\mathbf{y})$, the *posterior*, is given from Bayes' formula

$$\pi^{\mathbf{y}}(\mathbf{u}) \propto \rho(\mathbf{y} - \mathcal{G}(\mathbf{u}))\pi_0(\mathbf{u}).$$

The point of bayesian estimation is to determine the posterior. From this, point estimates can be obtained. Namely, the Conditional Mean (CM) and the MAP estimator. For the latter, an optimization problem is to be solved,

$$\mathbf{u}_{MAP} = argmax\, \rho(\mathbf{y} - \mathcal{G}(\mathbf{u}))\pi_0(\mathbf{u}).$$

This generalizes a well known deterministic approach. Indeed, consider a gaussian prior, $\mathbf{u} \sim \mathcal{N}(\mathbf{u}_0, \sigma^2\mathbf{I})$,

$$\pi_0(\mathbf{u}) \propto exp\left(-\frac{1}{\sigma^2}\left|\mathbf{u} - \mathbf{u}_0\right|^2\right)$$

If noise is also gaussian with zero mean and variance γ, we have

$$\mathbf{u}_{MAP} = argmax\left(exp\left(-\frac{1}{\gamma^2}\left|y - \mathcal{G}(\mathbf{u})\right|^2\right)exp\left(-\frac{1}{\sigma^2}\left|\mathbf{u} - \mathbf{u}_0\right|^2\right)\right)$$

$$\mathbf{u}_{MAP} = argmin \left(|y - \mathcal{G}(\mathbf{u})|^2 + \left(\frac{\gamma}{\sigma}\right)^2 |\mathbf{u} - \mathbf{u}_0|^2 \right).$$

Hence, $\mathbf{u}_{MAP}$ is Tikhonov'solution with regularization parameter, $\alpha = \left(\frac{\gamma}{\sigma}\right)^2$.

In our problem

$$\mathbf{u} = (A, \alpha, \omega, \delta)^t$$

and assuming gaussian noise we have,

$$\pi^y(\mathbf{u}) \propto exp\left(-\frac{1}{\gamma^2}|y - \mathcal{G}(\mathbf{u})|^2\right)\pi_0(\mathbf{u})$$

To sample the posterior we use emcee, an affine invariant Markov Chain Monte Carlo (MCMC) ensemble sampler, Foreman-Mackey et al. [3].

The prior may influence artificially the determination of the posterior, this is unwanted. In our case we are able to obtain satisfactory results with uninformative (uniform density) priors, only an estimate of the magnitude is required. We use uniform densities:

- $A \sim U[0.5\,g_m, 2.5\,g_M + 15]0.$
- $\alpha \sim U[0, 0.1].$
- $\omega_0 \sim U[0, 0.15].$
- $\delta \sim U[-2\pi, 2\pi].$

Where g_m, g_M are respectively the minimum and maximum of absolute values of G_0 shifted glucose concentration data.

Based on experiments, it was observed in Ackerman et al [1] that the parameter α is very sensitive to errors on G. Its use was not recommended for a diagnosis criterion. This is certainly true in a deterministic setting. In our approach we allow large observation errors with a gaussian model with standard deviation $\gamma = 5$, a typical value in the literature. An advantage of the bayesian approach, is that uncertainty of our point estimates is readily quantified by means of the marginal densities of the posterior. We found that densities for α are highly concentrated at the point estimates, mostly unimodal. For the patients above see Figures 4 and 5. Hence, we argue that the estimates are reliable.
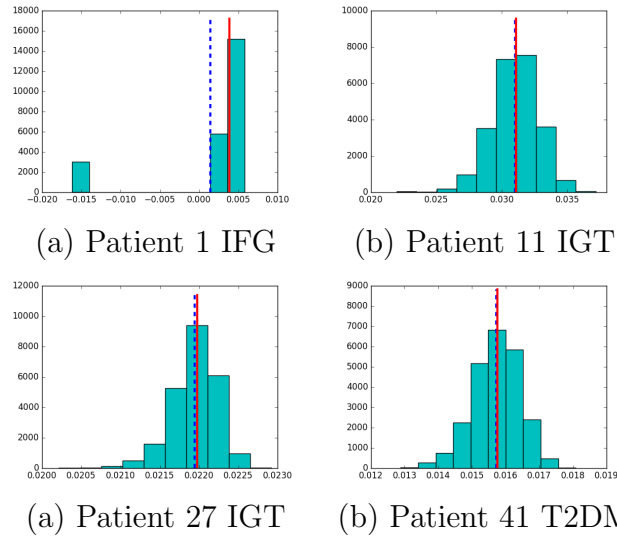
(a) Patient 1 IFG   (b) Patient 11 IGT

(a) Patient 27 IGT   (b) Patient 41 T2DM

Figure 4. Marginal posterior densities for healthy patients. MAP solid line. CM dashed line.

(a) Patient 53 IFG   (b) Patient 60 IGT
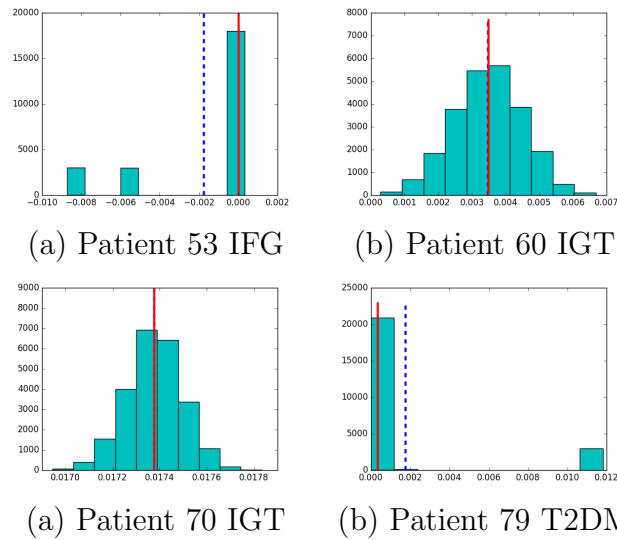

(a) Patient 70 IGT   (b) Patient 79 T2DM

Figure 5. Marginal posterior densities for ill patients. MAP solid line. CM dashed line.

5. Conclusions and future work

In this work we have considered a glucose-insulin interaction model for parameter estimation. The parameters are obtained from bayesian estimation constrained to a simple ordinary differential equation. Two of these parameters are chosen as patient's indices. A SVM classification algorithm show the potential of these indices to determine a patient's healthy or diabetic condition.

As a proof of concept for the introduced methodology, we started with the simplest of models obtaining satisfactory results. Consideration of more sophisticated glucose-insulin models is part of our current and future work. Also, more data is required to be conclusive.

Acknowledgements

References

[1] E. Ackerman, L. Gatewook,J. Rosevear, GI. Molnar; Blood glucose regulation and diabetes. Concepts and Models of Biomathematics, pp.131-156. (1969).

[2] A. Caumo, R.N. Bergman, C. Cobelli: *Insulin sensitivity from meal tolerance tests in normal subjects: a minimal model index*; J. Clin. Endocrinol. Metab. 85 4396. (2000)

[3] D. Foreman-Mackey, D. Hogg, D. Lang, J. Goodman; "Emcee: The MCMC Hammer." Publications of the Astronomical Society of the Pacific 125, no. 925 (2013): 306-12. doi:10.1086/670067.

[4] G. James, D. Witten, T. Hastie, R. Tibshirani; *An introduction to statistical learning* (Vol. 112). New York: springer. (2013).

[5] J. Kaipio, E. Somersalo; Statistical and Computational Inverse Problems, , Springer (2004)

[6] P. Palumbo, S. Ditlevsen, A. Bertuzzi, A. De Gaetano: *Mathematical modeling of the glucose-insulin system: A review*; Mathematical biosciences, 244(2), 69-81.(2013)

[7] G. Pillonetto, G. Sparacino, C. Cobelli; Numerical non-identifiability regions of the minimal model of glucose kinetics: superiority of Bayesian estimation; Math. Biosci., 184 (2003), pp. 53-67

[8] S. Rathee, Nilam; ODE models for the management of diabetes: A review; Int J Diabetes Dev Ctries 37: 4. https://doi.org/10.1007/s13410-016-0475-8 (2017)

[9] A. M. Stuart; Inverse problems: a Bayesian perspective. Acta Numerica 19 : 451-559. (2010)

M. A. Moreles, J. Peña, P. Vargas, Centro de Investigación en Matemáticas, Jalisco s/n, Valenciana, Guanajuato, GTO 36240,Mexico, *E-mail address*, moreles@cimat.mx:,

A. Monroy, Mexico General Hospital, Mexico City, Mexico, S. Alavez, Health Sciences Department, Metropolitan Autonomous University, Lerma, Mexico