Centro de Investigación en Matemáticas, A.C.

CIMAT

Statistical Analysis of OGTT Results

T E S I S

Que para obtener el grado de
Doctor en Ciencias
con orientación en
Probabilidad y Estadística

P r e s e n t a
Nicolás E. Kuschinski

Director de tesis:
Dr. J. Andrés Christen Gracia

Guanajuato, Gto. agosto de 2019

Statistical analysis of OGTT results

Nicolás E. Kuschinski

Abstract

Type 2 diabetes is a serious health condition that has only become more prevalent in recent years. One tool frequently used to help in its diagnosis is the oral glucose tolerance test, or OGTT. The methods currently in use for studying OGTT data, however, are basic, and do not take full advantage of the structure of the data created in an OGTT test. This work proposes a model for the mathematical analysis of OGTT data using Bayesian statistics on inverse problems. The main focus of the thesis is first to propose the model, and then to investigate various potential ramifications and improvements on it.

A model for OGTT data analysis is proposed, and tested on data from real OGTT tests with results that fit data well and closely match the intuition of medical collaborators.

A second test for diabetes has recently been proposed by researchers in Cuernavaca and Mexico City. Variable selection methodology is developed to study the ability of this tool to predict OGTT results. The results are unpromising. However, as a result of this analysis, a new technique for variable selection is proposed. This technique is christened FATSO and is useful for likelihood regularization in situations where intuitive parameter tuning is desirable. This is generally useful enough to give an importance to the new test which extends beyond merely OGTT analysis.

A potential improvement on the OGTT protocol is considered which would change the times at which OGTT data is collected. This modification is treated as a problem in the Bayesian design of experiments, and a new algorithm is developed for this purpose.

Another modification to the OGTT protocol is also considered which changes the method used for collecting samples, replacing it with one that is cheaper and which reduces patient discomfort. Although this method was thought to be too imprecise for the purposes of an OGTT, the mathematical model is adjusted to produce reasonable results from these data as well.

Agradecimientos

Muchas gracias al CONACYT por su apoyo económico durante la realización de este proyecto de investigación y al CIMAT, institución en la cual se realizó dicho proyecto.

En cuanto a individuos, agradezco primero que nadie a mi asesor, el Dr. Andrés Christen: Gracias por toda su paciencia y apoyo a lo largo de estos años.

De igual manera quisiera agradecer la participación de la Dra. Adriana Monroy por numerosas consultas acerca de los detalles médicos, y por la generosa y continua contribución de datos, a veces con algunas estructuras inusuales que se me ocurría pedir.

Agradezco a las siguientes personas por apoyar con consultas técnicas en diversas etapas del proceso de la investigación: El Dr. Peter Muller , el Dr. Marcos Capistrán, el Dr. Rogelio Ramos Quiroga, el Dr. Yussef Mazouk, y el Dr. Al Parker.

Agradezco a Yunuen Vital y a José Zubieta por su ayuda con la elaboración de algunos de los diagramas e imágenes.

Agradezco a Lucina Kathmann y al Dr. Nicholas Patricca por consejos relacionados con la redacción.

Agradezco también - de modo más indirecto - a todas las personas de mi comunidad quienes me brindaron apoyo moral en diversas etapas de la realización de este proyecto. Estos incluyen personas del Club de Go de Guanajuato, del CIMAT, de mi familia y de la comunidad en general. Son demasiados para mencionar, pero su apoyo fue invaluable y sentiría incompleta una sección de agradecimientos que no los mencionara.

Acknowledgments

I would like to offer many thanks to CONACYT for their financial support during this research project and to CIMAT, the institution where it was carried out.

Regarding individuals, first and foremost I would like to thank my advisor, Dr. Andrés Christen: Thank you for all of your patience and support during these years.

Similarly, I would like to thank Dr. Adriana Monroy for her contributions throughout several consultations about the medical details, and for her generous and continuous contribution of data with whatever unusual structures I thought of requesting.

I would like to thank the following individuals for offering their technical expertise at various stages in the process: Dr. Peter Muller, Dr. Marcos Capistrán, Dr. Rogelio Ramos Quiroga, Dr. Yussef Mazouk, and Dr. Al Parker.

I thank Yunuen Vital and José Zubieta for their help with the handling of some of the diagrams and images.

I would like to thank Lucina Kathmann and Dr. Nicholas Patricca for advice with the writing.

I would also – more indirectly – like to thank all the people in my community who offered moral support at various times during the completion of this project. This includes people from the Guanajuato Go Club, from CIMAT, from my family, and from the community at large. There are too many of them to name, but their support was invaluable and an acknowledgements section that did not include them would be incomplete.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Diabetes is a serious and potentially fatal illness that is on the rise and is expected to affect over 4% of the worldwide population by the year 2030 (Wild et al., 2004; American Medical Association, 2006). Diabetes occurs when the pancreas cannot produce enough insulin (a hormone which lowers blood glucose), or when the body is unable to efficiently use the insulin it produces, thereby reducing the effectiveness of the body's blood sugar regulation.

While type 1 diabetes is usually diagnosed very early in life, and is related to genetic disorders, the same cannot be said for type 2 diabetes, which is a far more common ailment that is acquired at an older age. It usually develops without any noticeable symptoms initially, but with time can become a serious problem which leads to severe complications, including death. Because of the long period of time without symptoms, it can often go undiagnosed for years. With a timely diagnosis and proper treatment, type 2 diabetes can change from a serious health risk to a relatively mild condition. Early diagnosis can also serve to identify patients who are at risk of developing type 2 diabetes and take steps to prevent this from occurring (American Medical Association, 2006).

In order to administer treatment or take preventative measures, it is of vital importance to have a good means of diagnosis. One common technique for diabetes diagnosis is the Oral Glucose Tolerance test (OGTT). To perform this test, a patient arrives after a night of fasting and has his/her blood glucose measured (in mg of glucose per dl of blood). The patient is then asked to drink a 75g glucose concentrate. Blood glucose is measured again at various times (typically over the course of two hours) and these measurements are used to study the body's ability to regulate sugar (American Medical Association, 2006; Jansson et al., 1980; Davidson et al., 2000).

The data produced by an OGTT are potentially very useful, but the classical tools for its analysis are crude. As a result, any conclusions drawn from OGTT tests do not use all of the information, and hence we find that the precision of diagnoses based on OGTT test results is often somewhat lacking. The primary aim of this thesis is to make significant contributions to the techniques used in the analysis of OGTT data, and to thereby improve their viability as a technique

for early diagnosis.

In chapter 2 we take the first steps to try to improve analysis of OGTT data. To do this, we create a dynamic model which represents some aspects of how the body processes blood glucose for the duration of the OGTT test. The critical point is that OGTT measurements are repeated measurements of a process over time and the dynamic model aims to represent the process itself. The dynamic model is based on a system of ODEs, which is solved numerically. Fitting the model to data is an inverse problem. To solve it, Bayesian priors are assigned to the relevant parameters, and posterior exploration is done via MCMC using the t-walk (Christen and Fox, 2010). The result of this modeling project is an improved model for the analysis of OGTT data which has the potential to lead to more accurate early diagnosis.

After the dynamic model has been developed, several further refinements are possible. The first issue of note is the inconvenience of performing the test to begin with. In chapter 3 we examine a new technique for breath analysis which was recently developed in UNAM, Cuernavaca and in Hospital General, in Mexico City (Gallego, 2016), and which may possibly allow for diabetes diagnosis without requiring a patient to go through the inconvenience of an OGTT. By merely requiring patients to breathe into a special bag, this technique involves measuring 92 metabolites in breath, and the researchers suspect that some of these may be indicative of diabetes.

Inferring which metabolites may be useful for detecting diabetes may be treated as a variable selection problem. One common technique for variable selection is likelihood regularization. In particular, one very common way to perform this kind of regularization is with the LASSO operator (Tibshirani, 1996). We extend the geometric mechanism of the LASSO operator to create a different operator which we call the FATSO operator. The FATSO operator is designed around the idea of making tuning parameters interpretable. FATSO is then used on simulated and real data with promising results.

To apply a linear selection operator to breath test data we require a single dimensional marker from OGTT data which is a stand-in for the presence of diabetes (or the severity thereof). For this purpose we have chosen the first time when blood glucose returns to its resting state. We estimate this marker for each individual for whom we have breath test data and we perform regression on the resulting dataset, applying the FATSO operator. The result is a set of candidate metabolites to consider as possible indicators of diabetes, but these do not match the set of candidate metabolites as selected by the proponents of the technique. For this reason, we choose not to pursue this particular avenue any further. The development of FATSO in chapter 3, however, is relevant on its own, and goes well beyond the analysis of OGTT data. In fact, it was tested on synthetic data as well as well known multivariate data from prostate cancer. In fact, we consider FATSO to be a potentially signficant and novel contribution to variable selection methodology in general, relevant in its own right, beyond the context of OGTT tests.

Returning to our investigation of OGTTs, we proceed to ask the question of whether OGTT tests themselves can be improved. In chapter 4 we investigate

this possibility by looking into ways of improving the testing protocol, restricting ourselves to cases that do not require changing the physical infrastructure or the associated technology. One way to do this is to reexamine the times at which blood samples are taken over the course of an OGTT test. There is no standard set of times, but locally, common practice is to measure blood glucose at the start of the test, one hour after drinking the glucose concentrate, and at the end of the test which is two hours after drinking the glucose (in this thesis, our data comes from more frequent measurements, but the data were collected specifically for research purposes). These times are somewhat arbitrary, so one possible improvement to the test is to select a new set of times at which to collect data.

The selection of measurement times is a problem of experimental design. To approach this problem we develop a new algorithm to compare Bayesian experimental designs, and then perform comparisons using this algorithm to find a good design. This approach is similar to some previous attempts at Bayesian experimental design, for instance see Christen and Buck (1998), but has a better mathematical justification. After selecting a design, some numerical experiments were done to validate it.

Changing the times for data collection is one way to improve the OGTT protocol to improve the accuracy of the collected data. Another way to change the protocol is to change the method for data collection. In chapter 5 we look into changes that alter the physical infrastructure required for the test. Specifically, we are interested in simplifying it. The usual OGTT protocol requires samples to be taken in a hospital, via a cannula, and then analyzed with a complex apparatus which requires significant time and effort to produce a result. There is, however, a very simple and inexpensive method to measure blood glucose, known as a glucometer.

Glucometer measurements are measurements of capillary rather than venous blood. They are known to be imprecise (Ginsberg, 2009), and were long thought to be insufficient for an OGTT. The new dynamic model, however, provides better information from OGTTs, and this yields some hope that we may be able to obtain enough information from glucometer measurements.

We attempt first to directly use glucometer data with the dynamic model, but we find that it is insufficient. We then investigate the difference between glucometer measurements and classical venous blood measurements. We identify two sources of error: A bias caused by the different blood type, and a greater variance caused by the measurement apparatus. After adjusting the model for these issues, we find that in many cases it is indeed possible to obtain sufficient information from glucometer measurements.

While the adjustment to the capillary glucose measurements is promising, there are still several cases where the results from capillary measurements do not match results from venous blood. To tackle this issue, we reexamine our error model for venous blood, requesting duplicate samples from our medical collaborator. These allow us to reformulate the error model for venous blood as well. The new model matches capillary inference in several cases which it did not before, providing further evidence that, with proper analysis, glucometer

measurements are a viable alternative to study OGTTs in most cases.

Overall, this work offers improvements which hope to greatly increase the descriptive power and scope of analytical methods for OGTT data. This thesis proposes a new dynamic model and uses it to analyze data. After, the thesis proposes two possible alternate versions of the OGTT test protocol which can be considered, in one case to improve the accuracy and in another to reduce the inconvenience in the test. It is expected that these results may serve to make improvements to OGTT protocols and analysis, and ultimately to improve our ability to diagnose diabetes. This thesis also explores an alternative to OGTT tests, and although this alternative is ultimately not chosen for further research, the investigation already conducted on this issue results in a contribution to the theory of variable selection.

Chapter 2

The ODE model and the inverse problem

2.1 OGTT tests

As explained in the introduction, for diagnosis of type 2 diabetes, one common test is the Oral Glucose Tolerance Test, or OGTT. For this test, a fasting patient arrives and his or her resting glucose is measured from a blood sample. The patient then drinks a 75g glucose concentrate and blood glucose is measured repeatedly over the course of the next two hours. The exact glucose measuring times vary depending on local practices. The results of these measurements are expected to provide some notion of how the patient's body handles the glucose (Jansson et al., 1980; Davidson et al., 2000; Anderwald et al., 2011).

In practice, the analysis of OGTT tests is usually done using very simple guidelines. Typically used markers include the average of the observed glucose measurements and/or the value of the first and last measurement. A patient is considered diabetic if the measurement chosen is above a certain threshold (typically $200 mg/dl$). While this analysis has proven to be useful, it disregards one of the primary qualities of OGTT test conditions: That they measure the *evolution* of a process over time (Davidson et al., 2000).

Accordingly, we propose a dynamic model based on Ordinary Differential Equations (ODEs) to model blood glucose during an OGTT. The idea of using mathematical models to analyze OGTT results is not new. Previously proposed models have not been used for inference, mostly because they lack the flexibility to explain many of the phenomena seen in OGTTs. For instance, Jansson et al. (1980) assumes that the body only lowers blood glucose, but in the course of measuring several patients we can see that this is not always the case (see real data in section 3.1)

In our approach, we use a dynamic model which was derived from the recommendations of our medical collaborators and follows the logic of previous related works such as Palumbo et al. (2013). It is flexible enough to describe most of

the observed behavior of glucose in real patient's OGTTs. Using Bayesian inference, we set appropriate priors on the parameters and fit this model to real data. We have been able to achieve good fits for observed data and our results match the intuition of our medical collaborators well enough that we consider our model a good candidate for serious analysis of OGTT data and, eventually, for early diasgnosis.

This chapter is organized as follows, in section 2.2 we present the dynamic model. In section 2.3 We develop a Bayesian statistical model that can be used to draw inference from the dynamic model. In section 2.4 we explain the details of how to perform inference from the model, and present results of said inference on real patients. Finally, section 2.5 concludes the chapter.

2.2 The dynamic model

Our model is based on the interaction of glucose, insulin and glucagon only. The glucose regulation system is far more complex but in the controlled environment of an OGTT these are by far the leading factors. Insulin is a hormone secreted by the pancreas which reduces blood glucose. Glucagon is also a hormone produced in the pancreas and has the opposite effect, it triggers the liver to produce glucose, thus increasing blood glucose levels. In simple terms, insulin is produced when blood glucose is high and glucagon is produced when blood glucose is low, making a feedback system of blood glucose level regulation (Jiang and Zhang, 2003; Palumbo et al., 2013).

Our dynamical model is represented by the following system of ODEs

$$\frac{dG}{dt} = L - I + \frac{D}{\theta_2} \tag{2.1}$$

$$\frac{dI}{dt} = \theta_0 (G - G_b)^+ - \frac{I}{a} \tag{2.2}$$

$$\frac{dL}{dt} = \theta_1 (G_b - G)^+ - \frac{L}{b} \tag{2.3}$$

$$\frac{dD}{dt} = -\frac{D}{\theta_2} + \frac{2V}{c} \tag{2.4}$$

$$\frac{dV}{dt} = -\frac{2V}{c} \tag{2.5}$$

where the meaning of each of the state variables and parameters is explained in table 2.1.

The heuristics behind this model are similar to other glucose-insulin models (Palumbo et al., 2013, for instance) and are as follows. There is a threshold level of glucose which the body hopes to maintain which is denoted by G_b. It is set at $80mg/dl$ for all examples in this thesis, but it can be adjusted or inferred otherwise if that is deemed appropriate. If blood glucose goes above G_b then insulin is produced, increasing $\frac{dI}{dt}$ as indicated by (2.2). As insulin is produced, this acts to reduce glucose concentration in the blood, reducing $\frac{dG}{dt}$ as indicated by (2.1). The opposite effect is achieved by glucagon, as seen in (2.3) and (2.1).

	Interpretation	Value
G	Blood glucose.	State variable
I	Blood Insulin.	State variable
L	Blood Glucagon.	State variable
D	Glucose in digestive system.	State variable
V	Glucose not yet in the digestive system	State variable
θ_0	Insulin responsiveness	Unknown par.
θ_1	Glucagon responsiveness	Unknown par.
θ_2	Glucose digestive system mean life.	Unknown par.
a, b	Insulin and Glucagon clearance mean life.	31 min.
c	Time taken to drink most of the glucose solution	5 min max.

Table 2.1: Meanings of state variables and parameters in the OGTT model.

Insulin and glucagon are both metabolized and decrease with mean lives a and b as seen in equations (2.2) and (2.3), respectively.

$D(t)$ and $V(t)$ represent glucose which is moving into the bloodstream. It begins outside the body, ie. the sugar concentrate $V(t)$, decreasing and moving into the digestive system, $D(t)$, as seen in (2.5) and (2.4), and then from the digestive system moving into the bloodstream, as seen in (2.4) and (2.1).

Time is measured in hours, and blood glucose is measured in mg/dL of blood. The units of glucagon and insulin are more abstract and can be thought of in terms of their effect on units of blood glucose. Insulin and glucagon responsiveness include both the generation of the hormone and also the response of the body to the hormone after production. The model is not intended for insulin nor glucagon level prediction and only glucose measurements are available, therefore in our model the units of I and L are not relevant and not directly interpretable.

a and b are extrapolated from best estimates of insulin and glucagon clearance time from Duckworth et al. (1998). Similarly, estimates exist on times of glucose absorption into the body (Anderwald et al., 2011), but these vary greatly from patient to patient and thus θ_2 is inferred. Jointly with θ_0 and θ_1, which are also inferred, these parameters represent the patient's condition in our model.

For the examples in this thesis, the system of ODEs is solved numerically (there is no known analytic solution). This is done by using the odeint function in the scipy package of the python programming language, (Jones et al., 01). This uses an implementation of the LSODA algorithm, described in Petzold (1983).

While some justification for the dynamic model comes from the heuristics, this is secondary to the real issue, which is whether its behavior can adequately represent what happens to glucose inside a patient's body. In figure 2.1 we see glucose curves which follow from the model. The first three curves all start at G_b, to show the behavior of the model when the patient is already stable. One of the purposes of asking patients to fast beforehand is precisely to obtain

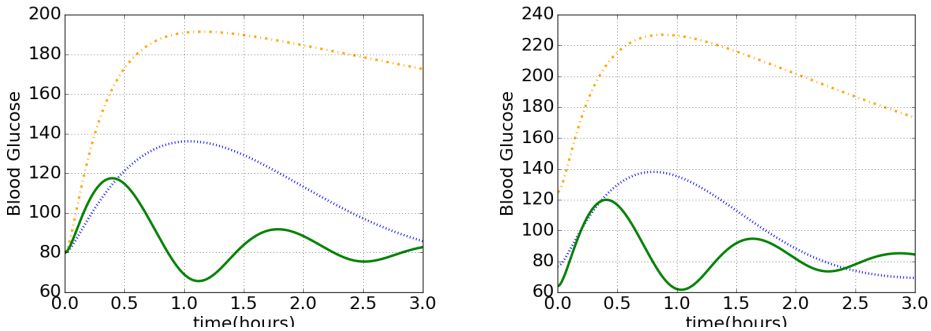


Figure 2.1: **Left:** Three curves produced by our model, all beginning with $G(0) = G_b = 80mg/dl$ these represent three kinds of patient: The dotted line is a healthy normal patient, the broken line is a diabetic patient who does not adequately regulate insulin, and the solid line is an oscillating patient, whose insulin and glucagon response is very strong. **Right:** Curves showing similar scenarios but with slightly different parameter values, including $G(0)$.

this behavior – however, particularly for diabetic patients, fasting may not be sufficient and glucose may begin elsewhere. The curves in the right panel of figure 2.1 represent a scenario wherein glucose begins somewhere other than G_b.

2.3 Statistical model for the OGTT data analysis: The Inverse Problem

In order to perform inference on the OGTTs of real patients, the model must be fit to the data, ie. the patient's glucose readings over the course of the test. For instance, for one real patient, at times $t =$ 0:00, 0:30, 1:00, 1:30, and 2:00 hours we obtained glucose measurements of $y = 81, 156, 141, 102$, and 89 mg/dl respectively. The intent is to use these data to infer the glucose curves. We assume data to be observations of $G(t)$ at the measured times and model the data y with

$$y_i = G(t_i) + \epsilon_i$$

where $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ and $\sigma = 5mg/dl$ for all examples in this chapter. This allows us to write the likelihood as

$$f(y|\theta) \propto \prod_i e^{-\frac{y_i - G(t_i|\theta)}{2\sigma^2}}.$$

Fitting this kind of model is considered an inverse problem; a non-linear regression problem with a complex regressor defined through a system of ODEs. These systems are frequently characterized by drastically different sets of parameters fitting well with the same data, leading to many explanatory scenarios.

For this reason, classical statistical estimators, such as the least squares estimator, only select one possible scenario, often quite an unreasonable one, following the data closely. There are several ways to address this issue, but one popular choice is to use Bayesian inference and encode some notion of what reasonable parameter combinations are in the prior distribution (Fox et al., 2013; Kaipio and Somersalo, 2006).

For all of the examples in this chapter, the following priors were used:

$$\begin{aligned}
\theta_0 &\sim Gamma(2,1) \\
\theta_1 &\sim Gamma(2,1) \\
\theta_2 &\sim Gamma(10, 1/20)\ \mathbb{1}\{\theta_2 > 0.16\} \\
G_b &\sim \mathcal{N}(80, 10000) \text{ truncated to } [30, 400].
\end{aligned}$$

The priors for θ_0 and θ_1 were chosen to give high probability to all values estimated from even the most extreme patients that have been analyzed in this way. The prior for θ_2 is chosen to match information in Anderwald et al. (2011). The prior for G_b is centered on healthy patients and is truncated since any patient whose initial glucose is outside of this range should not undergo an OGTT test but instead be placed into emergency care (Our medical collaborator performs a preliminary finger stick glucose test precisely for this purpose. For further information on finger stick tests, see chapter 5).

2.4 Inference

The object of interest is the function $G(t)$ for each patient and inference is performed on data from each patient separately leading to a separate posterior for $\theta_0, \theta_1, \theta_2$ and G_b for each patient. Our objective here is not a population study, and hence we concentrate on studying our model and its ability to fit OGTT data parsimoniously. Posterior exploration is achieved using MCMC techniques. Most MCMCs must be tuned to the posterior for each situation and in this case for each patient. A practical alternative is to use a self-tuning MCMC algorithm. One such algorithm is the t-walk, which is an MCMC algorithm that adapts to the scale of the target distribution. This is the algorithm that was chosen for this case, see Christen and Fox (2010).

Posterior exploration can be done in reasonable time even without high end hardware. All the examples in this chapter were performed on a laptop computer with an i5 processor and took less than 2 minutes to perform 15000 iterations of the t-walk. This represents 150 pseudo-independent posterior samples (using higher than necessary autocorrelation times, to account for patients with posterior distributions which are harder to explore than usual). This is quite an acceptable numerical processing time, since it takes 2 hours to gather the blood samples and processing is typically done overnight, depending on the availability of staff and laboratory equipment.

2.4.1 Results on real data

Figure 2.2 shows a posterior sample for three real patients. The curves fit the data well, even for the third patient (top to bottom), whose data would not fit a curve which does not account for glucagon. The first patient is a healthy patient, whose body handles glucose normally. The second patient is a potentially diabetic patient, whose glucose does not return to the baseline during the test. The third patient is a patient whose body responds rapidly to glucose, causing oscillations. Performing inference on many patients has shown that the latter is not an unusual or rare situation.

These curves display more nuance than current guidelines or practices for OGTT analysis. For instance, current practices would not distinguish between the first and third patients, despite their metabolism showing clearly different behavior, since the maximum measured value for both patients is similar (note also that although the first patient has higher glucose measurements, the third actually achieves a higher peak value in the estimated curve; this information can only be found by considering the temporal aspect of the measurements). Our model has strong descriptive power, giving reasonably small uncertainty for times in the measurement interval. It also has reasonable predictive power for a short time outside of the measurement interval as can be seen by prolonging the function $G(t)$ beyond the last measurement (in our graphs we prolong this an additional hour.) This can be thought of as a projection of what the patient's glucose *would* be if the conditions of the experiment were to continue. It is not clear, however, how long the dynamics of the system can be expected to remain intact, so this interpretation should only be considered over a short term.

Figures 2.3, 2.4 and 2.5 show histograms obtained from the MCMC posterior sampling for each of the model parameters for the patients from figure 2.2. The priors are represented with solid lines for reference. We may note that for patients without measurements below their resting glucose levels the data is uninformative about θ_1, which represents glucagon response. This is to be expected since in our model glucagon does not kick in unless blood glucose goes below G_b. Oscillating patients do provide data that is informative with regards to θ_1.

2.5 Conclusions

The diagnosis of type 2 diabetes is an important public health issue, and it requires a more sophisticated tool than the direct recording of values from the test, not only because these values are insufficiently informative, but also because they do not account for measurement error.

Our model shows that overall it is able to represent the results of OGTT tests for nearly all patients for whom a fit was attempted. For one patient for whom the fit failed, it was later discovered that there was an error when recording the data, and the failure of the model to fit was an indication that triggered this error's discovery. The model also displays significantly deeper nuance and detail

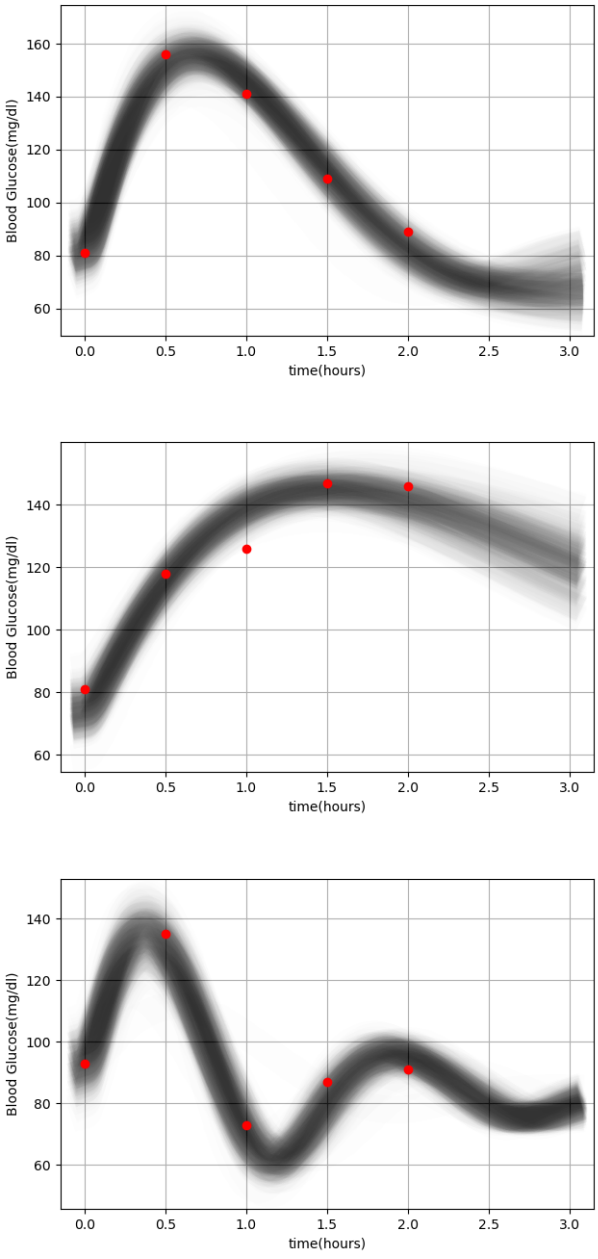


Figure 2.2: OGTT inference for three patients. The first appears to be a healthy patient, the second a diabetic and the third an oscillating case. The graphs show the posterior distribution of $G(t)$ over 3 hours. Each vertical slice is a kernel density estimate of the posterior distribution of $G(t)$ at that time. The dots are the collected data.

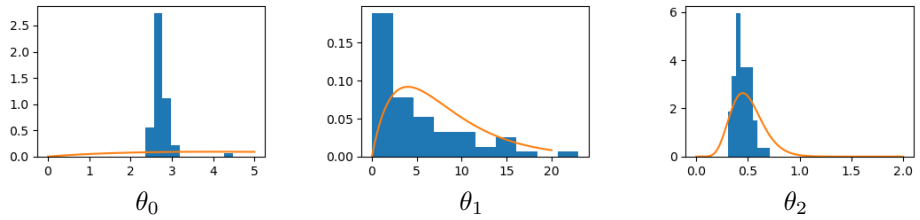


Figure 2.3: Histograms of posterior samples for the first patient. They are the parameters θ_0, θ_1, and θ_2 respectively. They are superimposed on a graph of the prior density of each parameter. In particular we note that for θ_1 the posterior matches the prior closely, and for θ_0, the data is extremely informative.

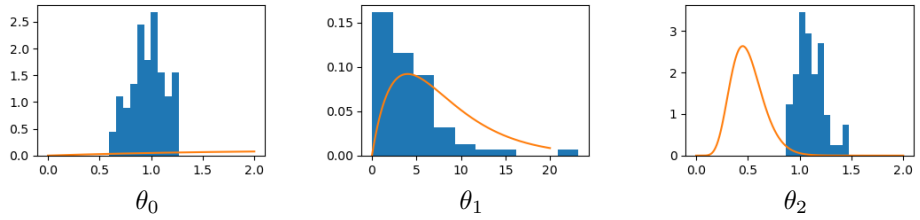


Figure 2.4: Similar graphs for the second (potentially diabetic) patient. Once again we note that θ_1 once again deviates very little from the prior. This is caused by having no data below G_b.

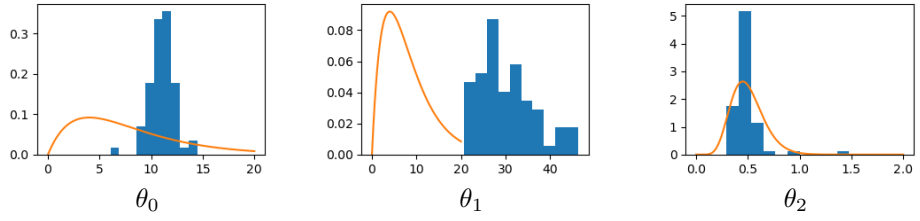


Figure 2.5: Similar graphs for the third (oscillating) patient. We note that in this case, the data that is below G_b gives us information about θ_1.

than previous analysis techniques could ever hope to represent.

At present, this model serves for the analysis of OGTT data, but not for diagnosis. The reason for this is that this model provides much more information than had previously been available, and our medical collaborators are - as of yet - uncertain about how to interpret this new information for which they have not been trained. Further study is required in order to transform full glucose curves into diagnoses or treatment recommendations. That said, this kind of a study should be well worth the effort.

At present, there is no known method which serves to diagnose initial stages of type 2 diabetes quickly, and accurate diagnosis may only be done by following a patient over time. Regarding our model, we can envisage a faster and simpler solution based on a single dimensional marker. One single dimensional marker that seems reasonable is $min\{t : G(t) = G_b \text{ and } G'(t) < 0\}$ (first return of blood glucose to the base level G_b), although simply using the marginal posterior distribution for θ_0, and comparing it with reference θ_0 values in healthy patients, might also be a possibility.

We consider this model a strong candidate for further research in the analysis of OGTT data. However, even if not this specific model, some sort of dynamic model with strong descriptive power is required for the important and delicate issues involved in the analysis of OGTT tests.

Chapter 3

Breath tests and FATSO

3.1 Motivation

Recently, a new mechanism for the analysis of metabolites in breath was developed in the UNAM, Cuernavaca and in Hospital General, in Mexico City. The technique involves collecting a sample from a patient's breath in activated carbon, and then studying the sample with gas chromatography. This process measures the prevalence of 92 organic compounds in breath. Gallego (2016) studied this technique as a potential way to predict diabetes. She collected a sample of 35 patients (27 healthy, 3 at risk and 5 critical diabetics) for whom a clinical diagnosis was available and performed this test. OGTT data for nearly all patients was also available.

The prospect of using breath metabolites as a test for diabetes is very exciting because it is noninvasive and fast. If effective, this would allow patients to be diagnosed without having to go through the hassle of an OGTT test. The only thing required of a patient is to rinse his/her mouth and then breathe into a bag.

The question of whether the data collected is a good predictor for diabetes was studied in Gallego (2016). The primary evidence in favor is a principal component analysis of the data. The data is a 35×92 matrix of metabolite counts. The results of a PCA analysis on this matrix were charted, and the 2 main components are charted in figure 3.1 (image taken directly from Gallego (2016)). Healthy patients are colored blue, patients at risk are orange, and critical patients are red. This chart appears to show some correlation between the severity of diabetes and the principal components of the data matrix, and the author went so far as to add colored ellipses to highlight the apparent division of the patient grouping.

There is, however, a problem with figure 3.1 that is not visible at first glance, and it is the conspicuous absence of patient number 5 from the chart. In figure 3.2 we use the same data to replicate the PCA analysis using the same coloring. We note the presence of a severe outlier. This is patient number 5, whose first

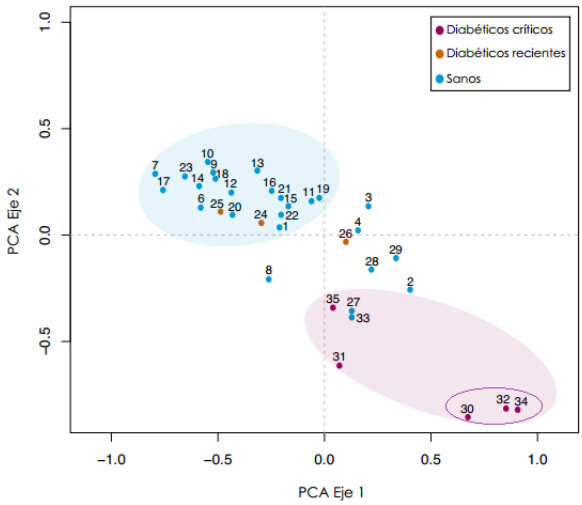


Figure 3.1: Figure from Gallego (2016) showing the principal component decomposition of 35 patients' breath test data and the severity of diabetes related illness. While it is difficult to spot immediately, patient 5 is missing.

component is an order of magnitude larger than any other patient's (And we note that in our first graph, larger values of the first component would appear to indicate higher severity of diabetes), however this patient was classified clinically as healthy. We used the data from this patient's OGTT to perform posterior inference, and the result is seen in figure 3.3, which is a normal OGTT curve which might be expected for a healthy patient. While this does not invalidate the medical physics in Gallego (2016), it does significantly undermine the conclusions about the relationship between the principal components of the breath test and diabetes. In particular, it is well known that PCA analysis is sensitive to outliers, and hence, even if this outlier is discounted, it severely affects which components are being plotted.

These considerations aside, we attempted to use the data for inference. With a $35x92$ matrix this is a $p > n$ problem, and there is an infinite number of linear combinations of the metabolites which generate any vector of length 35. We hence approach this as a standard $p > n$ linear regression problem, and investigate selection operators and regularization.

The chapter is organized as follows. First, in section 3.2 we introduce the topic of variable selection in linear models. In section 3.3 we establish our notation and investigate the geometric properties of the LASSO operator. In section 3.4 we introduce the FATSO operator, which is based on the geometrical properties discussed in section 3.3. In section 3.5 we study the behavior of FATSO and see how it addresses the issue of parameter interpretability. Section 3.6 discusses the differences between FATSO and various other extensions of LASSO. In section 3.7 we look into the observable effects of the parameters in

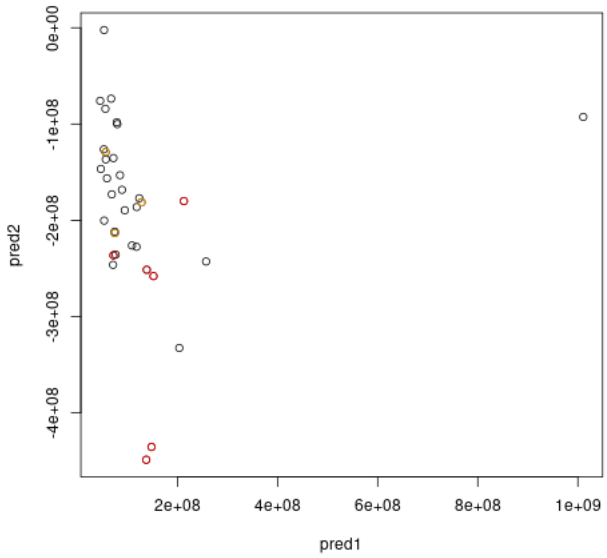


Figure 3.2: Our own principal component analysis, using the same data as from figure 3.1. We use the same coloring to indicate the severity of diabetes related illnesses, and we note an outlier. This is patient 5, who is classified as healthy!

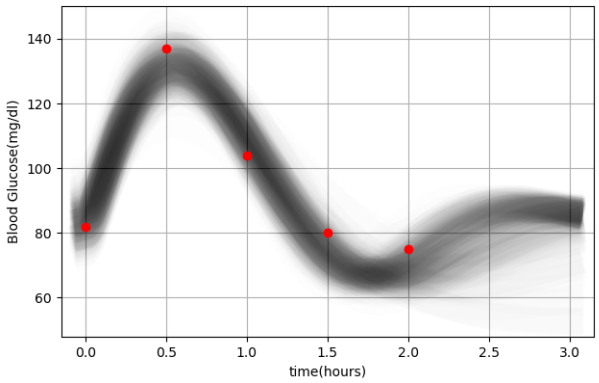


Figure 3.3: OGTT inference for patient 5, who is the outlier in the breath tests. As we see, this patient's OGTT shows him/her as being normal, as does his/her clinical history.

numerical examples and with real data. Section 3.8 explores the results of using FATSO on the data from the breath tests, and finally, section 3.9 gives some final thoughts.

3.2 Variable selection and LASSO

In standard linear models, it is not uncommon to have prior knowledge that several of the regression coefficients should be zero. This happens, for example, when it is suspected that most of the factors considered in a large model are not relevant. The identity of which factors *are* relevant, however, is not known beforehand. It is of interest to estimate model parameters and to identify which coefficients are nonzero. This is a classical problem with a classical solution, wherein regression is performed on the model and then the parameters are tested one at a time to determine whether they are significantly nonzero. Often this is followed by a second round of inference using only those parameters determined to be nonzero the first time through (see Rencher, 2008, for example).

This procedure works well for large sample sizes and small dimensions, but for high dimensions (large numbers of explanatory variables X) or small samples sizes, eg. the $p > n$ problem, it becomes impossible to perform linear regression using classical techniques since the response Y is typically found exactly inside of the column space of X, and there are infinitely many exact solutions.

There are several ways to handle to this problem, but many of them center only around estimation and do not intend to identify relevant factors. Methods that do intend to separate relevant from irrelevant variables are known as *variable selection* methods. There is a broad body of recent literature on the subject of variable selection in extremely high dimensional problems, such as those which are frequently encountered in gene selection and microarray data (Guyon and Elisseeff, 2003). In this chapter we will focus on linear regression problems, usually with a more manageable (if still large) number of dimensions. We will also find some justification for using variable selection techniques even in low dimensional problems.

In order to obtain good estimates in these situations, one common solution is to use regularizing operators. One popular such operator is the LASSO operator (Tibshirani, 1996), which is designed to yield point estimates which are frequently exactly zero. Tuning the degree of selectivity of the LASSO operator, however, is not very fluid. The degree of selectivity is tied to shrinkage of the estimators and it is difficult to interpret.

While LASSO is a very popular operator for variable selection in linear models (and has been tried in non linear models also, see Ribbing et al., 2007), several other regularization methods exist, such as ridge regression (Hoerl and Kennard, 1970), bridge regression (Park and Yoon, 2011), elastic net (Zou and Hastie, 2005), etc. While these operators have several important virtues, none of them address the issue of interpretability in variable selection.

In a Bayesian setting, a large number of selection operators have been suggested recently in the form of scale mixtures of normals. For instance the

horseshoe prior (Carvalho et al., 2010), Dirichlet-Laplace priors (Bhattacharya et al., 2014) and others (Liang et al., 2008). These approaches also do not focus on interpretability issues. We suggest a selection operator that is not scale a mixture of normals; we follow a different strategy.

In the remainder of this chapter, we look into the geometric mechanism by which LASSO promotes regression estimators to zero, and we study some of the consequences. Using this information we propose a new family of operators which use the same geometric mechanism as LASSO, but provide an extra parameter which permits fluid and intuitive tuning of the degree of selectivity separately from shrinkage. We prove that this family of operators corresponds to a large family of Bayesian prior distributions, and we study the relationship between the geometry of the priors and the meaning of the parameters in a Bayesian context.

3.3 The LASSO operator

Consider a standard linear model of the form

$$Y = X\beta + \epsilon$$

where Y is the $1 \times n$ data vector, β a $1 \times p$ vector of parameters, X^T is the design matrix and the errors are $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$. In a situation in which we suspect that many of the coefficients in β are 0 the problem of interest is estimating which coefficients are nonzero along with their value. When the dimension of β is high in relation to the sample size, classical inference does not work, so this problem becomes a problem of variable selection. For this purpose, one common technique is to use the Least Absolute Shrinkage and Selection Operator (LASSO), see Tibshirani (1996). In classical statistics LASSO is seen as a likelihood penalization, and in Bayesian statistics it is treated as a Laplace prior (Park and Casella, 2008). In the Bayesian setting, the MAP corresponds with the classical estimator, namely

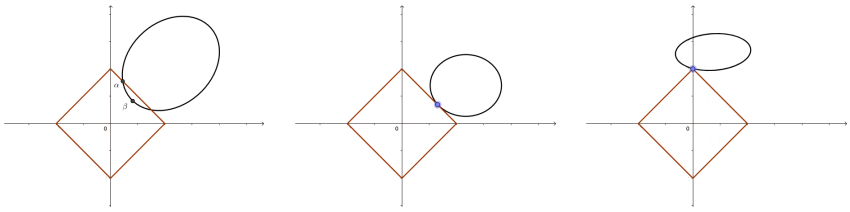
$$\hat{\beta} = argmax_\beta \left[\mathcal{L}(\beta, X, Y) - \lambda \sum |\beta_i| \right]$$

where $\mathcal{L}$ is the Gaussian log-likelihood function. The expression $\lambda \sum |\beta_i|$ is the LASSO operator and it depends on the value of a parameter λ.

A popular alternative parametrization for LASSO is to write the operator as $k/\sigma^2 \sum_{i=1}^p |\beta_i|$, which makes the LASSO estimator equal to

$$\begin{aligned} \hat{\beta} &= argmax_\beta \left[-\frac{1}{2\sigma^2} ||Y - X^T\beta||_2^2 - \frac{k}{\sigma^2} \sum_{i=1}^p |\beta_i| \right] \\ &= argmax_\beta \left[-\frac{1}{2} ||Y - X^T\beta||_2^2 - k \sum_{i=1}^p |\beta_i| \right]. \end{aligned}$$

Thus, $\hat{\beta}$ no longer depends on σ.



(A) (B) (C)

Figure 3.4: The three forms of intersections of level curves of the likelihood (elipses) and the LASSO operator (squares). (A) cannot happen at the MAP estimator, (B) corresponds to the likelihood curve being tangent to the slope of the prior and (C) corresponds to the likelihood curve intersecting the prior at an extreme, in this case the parameter in the x axis is sent to exactly zero.

The reason why LASSO produces parameter estimations that are *exactly zero* can perhaps best be understood by examining its level curves. In the case where β is bivariate there are three possibilities for the geometry of the level curves at the estimator (see figure 3.4). The level curves of the likelihood and the operator may cross (type A), may be tangent (type B), or they might meet at a point at which the curves of the operator are non-differentiable (type C). We note that type A cannot be the estimator by a simple argument: The point marked α cannot be the estimator since at the point marked β the value of the likelihood is the same, but the value of the operator is greater. Hence, the estimator must be either type B or type C. It is a $\hat{\beta}$ of type C that interests us given that these situations make the MAP estimator of one parameter exactly equal to zero.

3.3.1 How LASSO promotes variable selection

When considering whether $\hat{\beta}$ is of type B or C, we find that it depends on the value of λ, and this dependence has a notable property.

For fixed X, with probability 1, random data Y will allow the lasso estimator to fulfill the following criterion: There exists ν such that if $\lambda > \nu$ then $\hat{\beta}$ is of type C.

Proof. Note that the LASSO operator may be viewed as the Lagrangian for the restricted maximization of the likelihood subject to $\sum |\beta_i| < t$ for some t. The larger the value of λ, the smaller the value of t, and when $\lambda \to \infty$ then $t \to 0$.

For the bivariate case, consider the slope of the level curve of the likelihood function at the origin. With probability 1, this slope will be neither 1 nor -1.

Note that the level curves of the likelihood function are concentric. Hence, there is an open ball around the origin where the level curve does not have a slope of 1 or -1 either. In this area, it is impossible for $\hat{\beta}$ to be of type B. Therefore, for large enough λ, $\hat{\beta}$ must be of type C.

Now for the general case, note that all of the bivariate marginals behave as the bivariate case just explained. □

For a fixed X and a fixed random Y, with probability 1 there exists ν such that if $\lambda > \nu$ then $\hat{\beta}_i = 0$ for all except one value of i.

Proof. ν is the maximum threshold for each pairwise comparison of β_i vs β_j. □

In other words, there is almost certainly a threshold for which any λ above this threshold will make all estimators zero except for one.

In general, there is no clear way to choose λ so as to select variables in any controlled way. In other words, we know that when λ grows, our selection becomes tighter and tighter, discarding more and more variables, but there is no interpretable measure of how *much* tighter. In other words, the choice of λ can run the gamut from allowing all coefficients to be nonzero to allowing only one of them, and no good way to control its degree of selectivity.

In practice, the most common method for selecting λ is to use data-driven techniques such as cross-validation (Obuchi, 2016; Ribbing et al., 2007).

In passing, we note the following important point: A known issue with LASSO is that estimations depend on the scale of the variables, so it is common practice to center the covariates and standardize them so that $\sum_i x_i^2 = 1$ (Ribbing et al., 2007), although recently there have been alternative suggestions on how to rescale the variables (Sardy, 2008). Regardless of the specific method, something must be done unless the scale of the covariates is carefully chosen. This point is critical not only in LASSO, but in other selection operators as well. For this chapter, we will assume that prior to any regularization, covariates have been centered and standardized in the way described above. This will become important when performing calculations related to our proposal later on, but it is equally critical in LASSO, so we mention it now.

3.4 An alternative proposal

We note, as seen in the proof of lemma 3.3.1, that the behavior of the level curves of the likelihood at zero is directly related to what variable selection choice will be made by the LASSO estimator. Essentially, the LASSO estimator will be either at a point where the level curves of the likelihood are at a 45 degree angle, or it will make a selection. The only time that it will select both variables regardless of λ is if the likelihood level curves are at a 45 degree angle exactly at 0 (for Gaussian data, the probability of this occurring is zero). In figure 3.5 we see a graphical representation of exactly where the LASSO estimator may be located (depending on choice of λ).

In order to address the issue of selectivity, we propose to alter the LASSO level curves. The idea is to propose a new set of level curves directly, and to build a selection operator from this proposal. The objective is to adjust the slopes of the level curves such that they span a continuous range. If the slope

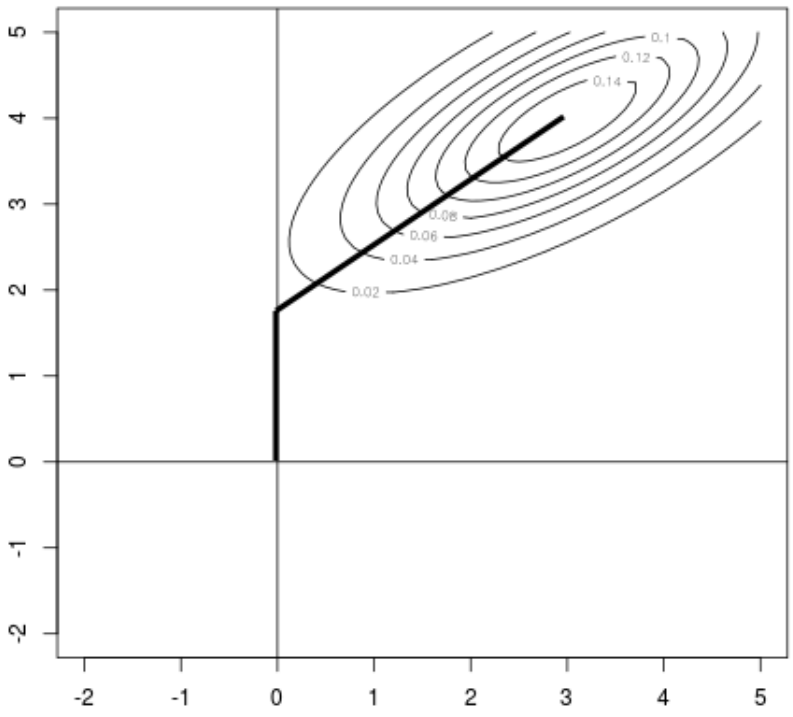


Figure 3.5: The possible locations for the LASSO estimator, as determined by the level curves. Which specific location corresponds to the LASSO estimator depends on λ. The dark line runs from the MLE along the points where the level curves are at a 45 degree angle, until it reaches an axis.

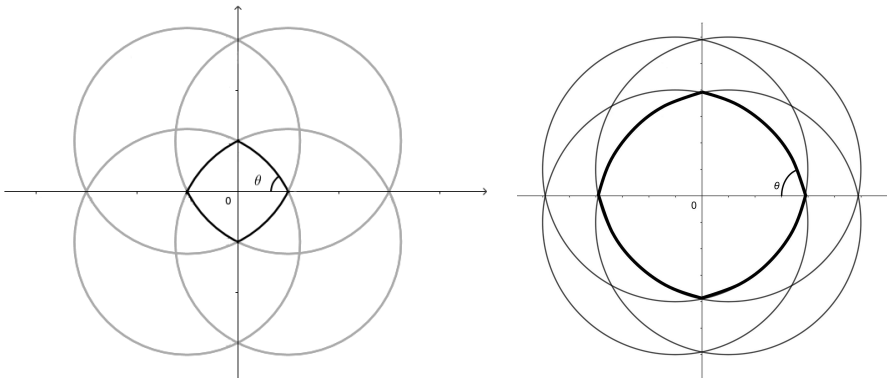


Figure 3.6: The proposed operator's level curves are the boundary of the intersection of disks. If the angle of the likelihood level curves at 0 falls between $\theta - \frac{\pi}{2}$ and θ then one variable will not dominate the other regardless of the level of shrinkage (λ in LASSO). We will refer to the angle θ introduced in this figure, and shown again in figure 3.7, in many parts of the chapter. The value of the level curve corresponds to the level of shrinkage, but a new parameter ρ is introduced, to change the geometry and the angle θ, which controls the position and size of the circles. The two images show the geometry with a different ρ and θ.

of the likelihood level curves at zero is in this range, then one variable will not dominate the other.

If the slope of the level curve at zero is in this range then $\hat{\beta}$ will be of type B regardless of the degree of shrinkage.

The geometry of the proposed level curves is the perimeter of the intersections of disks, as illustrated in figure 3.6 (or in general the boundary of the intersection of d-balls in dimension d). If the angle θ in the figure is the same for all level curves, then if the angle of the likelihood level curves at the origin is between $\theta - \frac{\pi}{2}$ and θ, then both variables will be selected regardless of the degree of shrinkage (parameter λ for LASSO). This construction will introduce a second parameter ρ, which determines θ, and which will be used in addition to a shrinkage parameter.

For the construction to make sense, the angles of intersection of the level curves with the axes must not depend on the degree of shrinkage. Consequently, the center of the corresponding circle will vary depending on which level curve we are on. We proceed to explore the necessary calculations for the construction of an operator from this idea.

Figure 3.7 shows the essential geometry used to calculate the location of the center of each curve. We note that triangles abc and AbC share intersection b and we also note that the angle at c is the same as the angle at C so these triangles are similar. We can therefore characterize the angle c by $\rho = ||ac||/||ab|| = ||AC||/||Ab||$. We can now write $a = a_\beta = \alpha\mathbf{1}$ where $\mathbf{1}$ is a $1 \times p$ vector of

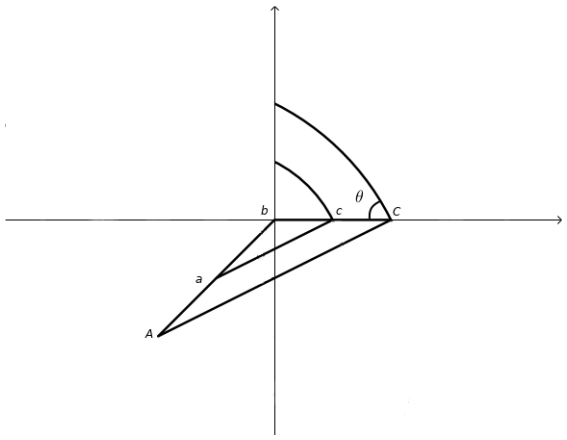


Figure 3.7: The geometry required for calculating the value of the operator. A and a are the centers of the circles , the arcs of which intersect the horizontal axis at C and c, respectively. Note that triangles abc and AbC are similar. This figure is a reference for several calculations throughout the chapter.

ones, and α is the distance from a to the origin along any given axis. ρ is the additional parameter in our operator, which will be directly related to the desired level of selectivity.

With this notation, and using d for the dimension of β it is now possible to write out the calculation

$$||ac|| = \rho||ab||$$

$$\sqrt{\sum_i (|\beta_i| + \alpha)^2} = \rho\sqrt{p\alpha^2}$$

$$\alpha^2 \left(p\left[1-\rho^2\right]\right) + \alpha\left(2\sum|\beta_i|\right) + \sum|\beta_i|^2 = 0.$$

In the range of interest, $\rho > 1$ and $\alpha > 0$ so we solve this equation to find a closed form expression for α

$$\alpha = \frac{-2\sum|\beta_i| - \sqrt{\left(2\sum|\beta_i|\right)^2 - 4\left(\sum|\beta_i|^2\right)\left(p\left[1-\rho^2\right]\right)}}{2\left(p\left[1-\rho^2\right]\right)}.$$

We must remember that in this section we have written a and α out of notational convenience, but that they depend on β and on ρ, so it really is $\alpha_{\beta,\rho}$ and $a_{\beta,\rho}$.

Now that we have computed the geometry of the problem, the remaining issue is to use this geometry to construct an operator (in this case one that will also match a prior distribution). Any probability distribution for which the level curves of the density function are concentric circles (or higher dimensional equivalent) centered at the origin may be used as a basis for the construction of

the operator. If the density function is $f(\beta)$ then we can construct a distribution with density function $g(\beta) \propto f(|\beta|+a_{\beta,\rho})$. This does not have a scale parameter unless f does, but most useful distributions do have one. We will refer to this family of priors as FATSOs or Flexible Axis-Thickened Selection Operators, and the basic form of FATSO will be based on a Gaussian distribution.

The FATSO will always be a probability distribution so long as f is also a probability distribution, and will have finite moments whenever f does since

$$\int h(\beta)g(\beta)d\beta \propto \int h(\beta)f(\beta+\alpha_{\beta,\rho})d\beta \leq \int h(\beta)f(\beta)d\beta.$$

The full formula for the Gaussian FATSO will have log density

$$log[g(\beta)] = K_{\rho,\lambda} + \lambda \sum (|\beta_i| + \alpha_{\beta,\rho})^2$$

for some normalizing constant K, which does not have to be computed since it does not depend on the β_is.

This distribution also has the following useful property:

The negative log density of the Gaussian FATSO is concave.

Proof. Note that $-log(\phi(x))$ (where ϕ is the univariate Gaussian density) is an increasing function for positive x. Note also that $-log[g(\beta)] = -log(\phi(\beta + \alpha_{\beta,\rho}))$. We also observe that $\alpha_{\beta,\rho}$ is convex when seen as a function of β, so the result follows. □

A trivial corollary is that, since the likelihood function for linear regression is also log-convex, then the posterior is log-convex and the calculation of the MAP is a convex optimization problem. Unfortunately most convex optimization algorithms require the use of gradients, and FATSO is not differentiable at any point where some $\beta_i = 0$, so the gradient does not exist at the expected optimum. That said, the convexity of the target function guarantees a unique maximum, and other desirable properties for optimization. Almost any optimization technique which does not depend on differentiability at the optimum will calculate the FATSO estimator effectively.

3.5 Interpreting FATSO and selecting parameters

The design of FATSO is based around the idea of reducing the collection of level curves for which parameter estimates are zero in a controlled way. Namely, the issue is the slopes of the level curves of the likelihood function at zero. By adding the parameter ρ, we have allowed an interval of these slopes to produce nonzero parameter estimates, rather than a single slope. This seems promising, but in order to be of real use, we need a proper way to interpret this slope and assign ρ (and λ in most cases) to fit our problem.

As we have previously observed, in the bivariate case, if the angle of the level curves is between θ and $\frac{\pi}{2} - \theta$ then both variables will be selected. For

interpretative purposes, let m be the slope $m = tan(\theta)$; θ as in figure 3.6. Following the geometry from figure 3.7 we can observe that b is a known angle $\left(\frac{3\pi}{4}\right)$. Using $sin(b) = \frac{1}{\sqrt{2}}$ allows us to calculate $\rho = \frac{|AC|}{|Ab|} = \frac{sin(b)}{sin(c)} = \frac{1}{\sqrt{2}}$ and $\theta = cosec^{-1}(\sqrt{2}\rho)$, so we have

$$\rho = \sqrt{\frac{1+m^2}{2}}.$$

We now have an easy conversion between θ, m and ρ, but on its own this brings us no closer to interpreting m nor to being able to set m (ie. ρ) in the FATSO operator.

The key to this crucial step is to calculate the slope of the likelihood level curve at zero. We note that the level curves are perpendicular to the gradient, so it is possible to study this slope by considering the gradient of the likelihood function at zero.

We observe that for the standard linear regression problem, the likelihood function is integrable, and a flat prior can be used to obtain a Gaussian posterior (Box and Tiao, 1992). We will not actually use a flat prior nor treat the result as a posterior, but for mathematical convenience, we can think of the likelihood function as if it were a Gaussian density $\pi(\beta)$ with mean μ at the MLE and covariance matrix $\Sigma = \left(X^T X\right)^{-1} \sigma^2 = \begin{bmatrix} \varsigma_i^2 & \varsigma_{ij} \\ \varsigma_{ij} & \varsigma_j^2 \end{bmatrix}$.

The gradient of $\pi(\beta)$ of a Gaussian density is (Petersen et al., 2008)

$$\frac{d\pi(\beta)}{d\beta} = -\pi(\beta)\Sigma^{-1}(\beta - \mu).$$

When we reduce it to the bivariate case, the slope of the gradient at zero is

$$\frac{\mu_j \varsigma_i^2 - \mu_i \varsigma_{ij}}{\mu_i \varsigma_j^2 - \mu_j \varsigma_{ij}}.$$

As previously explained, the covariates have been standardized so $\sum_i x_i^2 = 1$ and hence it is easy to observe that $\varsigma_i = \varsigma_j$, so we can write this quantity with ς, obtaining

$$\frac{\mu_j \varsigma^2 - \mu_i \varsigma_{ij}}{\mu_i \varsigma^2 - \mu_j \varsigma_{ij}}.$$

In the independent case, where $\varsigma_{ij} = 0$ (which can only happen if there is no intercept: If both the covariates and the response variable are centered then the intercept is always 0 anyway) this result is simply the ratio of the signals of the two parameters (note that with standardized covariates these are the pure effects on Y, free from the units of measurement; this can be seen easily since $\varsigma_i = \varsigma_j$ so $\frac{\mu_j}{\mu_i} = \frac{\mu_j \varsigma_i}{\mu_i \varsigma_j}$ or the quotient of signal to noise ratios, which are unitless). This corresponds well with an intuitive notion of the relative importance, or difference from zero, of one parameter to the other. In other words, this gives us an interpretation for the slope of the likelihood level curve at zero.

This intuitive notion is quite reasonable in the case of β_i and β_j are independent, but when they are correlated then it is lacking. If β_i and β_j tend towards zero together, for instance, then we would hope our notion of relative difference from zero would reflect that.

One way to attempt to correct this is to consider instead the conditional distribution of one variable given the other (Eaton, 2007),

$$\beta_i|\beta_j \sim \mathcal{N}(\mu_i + \frac{\varsigma_{ij}}{\varsigma_j^2}(\beta_j - \mu_j), (1 - \frac{\varsigma_{ij}^2}{\varsigma_i^2\sigma_j^2})\varsigma_i^2),$$

and calculate the conditional equivalent which we will call r_{ij}

$$r_{ij} = \left|\frac{\mathbb{E}(\beta_i|\beta_j = 0)}{\mathbb{E}(\beta_j|\beta_i = 0)}\right|.$$

This would give a more accurate representation of the relative difference from zero of the two variables, since it is the quotient of the means in the particular case of interest in which the other variable is zero (Also, $var(\beta_i|\beta_j = 0) = var(\beta_j|\beta_i = 0)$ so this is still unitless).

Figure 3.8 gives some intuition to show how the conditional distribution is a better choice than the marginal distribution. Both of the Gaussian distributions shown have the same marginal density, but in one case they are independent and in the other they are highly correlated. The difference between the relative importance of the two variables is visually apparent: If one of the variables is set to be zero, the other should be small as well.

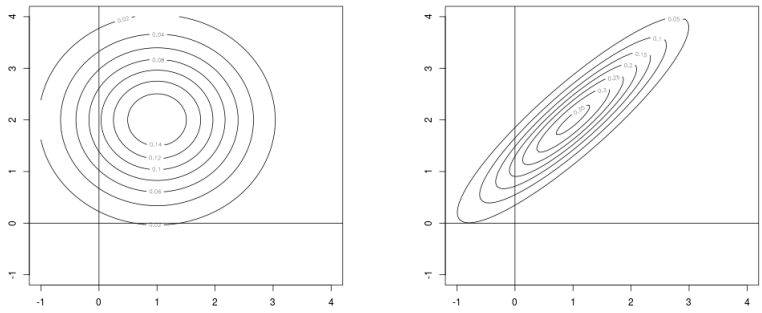
When we calculate r_{ij}, the result is exactly $\left|\frac{\mu_i\varsigma^2 - \mu_j\varsigma_{ij}}{\mu_j\varsigma^2 - \mu_i\varsigma_{ij}}\right|$, which is precisely the slope of the gradient of the likelihood at zero.

In other words, regardless of correlation, the slope of the level curves of the likelihood at zero matches the conditional signal ratio, r_{ij}, which is a good intuitive measure of relative importance between variables in a regression problem.

The user therefore assigns m as the circumstances require so that, regardless of λ, both β_i and β_j are selected if $r_{i,j}$ is between m and $\frac{1}{m}$. Our previous calculation allows us to set ρ when m is known, although it is also possible to simply use an alternate parametrization, working with m directly instead of ρ. This parametrization is easier to interpret and will be used from here on out.

This is nicely interpretable in two dimensions. In higher dimensions the structure is analogous and the mathematics are identical (One can Simply do the calculation with the marginal distribution of the two intended variables). The interpretation of the slope is slightly less intuitive, since the direction is determined by a vector rather than by a single number. The relationship of the corresponding components of the gradient, however, still matches $r_{i,j}$.

We have a way to interpret m. For full Bayesian inference, one would simply select m but it may also be reasonable to choose another path and simply try out values of m. Since the computational cost is low (unless the number of parameters is truly huge), a fair amount of information about the behavior and relative importance of parameters can be gleaned in fairly little time. In table



(A) (B)

Figure 3.8: Two Gaussian bivariate densities with the same marginals and different degrees of correlation. In (A) the variables are independent and (with the vertical variable as β_i and the horizontal variable as β_j) we have $r_{ij} = 2$. In the second case, the two variables are strongly correlated. When one variable tends towards zero, the other is also very small. In this case, intuitively, the variables are closer and there is less of a reason to prefer one over the other. This intuition is reflected by $r_{ij} = 1.375$.

3.2, from section 3.7.2, we can see an example of what such an exploration might look like.

One final practical note on the subject of the selection of m is based on the fact that it is independent of units. Since it means the same at all scales, one can think that reasonable values for m should be between 1.1 and 15. If m is less than 1.1 then there is very little difference between the geometries of FATSO and LASSO, whereas if m is greater than 15 one is hardly performing any variable selection at all.

3.5.1 λ and prior conditional variance in Gaussian FATSO

We now have a handle on m but Gaussian FATSO has a second parameter λ. If $\lambda \to 0$ then we end up with a flat prior which may be suitable to variable estimation without any selection. On the other hand if $\lambda \to \infty$ then the prior will be concentrated around zero. This yields higher selectivity, but also shrinks the value of all estimations.

One way to think about selecting λ is to think of the FATSO less as an operator and more as a prior distribution. We can then study the properties of FATSO as a probability distribution, in which case λ may be interpreted as related to the variance of this distribution. Following the geometry from figure 3.7 we have the next lemma.

For a Gaussian FATSO, the prior distribution of $(\beta_i | \beta_j = 0 \forall j \neq i)$ is a Gaussian random variable with mean zero and (prior)variance

$$\lambda^{-1}\sqrt{2}sin\left(\theta - \frac{\pi}{4}\right)$$

where θ is the angle as shown in figures 3.6 and 3.7.

Proof. We note from figure 3.7 and Tales's theorem, that the ratio of AC to BC is the same regardless of how far C is along the horizontal axis. Then we have the relationship

$$\lambda||C - A||^2 = \lambda k||C - B||^2$$

where the left hand side differs by a constant from the value of the FATSO prior log-density calculated at the point C and the right hand side differs by a constant from a Gaussian log-density calculated at the same location (B is the origin).

Some trigonometry will then yield the value of $k = \sqrt{2}sin(\theta - \frac{\pi}{4})$, which proves the claim. □

When $\theta \to \frac{\pi}{4}$ then the variance goes to zero, and the geometry of FATSO approaches the geometry of LASSO.

Of note, if θ is close to $\frac{\pi}{4}$ then $sin\left(\theta - \frac{\pi}{4}\right)$ can become very small, and as a result m will have an effect on shrinkage of estimators unless λ is adjusted to compensate. This is not a very significant issue unless $m < 1.1$

If FATSO will be used for Bayesian analysis, this shows the effect of λ on the FATSO prior. The conditional variance of one parameter given all others are zero is a reasonable way to establish prior variability. λ should be selected accordingly.

Departing from a full Bayesian prior statement, one reasonable way to select λ is to use data-driven techniques such as cross-validation, but these may come at a significant computational cost, or the sample size may be too small for cross-validation to be a reasonable choice.

If we want to set λ using heuristics, we will turn to the observed data Y for some guidance. Note that if only β_i is active and all others are equal to zero, then if we write X_i as the ith column of X we have

$$Y = X_i\beta_i + \epsilon.$$

Now, using the Bayesian interpretation (even if we are not going to perform Bayesian inference), we can think of β_i as a random variable (a priori independent of ϵ) and write

$$var(Y) - \sigma^2 = var(\beta_i)X_i^T X_i,$$

and here we use the fact that X was standardized so that $\sum_j X_{j,i}^2 = 1$.

$var(Y)$ is not known, but it can be estimated with the sample variance $var(Y) \approx \sum \frac{(Y_i - \bar{Y})^2}{n}$. Hence, if $var(\epsilon)$ is known we can calculate one choice for λ as follows

$$\lambda = \sqrt{2}sin\left(\theta - \frac{\pi}{4}\right)[var(Y) - \sigma^2]^{-1}$$

However, it is worth observing that in practice, the value of λ has a relatively small effect on point estimates, as will be seen empirically in the results section of this chapter. λ acts more as an on/off switch than a dial, and hence it is not too important to worry about its exact value. One has only to find something in the (usually very large) reasonable range. The above method for selecting λ is not meant to be taken as a precise value, but merely to give a notion of where the reasonable range might be.

A second option, as is done in LASSO, mentioned in section 3.3 is to parameterize not with λ but with k/σ^2, yielding estimates which no longer depend on σ. Of course, this comes at the cost of being able to use knowledge of σ in order to select the parameter, as was done above with λ. Even if we prefer to use λ, however, this shows us that we can scale λ with the inverse of the standard deviation of the noise to achieve similar results.

3.6 Comparison to other LASSO extensions

FATSO is not the first attempt to extend the ideas of LASSO in a new direction. There are several other regularizations which have been attempted and which yield different benefits. We make no claims that FATSO is necessarily any better than any of these, but only that the issues it aims to address are different.

3.6.1 Ridge and Bridge regression

Ridge regression, also known as *Tikhonov regularization* in inverse problems (Fox et al., 2013) and is an older idea than LASSO. It is also closely related to the use of Gaussian priors in Bayesian regression. It essentially aims to estimate the regression coefficients with

$$\hat{\beta} = argmax_\beta \left[\mathcal{L}(\beta, X, Y) - \lambda \sum \beta_i^2 \right],$$

where $\lambda \sum \beta_i^2$ is the Ridge operator. One idea which places LASSO at one end and Ridge regression at the other is called Bridge regression, which changes the operator to $\lambda \sum |\beta_i|^\alpha$ for $\alpha \in (1, 2)$. Of note, however, for any value of $\alpha > 1$ the slope of the level curves at 0 is exactly zero. Hence, Ridge and Bridge are not selection operators in the sense that the resulting estimators are not zero (Hoerl and Kennard, 1970; Park and Yoon, 2011).

3.6.2 Group LASSO

One common extension to LASSO is the group LASSO, which separates the columns of X into groups and which promotes the selection of groups of variables together. While this does extend the ability of LASSO to handle more complex situations, it also requires some degree of understanding of the relationships between covariates, which is not the goal of FATSO. In another sense, however, group LASSO is more closely related to FATSO than the other LASSO

extensions since it aims to incorporate information about parameter grouping that is not immediately visible in the data but which is understood by the user (Yuan and Lin, 2006).

3.6.3 Scale mixtures of Normals

In recent Bayesian literature, there has been an explosion of selection operators proposed with the theme of corresponding to priors which are scale mixtures of normals (Carvalho et al., 2010; Bhattacharya et al., 2014; Liang et al., 2008). A scale mixture of normals is a random variable X which can be represented as $X = Y\sigma$ where Y is a random variable with a standard normal distribution and σ is some other (continuous or discrete) random variable (West, 1987). LASSO itself is closely related to this family, since it corresponds to a Laplace prior and a single variate Laplace prior is a scale mixture of normals with σ a Gamma distributed random variable. There are various motivations for the proposed operators, but they generally are focussed on some form of asymptotic convergence either of the entire posterior distribution or of some point estimate derived from it. We are unaware of any which ease the interpretation of tuning parameters.

3.6.4 Elastic net

The idea with the most similar behavior to FATSO is the elastic net. The elastic net uses as a regularization operator $\lambda_1 \sum |\beta_i| + \lambda_2 \sum \beta_i^2$ (and then applies a correction to the estimator), essentially working as a linear combination of the Ridge and LASSO operators. The first thing to note about the elastic net operator is that the level curves are not concentric, and the slope of the curves' intersection with the axes depends on the curve. For distant curves, the geometry of Ridge is dominant, whereas with curves closer to the origin the geometry is closer to that of LASSO.

While elastic net does not maintain the concentric level curves of FATSO, it does allow for variable selection with less stringent selectivity than LASSO, so it behaves in a similar way. In elastic net, however, the degree of selectivity is moderated very obscurely by the interplay of λ_1 and λ_2. The common recommendation is to select both parameters by data-driven techniques, such as cross-validation. This is a valid approach, but does not allow users to make informed decisions about the desired degree of selectivity based on their own expertise. Given that the $p > n$ scenario is one where data is known to have very little information, the goal of allowing human knowledge to participate is very sensible.

While FATSO is in no way intended to replace the elastic net, it is worth noting that the two main issues with LASSO which the elastic net aims to solve are both addressed by FATSO as well. The first of these issues is that in $p > n$ situations, LASSO cannot select more than n variables, and in the following section we will see an example where FATSO selects more than n covariates. The second issue is that when several covariates are highly correlated LASSO

Table 3.1: Results of estimations using the FATSO operator using various values for ρ and λ. We see that ρ affects selectivity smoothly, while adjusting λ does not allow for very flexible tuning.

m	ρ	λ	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	Other βs
500	353.6	0.1	0.697	0.329	0.543	Many are of similar order of magnitude
30	21.21	30	0.709	0.336	0.548	$\lvert\beta_6\rvert, \lvert\beta_7\rvert, \lvert\beta_{12}\rvert, \lvert\beta_{16}\rvert, \lvert\beta_{20}\rvert$ also active
3	2.236	30	0.814	0.397	0.602	β_7 and β_{17} are active
2	1.581	30	0.871	0.464	0.671	Others inactive
2	1.581	0.01	0.857	0.442	0.65	β_{17} active
2	1.581	1	0.868	0.459	0.69	Others inactive
2	1.581	100	0.868	0.46	0.658	Others inactive
2	1.581	1000	0.71	0.249	0.437	Others inactive

tends to select only one of them. It is proven in the original elastic net paper (Zou and Hastie, 2005) that any strictly convex regularization will solve this issue and FATSO is strictly convex.

FATSO does not aim to compete with the elastic net in terms of computational tractability or in terms of asymptotic error reduction, so while the behavior of the two operators is somewhat similar, their ultimate objectives are different.

3.7 Numerical results of FATSO

3.7.1 Simulated Data

15 observations of a 20 dimensional regression problem were simulated. The true values of the regressors β were zeros except for three variables. These variables were $\beta_1 = 0.9$, $\beta_2 = 0.5$, and $\beta_3 = 0.7$ (noise standard error was 0.1). Using these same data, FATSO estimates were calculated using different values of ρ and λ. Table 3.1 shows maximum a posteriori estimates of these data based on various values of ρ and λ using Gaussian FATSO.

With a high value for m and a low value for λ, the FATSO prior is nearly flat. In these situations the estimator is nearly the MLE, and since the dimension is greater than the sample size, the MLE does not give any real information about the parameters. As FATSO becomes more informative, so do the estimations. Similarly, we note the effect of λ and m act independently. We note that the effect of λ acts almost as if it had a threshold. For a fixed m of 2, then for any $\lambda \geq 1$ the exact value does not seem to have very much effect. λ at 1 and at 100 both yield very similar estimations for the parameters, and it is not until λ is extremely large (1000) that the effect of shrinkage becomes noticeable. With very small λ, however, the effect is lost somewhat (the extreme case being $\lambda = 0$

where we are left with the MLE again). The same cannot be said for m (or ρ), which affects estimations much more fluidly. We see that with a fixed λ of 30, the effect of m on selectivity is very clear. As m approaches 1 the selection is more strict, and as m grows then selection is looser. This confirms that the parameter m permits tuning the degree of selectivity in a fluid way that is not possible with a shrinkage parameter alone.

It is very tempting to be seduced by good results with $m = 2$ and reasonably high λ since the estimations are so close to the truth, but we must remember that these are synthetic data. The estimates with higher values of m are also estimates that could easily produce the same dataset, but in this particular case did not. When the dimension of the parameter space is larger than the sample size, then the data will be in the column space of the design matrix. The choice of one set of estimators over another is not information that is really in the data at all. With lower m, FATSO will tend to choose smaller sets of covariates which explain the data, but whether that is desirable or not is really an issue for the user to judge.

In order to illustrate this case, data was simulated with 17 nonzero variables rather than 3. It is known that LASSO selects at most as many variables as the sample size, so using LASSO here will select for at most 15 of the 17 active variables. The variables that were zero were β_1, β_{11} and β_{19} and inference was performed in the same way. With $\rho = 1.2$ and $\lambda = 100$ (a highly selective combination with a geometry similar to LASSO) we estimate 6 inactive variables. These are β_4, β_6, β_{10}, β_{11}, β_{12}, and β_{19}. As we see, not only are the variables being selected more strictly, but the variable choice is simply wrong. This is caused by the insistence on a high level of selectivity. With $\rho = 3$ and $\lambda = 100$ we estimate two inactive βs, and these are β_{11} and β_{19}. In this case the random simulation turns out to be unusually highly correlated with β_1 simply by chance, so β_1 was not selected against. We note that specifying too stringent selection criteria forces the model to shift away from the true values of the parameters, but allowing more nonzero entries yields a very good selection of variables. The difference between this situation and the last one is subtle, and it may often be a good idea to make the selection based on human understanding of the situation rather than on data which is necessarily insufficient.

3.7.2 Real Data

We use data from Stamey et al. (1989) to evaluate the performance of FATSO and the effect of parameter adjustment. These data are often used for LASSO demonstrations. The data are a 9 column matrix which describe prostate cancer data in 97 patients. The first 8 columns describe characteristics of the tumor and the last column is the response variable: An antigen. The data is available in the R package *lasso2* (Lokhorst et al., 2014)

Since the data was sorted by the response variable, the rows of the matrix were permuted randomly. There are 97 rows; estimation was done using the first 67 rows and used to estimate the remaining 30. Table 3.2 displays the results of inference on this collection of data using different values for the parameters.

Table 3.2: Results of estimations using the FATSO operator on the prostate cancer data. Once again we see that ρ can be adjusted to tune the degree of selectivity with some reasonable degree of control.

m	λ	active βs	MSE	Observations
100	0.0001	all except β_8	0.60516	Almost exactly simple linear regression
5	0.5	all except β_8	0.61363	
3	1	$\beta_1, \beta_2, \beta_4, \beta_5, \beta_6, \beta_7$	0.63189	Removing one variable does not greatly increase the error
2	1	$\beta_1, \beta_2, \beta_4, \beta_5, \beta_6, \beta_7$	0.65002	
1.2	1	$\beta_1, \beta_3, \beta_6$	0.78280	With a higher error we can remove many more

The main takeaway from this experiment is the fluid way in which we can select the β_is. Since this is not a high dimensional problem, the MLE is a good estimator, but by tuning m we can pick a simplified model which selects more or fewer variables. Removing variables comes at a cost, but we can see exactly how costly this removal is. Using this information it is possible to manually tune our model to whatever balance of parsimony and accuracy we want. Hence, even in this relatively low dimensional scenario, there is something to be gained by having a fluid selection operator.

3.8 Use of FATSO on breath test data

Once FATSO has been developed, we proceed to use it on the breath test data. Our first objective is to have an appropriate single dimensional marker for breath test data to use as the response variable. One choice is to use our clinical results which classify patients into 3 categories (healthy, insulin resistant and severe diabetic), but although this clinical diagnosis is probably our most reliable data, it is categorical. For a linear model, it is preferable to use data which varies on a continuous scale. We have complete OGTT data for 30 of the 35 patients for whom there is breath data available. The remaining 5 patients had severe health complications, so performing an OGTT test was considered dangerous. These patients were instead placed in emergency care.

The dynamic model was used to fit curves to each of the patients for whom complete OGTT data is available. The resulting curves were compared with the clinical diagnosis for each patient, and it was found that the geometry of the curves matched the diagnosis well for all patients. We therefore consider it reasonable to use the OGTT data to construct a linear model.

The next task is to select a single dimensional response which represents the risk of diabetes for each patient. There are many plausible choices, but one

reasonable option is

$$Y_i = min\{t : G(t) = G_b \,\text{and}\, G'(t) < 0\}$$

for patient i, which is the option that was chosen. This datum was estimated for each patient using the MAP estimator.

For the matrix X of metabolite counts, some minor preprocessing was also required. For some patients, the breath test had been performed more than once, and for these patients the average of the counts for each metabolite was used. Also, several entries were found to be missing, but after verifying with the team in charge of breath tests, these missing data are interpreted to be indications that the metabolite was not found in the breath test, and hence they are set at 0.

Regression was performed using FATSO, with high, low, and intermediate degrees of selectivity. Even with high levels of selectivity, at least 18 explanatory variables were selected. These 18 variables were compared with the variables that are most active in the principal component analysis from section 3.1 and were found not to match. Although this does not necessarily invalidate the idea that breath tests may be reasonable predictors for diabetes, it does reduce the value of the principal component analysis as evidence thereof.

3.9 Conclusions

The results of using FATSO on breath test data are not particularly promising. While they are not sufficient to completely rule out breath tests as a way to diagnose diabetes, they significantly reduce the credibility of the initial principal component analysis as evidence thereof. This aforementioned result from principal component analysis is the main piece of evidence in favor of considering breath tests as an alternative test for diabetes, and therefore it seems unreasonable to pursue breath tests any further. Instead, we will center our attention back on OGTT tests, for which we have more promising results.

The above notwithstanding, the methodology for variable selection that was developed has a scope which reaches far beyond any applications to diabetes.

In situations with high dimensional data, where $p > n$, there are infinitely many parameter combinations which might yield the observed data. In these situations, the data does not clearly favor one choice of parameters over another, so in order to make a selection, some measure of human choice is required. LASSO and other similar regularization operators are means by which a form of preference is given to one kind of solution over another. These systems all have parameters which affect – in some sense – how this choice is made. The selection of these parameters by data-driven techniques is appealing, but the information to make the choice is not really in the data. As a result, it becomes desirable to understand the meaning of the parameters and the effect of their choice on the resulting inference. This problem is particularly serious in the Bayesian setting since the operators correspond to prior distributions and it is invalid to assign priors using the data that these priors are chosen to analyze.

This last issue is not a vague or theoretical one since both LASSO and Elastic Net have been adapted for Bayesian inference regardless of the difficulty in assigning parameters (Park and Casella, 2008; Li and Lin, 2010).

While significant effort has been made to improve the data-driven techniques for adjusting parameters, this effort has done little in the sense of improving the interpretability of the parameters for human users who have additional information. In this sense, the elastic net is the system which boasts the lowest mean squared error for theoretical purposes, but it is also gives perhaps the least interpretable combination of parameters.

FATSO is an attempt to offer a means of setting the degree of selectivity by hand. While it is theoretically possible to use data-driven techniques to assign m, if data driven techniques are preferred, then one would probably be better served using another regularization operator. On the other hand, in situations where one intends to choose the degree of selectivity using outside knowledge, FATSO is recommended to set the selectivity in a way that is understandable and meaningful.

Chapter 4

Design of experiments

4.1 Improving OGTT tests

As seen in chapter 2, The use of a dynamic model to analyze the results of OGTT tests represents a significant potential improvement over the current guidelines since it attempts to describe how the patient's body handles the ingested glucose over the duration of the test. The use of this ODE model may help to improve the analysis of OGTT data.

We have also investigated an alternative test which was suggested, and our study found that this test is not a promising avenue of study. As a result, we have reaffirmed our interest in OGTT tests as a means for diagnosis of type 2 diabetes. This leads us to continue to investigate OGTT tests directly.

Our dynamic model has improved our analysis of the results of the test, but the test itself remains, for the most part, unaltered. A natural question is whether the model yields further information on the test itself and, if so, whether this information can be used to improve it. There are two main kinds of changes that can be considered. The first is a change to the testing protocol, altering none of the related infrastructure or equipment, and the second is a change to the equipment. In this chapter we will focus on the first kind of change.

In particular, the issue we are interested in is the set of times t_i at which glucose is measured. At present, these are assigned arbitrarily, according to local practices, rather than systematically, to make the most of the information gained by the test. These times vary from location to location, but common practice is to measure glucose at $t_1 = 0$ (arrival), $t_2 = 1$ and $t_3 = 2$ hours. Throughout this thesis we use data taken at more frequent intervals (up to every 15 min in section 4.5.1). These more frequent measurements are uncommon, but are occasionally taken for research purposes (OGTTs performed by our medical collaborator).

The selection of measurement times is not a trivial issue. Of note, it is not difficult to come up with *bad* sets of times at which to measure. For instance,

if we look back to figure 2.1 we can see that the red and blue lines cross around time $t = 0:30$ hours and that they converge again around time $t = 3:00$ hours. If we only measure at times $t_1 = 0:00$, $t_2 = 0:30$ and $t_3 = 3:00$ then it will not be possible to distinguish between these two significantly different scenarios. It is also worth noting that, without referring back to the curves as displayed, this set of times is *not* obviously problematic.

It is clear that any testing scheme can be improved by increasing the number of times at which a measurement is taken (the patients from figure 2.2 had samples taken at $t =$ 0:00, 0:30, 1:00, 1:30, 2:00 hours) but it would be desirable to know how many times are enough as well as what times these should be.

Selecting the correct times is a problem of experimental design. Since the model is fit to data using Bayesian inference, We will use Bayesian experimental design to propose a new set of times with the aim of improving the test. In this chapter we will investigate the theory of Bayesian experimental design, and look into the previously proposed techniques for this purpose. All of these techniques tend to share one significant flaw, which is that they make strong demands on the structure of the design problem or on the geometry of the utility function. Since we cannot assure that our particular design problem meets these demands, we will propose a method for experimental design which works to compare designs pairwise and which does not make these assumptions. A suitable algorithm is also proposed to implement this method in practice. The method is not particularly different from previously explored methods (Christen and Buck, 1998), but significant effort is made to improve the theoretical backing of this method.

Another issue at play is the choice of priors for the design problem. Section 4.4 explores the pragmatic issues relating to the choice of priors and explains the unusual choice which is made for this work. While the particular choice we make here may seem odd, we believe that the motivation for our choice represents a common issue and that the choice we make may be thought of as a reasonable option in other situations as well.

The new design problem for OGTT tests is posed, and the method we developed is used to make a recommendation about a good set of times to get high quality information from OGTT data. Several numerical experiments are then done to validate the design choice.

This chapter is organized as follows: Section 4.2 presents the theory of Bayesian experimental design. In section 4.3 we offer a new algorithm to decide between designs. Section 4.4 discusses the issue of prior selection for diagnosis of diabetes using OGTT tests, particularly in the context of experimental design. Finally, section 4.5 presents the results of using the algorithm to study the problem of design for OGTT tests.

4.2 Experimental design

Most of statistics is concerned with inference from collected data, ignoring the issue of how the data is collected to begin with. This is, overall, a significant

omission, since the quality of inference typically depends heavily on the quality of the data. Experimental design is the use of statistical techniques to improve the quality of the data that will be collected, and thus improve the quality of the resulting inference (Berger, 1993).

Unlike most areas of statistics, experimental design is concerned with what happens *before* any data becomes available. In Bayesian statistics in particular, the information available before the presence of data is encoded in the prior distribution. While Bayesian inference is concerned with the results of studying the interaction of the prior distribution and data, experimental design is primarily interested in studying the properties of the prior distribution itself, considering what it says regarding the data that might be obtained when it is actually collected.

In the case of OGTT tests, a design is $d = (t_1, \dots t_n)$, the times at which blood samples are drawn for testing.

For a given design d, it is common practice to write $\pi(y|\theta, d)$ as the likelihood function for the data y and the parameter θ when using the design d, see for example Huan and Marzouk (2014) for one case where this is used. This is not actually a conditional probability in the strict sense of the word: d is not a random variable, and $\pi(d)$ does not exist. This practice is therefore notational abuse. It is, however, standard, and here we conform to this notation.

4.2.1 The main idea: Utility functions

Let d_1 and d_2 be designs which we want to compare. We define a utility function u, which assigns a value to the result of an experiment. In the most general sense, the utility is a functional from the space of posterior distributions to $\mathbb{R}$, however we need not worry about this representation since the posterior distribution is determined by finite dimensional data. It is therefore possible to write the utility as a function of data directly $u(y)$. Note that since $u(y)$ depends on random data, a priori it is itself a random variable.

A utility function may be, for instance, equal to minus the posterior variance of one component of the parameter, or minus some norm of the difference between the predictive distribution and the true distribution which generates data (assuming such a thing exists), etc. Another choice is to use the K-L divergence between the prior and posterior distributions, the idea being that the bigger the difference between these distributions, the more information was acquired from the data; see Huan and Marzouk (2014); Zhang (2006); Christen and Buck (1998); Chaloner and Larntz (1989); Anand et al. (2010); Gilmour and Trinca (2012); Alexanderian et al. (2016); Weaver et al. (2016); Solonen et al. (2012) for several examples of utility functions and Bayesian desing problems

We write $u(y|d)$ as the utility for the data y collected using an experimental design d. Our selection between d_1 and d_2 is based on which of these designs maximizes the expected utility given our prior distribution, namely

$$U(d|\pi) = \int u(y|d)\pi(y|d)dy.$$

The distribution under which the expectation is calculated is the predictive prior distribution of the data under the design d:

$$\pi(y|d) = \int \pi(y|\theta, d)\pi(\theta)d\theta.$$

While it is common to simply write $U(d)$, note that $U(d)$ is also dependent on the prior $\pi(\theta)$.

The main goal of experimental design is to find good designs, which means we want a design d such that $U(d)$ is as high as possible. Unfortunately, in our case, $U(d)$ is not tractable. The difficulty lies not only in calculating the integral $\int u(y|d)\pi(y|d)dy$ since even with specified data y, it is usually not possible to calculate $u(y)$ exactly.

The real difficulty in this experimental design, and a common issue in Bayesian experimental design in general, lies in optimizing a function which cannot be evaluated exactly. The approach we take here is to find good Monte Carlo estimators for $U(d)$ and use them for design comparisons on a comprehensive discrete grid of possible designs. The optimization is then taken, not over the continuous time space, but only over a discrete space of 15 min intervals, proceeding by semi-brute force maximization. We explain the details of this approach in section 4.3 and onwards. Meanwhile, in the next section we explain briefly other common approaches for implementing Bayesian experimental designs and why these are not suitable for our design problem.

4.2.2 Other approaches

The number of published papers on Bayesian experimental design is very small relative to the amount of research done in Bayesian statistics and also very small relative to the amount of research done in experimental design in general. Nonetheless, some techniques have been proposed to optimize design parameters. All of these approaches are based on the idea of finding

$$d^* = argmax_d U(d)$$

using some kind of optimization algorithm.

Most traditional optimization techniques are not useful with a function as poorly understood as $U(d)$ (classical optimization techniques have been attempted using random estimations of $U(d)$, Anand et al., 2010, but this does not have adequate theoretical justification), so specialized approaches must be taken. There are two main ideas to try to circumvent this problem.

1. **Asymptotic estimations of $U(d)$:** A well known result of Bayesian statistics states that under certain regularity conditions, the posterior distribution approaches a Gaussian as sample size goes to infinity. If we assume that the sample size is large, then for certain utility functions it is possible to calculate the asymptotic value of $U(d)$. Then $U(d)$ is optimized in the asymptotic regime (Zhang, 2006; Chaloner and Larntz, 1989; Gilmour and Trinca, 2012).

While this approach may be reasonable under some circumstances, it is worth noting that the entire problem of experimental design is most interesting precisely when sample sizes are *small*. Under most circumstances, increasing the sample size is an easy way to improve a design, and the need for a well-designed experiment arises only when circumstances indicate that large sample sizes are impossible to begin with.

2. **Stochastic approximation:** There is a class of numerical techniques called *stochastic approximation* techniques which deal with functions that cannot be directly measured. Some of these techniques are used for optimization, and these have been used to attempt to optimize $U(d)$ without requiring a large sample size. By far the most commonly used of these techniques is the Robbins-Monro algorithm (Robbins and Monro, 1951; Huan and Marzouk, 2014; Duflo, 1997).

   Most stochastic approximation techniques (including Robbins-Monro) require an unbiased estimator of the gradient $\frac{\delta U}{\delta d}$. With certain utility functions it is possible to obtain this estimator, but this is not universal.

   Although there are some derivative free stochastic approximation techniques (Duflo, 1997), a more serious issue than the requirement of gradients is the fact that all of these algorithms only perform local optimization. In fact, to our knowledge, the convergence of stochastic optimizers has only been proven for strongly convex functions (Duflo, 1997). If $U(d)$ is a well-behaved strongly convex function then this is not an issue, but in our case we have no reason to believe that our expected utility $U(d)$ belongs to such a class.

   Furthermore, stochastic approximation algorithms are based on simulations and estimations. They are therefore subject to error based simply on the randomness of the estimators. While avoiding this kind of error altogether is impossible, it would be extremely desirable to control - or at least quantify - the uncertainty in our eventual conclusions as a result of these errors.

Recently an alternative has been proposed which approximates the utility function using Gaussian processes. This estimation is not asymptotic with sample size and may be a more robust alternative to asymptotic estimations (Alexanderian et al., 2016; Weaver et al., 2016).

We propose a different alternative which does not have these problems, but which suffers from a different set of limitations. Rather than attempting to find an optimum design, we simply propose a good way to decide between any two designs, and then perform many comparisons, optimizing by semi brute force. This is generally not a good technique for optimization, but in this case its use is warranted since it allow us to perform comparisons that do not depend on sample size or any special properties of the function U, to avoid bad local maxima, and also to control and quantify the uncertainty in our conclusions. While we may not be able to reach any definite conclusion about what design

is actually the best, we will be able to achieve arbitrarily high confidence in our claims regarding what designs are good.

4.2.3 An unusual generalization

We make a generalization to the usual scheme of experimental design, which is to allow for two different priors over the same parameter: One for design ($\pi_{\mathcal{D}}$) and one for inference ($\pi_{\mathcal{I}}$). This generalization is admittedly unusual: It is not clear that anyone would ever *want* to allow $\pi_{\mathcal{D}}$ and $\pi_{\mathcal{I}}$ to be different. For now we limit ourselves to indicate that this generalized problem is indeed well-defined. The motivation for this generalization is discussed in section 4.4, where we see how, in some situations, this may indeed be desirable.

Accordingly, let $\pi_{\mathcal{I}}(\theta)$ be the prior which is used for performing inference, and let $\pi_{\mathcal{D}}(\theta)$ be the prior used to design the experiment. If $\pi_{\mathcal{D}} = \pi_{\mathcal{I}}$ then we have the usual problem, as described above. The general function which we wish to optimize is $U(d|\pi_{\mathcal{D}})$ which can be written as

$$U(d|\pi_{\mathcal{D}}) = \int u_{\mathcal{I}}(y|d)\pi_{\mathcal{D}}(y|d)dy,$$

where the utility $u_{\mathcal{I}}(y|d)$ is calculated using the posterior distribution generated by the inference prior $\pi_{\mathcal{I}}$.

For the remainder of this chapter we will write $U(d|\pi_{\mathcal{D}})$ as simply $U(d)$ and $u_{\mathcal{I}}(y|d)$ as simply $u(y|d)$. A design will be selected in a way that works regardless of whether or not the priors are different.

4.3 The algorithm for design selection

4.3.1 Estimation of $U(d)$

We require some technique for estimating (but not necessarily directly calculating) $u(y|d)$ for any given y and d. Note that unless y is fixed, $u(y|d)$ depends on random data y, and is hence a random variable itself. Since u is typically a functional of the posterior distribution, we assume that we can estimate $u(y|d)$ with a posterior sample ϑ produced by some posterior sampling algorithm such as an MCMC chain. We call this estimator $\hat{u}(y|d)(\vartheta)$. Note that $\hat{u}$ depends on the random data y and also on the random posterior sample ϑ. We have two requirements on $\hat{u}$. The first is that it is unbiased for fixed d and y (this requirement can be relaxed to allow for asymptotically unbiased estimators, but this comes at a cost, see section 4.3.3), and second that it have finite second moment. In other words, that $\mathbb{E}_{(Y|d)_{\mathcal{D}}}[\hat{u}(y|d)^2] < \infty$. In our case, $u(y|d)$ is minus the mean squared error of $G(t)$ integrated from $t = 0$ to $t = 3$ hours. This can be estimated using Monte Carlo samples, as described in section 4.5

We now propose the following sampling algorithm for estimating $U(d)$ for a given d:

Fix two constants T_1 and T_2:
for *i from 1 to* T_1 **do**
  Sample data $y^{(i)}$ from the predictive prior $\pi_{\mathcal{D}}(Y|d)$;
  Generate a sample $\vartheta^{(i)} = \{\theta^{(i,j)} : j = 1\ldots, T_2\}$ from the posterior $\pi_{\mathcal{I}}(\theta|y^{(i)}, d)$;
  Calculate $\hat{u}_i = \hat{u}(y^{(i)}|d)(\vartheta^{(i)})$;
end
Calculate $\hat{U}(d) = \frac{1}{T_1}\sum_i \hat{u}_i$

We now calculate the expected value of our estimator $\mathbb{E}_{\mathcal{D}}(\hat{U}(d))$:

$$\begin{aligned}\mathbb{E}_{\mathcal{D}}(\hat{U}(d)) &= \frac{1}{T_1}\sum_i \mathbb{E}_{(Y|d)_{\mathcal{D}}}(\hat{u}_i)\\ &= \frac{1}{T_1}\sum_i \mathbb{E}_{(Y|d)_{\mathcal{D}}}\hat{u}(y_i|d)(\vartheta^{(i)}))\\ &= \mathbb{E}_{(Y|d)_{\mathcal{D}}}\hat{u}(y|d)(\vartheta)\end{aligned}$$

where ϑ is a random variable with the same distribution as any ϑ_i

Now we observe that since $\hat{u}(y^{(i)}|d)(\vartheta)$ is unbiased then

$$\mathbb{E}_{\mathcal{D}}(\hat{U}(d)) = U(d)$$

so $\hat{U}(d)$ is an unbiased estimator.

Moreover, observe that $\hat{U}(d)$ is an average of *iid* random variables, each of which is distributed as $\hat{u}(y|d)$, which has finite second moment. Hence, $\hat{U}(d)$ is subject to the central limit theorem, so as $T_1 \to \infty$ we have

$$\mathbb{P}\left(\frac{\hat{U}(d) - U(d)}{\sqrt{var(\hat{U}(d))}} < \alpha\right) \to \Phi(\alpha).$$

Now $var(\hat{U}(d)) = var(\frac{1}{T_1}\sum_i \hat{u}_i) = \frac{1}{T_1}var(\hat{u}_i)$. We can estimate $var(\hat{u}_i)$ with its sample variance, and arrive at a normal asymptotic distribution for $\hat{U}(d)$.

Similar estimators have been proposed in the past (Christen and Buck, 1998; Anand et al., 2010) but the properties of the estimators (such as their asymptotic distribution) were not studied. In the following section we explain how this distribution can be used to quantify uncertainties in comparisons between designs.

4.3.2 Numerically deciding between d_1 and d_2

Now that we are able to estimate $U(d_1)$ and $U(d_2)$ the simplest idea is to select the design with the higher estimator. This is not satisfactory, however,

unless we have a proper way of controlling the uncertainty in this choice. Since our estimators depend on random simulations, we want to be certain that the difference between these estimators corresponds to an actual difference in the expected utility of the experiments and is not solely the result of the random nature of the estimators.

We have an asymptotic distribution for $U(\hat{d}_1)$ and $U(\hat{d}_2)$. Hence, so long as T_1 is sufficiently large, we can consider the comparison of two expected utilities to be a comparison of the means of two normally distributed random variables with known variance. (Technically the variance is unknown, but if T_1 is large enough this is not a problem. Theoretically, errors in variance estimation can be handled using a t statistic rather than a normal statistic, but the distribution of the statistic depends on the sample size. Furthermore, for large samples, the resulting t distribution is almost identical to a normal one anyway.) This is a well-studied classical problem.

Assume, with no loss of generality, that $U(d_1) \leq U(d_2)$. Now we fix a value $0 < \alpha < 1$ and we wish to make sure that the probability of wrongly concluding that $U(d_2) > U(d_1)$ is at most α. This can be done by considering the variable

$$Z = \frac{\hat{U}(d_1) - \hat{U}(d_2)}{\sqrt{var(\hat{U}(d_1)) + var(\hat{U}(d_2))}}.$$

Z is asymptotically normally distributed with mean 0 an variance 1 (DeGroot and Schervish, 2011), so we can conclude that $U(d_1) < U(d_2)$ if Z is less than the $\alpha/2$ quantile of a standard normal distribution.

It is still possible that this problem will not be completely solved since testing for $U(d_1) < U(d_2)$ and also testing for $U(d_2) < U(d_1)$ may both produce inconclusive results. This does not necessarily mean that the two designs are of equal (or even of approximately equal) expected utility, but rather that the variance of our estimators is still too large to be able to choose with the required degree of certainty. In section 4.5 we discuss how this does not represent a problem in our case, although in some other situations it might become an issue.

If reaching decisive conclusions is required, then it is possible to increase the sample size and test again. This presents a problem; the probability of error when testing repeatedly is greater than the probability of error when testing once since the error could have been committed at any of the tests. However, it is possible to implement a sequential testing scheme in the style of the Sequential Probability Ratio Test (Wald, 1945). The classical form of the Sequential Probability Ratio Test requires knowledge of the power of the test, which is unavailable in our situation, but it can be modified slightly to work in this situation as well. We have explored some implementations of this idea, but it is not yet clear how to accomplish this task efficiently.

4.3.3 How the choice of T_1 and T_2 affects estimation

Note that when we perform sequential testing, the way we reduce the variance of our estimator is to increase T_1, but it is also possible to reduce the variance by increasing T_2. However, T_1 and T_2 have very different effects on the distribution of $\hat{U}(d)$.

Our first observation is that increasing T_2 only reduces the variance of $\hat{u}(y^{(i)}|d)(\vartheta^{(i)})$ by increasing the size of the sample $\vartheta^{(i)}$, but even if we were able to calculate $u(y^{(i)}|d)$ exactly for each i, that still will not reduce $var(\hat{U}(d))$ to zero, since $y^{(i)}$ is still random. In other words, T_1 absolutely must be increased to assure that one of the models is eventually selected. Increasing T_2, however is not strictly required. $var(\hat{U}(d)) \to 0$ is assured so long as $T_1 \to \infty$

Proof. $var(\hat{U}(d)) = var\left(\frac{1}{T_1}\sum_i \hat{u}_i\right) = \frac{var(\hat{u}_1)}{T_1} \to 0$ □

That is, remembering what T_1 and T_2 are, increasing the number T_2 of (MCMC) samples for each posterior given a simulated sample does not assure that our estimator of the design utility $\hat{U}(d)$ tends to zero. On the contrary, only the number of simulated samples T_1 for the design d needs to increase and T_2 could be kept fixed, and possibly low, as we discuss next.

The second observation is that if T_2 is unchanged then it is possible to continue the algorithm, drawing more samples from the predictive prior. These can be used to increase T_1 without discarding the previous sample. It is not possible to do this if we attempt to increase T_2 for the new sample points since altering T_2 changes the sample size from which $\hat{u}(y^{(i)}|d)$ is calculated and therefore alters the distribution of the estimator. These simple and useful observations also apply to many similar algorithms but were overlooked by previous authors (Christen and Buck, 1998; Anand et al., 2010).

In general, the effect of T_1 and T_2 to reduce $var(\hat{U}(d_k))$ depends heavily on the loss function and the model, but the previous two observations make it seem reasonable to suppose that it is a good idea to have T_2 be "fast" (of course, it must be a bare minimum large enough to obtain an unbiased estimator $\hat{u}(y|d)(\vartheta)$) and allow T_1 to increase dynamically.

Note that if $\hat{u}(y|d)(\vartheta)$ is asymptotically unbiased – rather than unbiased for finite sample size – then this does not work equally well. The central limit theorem only states

$$\mathbb{P}\left(\frac{\hat{U}(d) - \mathbb{E}(\hat{U}(d))}{\sqrt{var(\hat{U}(d))}} < \alpha\right) \to \Phi(\alpha)$$

where for unbiased estimators $\mathbb{E}(\hat{U}(d))$ can be replaced by $U(d)$. For asymptotically unbiased estimators, the usefulness of the approximation depends on the quality of the approximation $\mathbb{E}(\hat{U}(d)) \approx U(d)$. This in turn depends on $\mathbb{E}_{(Y|d)_{\mathcal{D}}}(\hat{u}_i) \approx u(y^{(i)})$, and the quality of this approximation depends on the sample which is used to calculate it. That sample is of size T_2. Hence, if we intend to use an asymptotically unbiased estimator for the utility, then the quality

of our estimation depends on T_2. This is a strong reason to prefer an unbiased estimator if one is available.

4.3.4 Special considerations for MCMC type samplers

The aforementioned method for hypothesis testing does not depend on the technique used to obtain a posterior sample, but in practice the most common method is the use of MCMC algorithms such as the Gibbs Sampler or the Metropolis-Hastings algorithm.

There are two issues which are of particular interest when using MCMC for sampling. The first is the issue of obtaining a proper estimator $\hat{u}(y|d)$. It is worth noting that proximal iterations of an MCMC chain are usually strongly correlated. There has been much debate as to whether an MCMC sample should be "thinned" by taking only one iteration every so often (to avoid correlation of proximal iterations) in the chain for posterior inference or if it is OK to treat the full chain as the posterior sample of interest.

The answer to the thinning question in general depends on what information is desired from the posterior. In this particular case, what is needed is an unbiased estimator for $u(y|d)$. Common cases of unbiased estimators require an *iid* sample, and hence, most of the time the MCMC chain *must* be thinned.

The second issue of interest in an MCMC algorithm relates to burn-in times. When running an MCMC algorithm there are two parameters of note which affect the running time for the posterior estimation: These are the autocorrelation time and the burn-in time. When the reason to generate a posterior sample is to perform inference, the time which is most important to reduce is autocorrelation time, since for a size m sample the algorithm must run through the autocorrelation time $m-1$ times, and the burn-in time only once.

For this form of experimental design, however, large burn-in times can also be very problematic since an MCMC chain must be run T_1 times to obtain an estimator of $U(d)$ for a single design. If burn-in times are significant then this can be a problem. Luckily in this situation it is possible to start the MCMC chain close to regions of high posterior probability since the parameters used to generate the sample of the predictive prior are known (the data were simulated; see the algorithm in section 4.3). Since the chain can be started immediately at the true values of the parameters, the burn in time is all but eliminated; the only exceptions being rare cases where the data is very unusual for the parameters which generated it.

4.4 Selecting $\pi_{\mathcal{I}}$ and $\pi_{\mathcal{D}}$

Having developed a tool to compare designs, we return to the problem at hand: Improving the design of OGTT tests.

We have discussed, in general, inference on OGTT data but we have as yet to fix the joint prior distribution that is to be used for θ_0, θ_1, θ_2, and $G(0)$

We do not want misdiagnosed patients and we must make the best of available data to provide our inferences. An added difficulty is that the sample sizes involved are quite small. Testing repeated blood samples from a patient requires a significant amount of work from the laboratory staff, and requiring them to test a large number of blood samples is not reasonable (this may sometimes cause increased discomfort to the patient as well, although this is rare since the most common practice is to use a cannula). In practice the typical sample size is 3, although in some special research cases it may go up to 9. Consequently, priors must be chosen with a small sample size in mind. In particular, when performing inference, an informative prior is likely to overwhelm the data, and may lead to a diagnosis that is based mostly on the prior, rather than on the sample.

Assigning a prior distribution for inference which will serve for any patient is difficult. To avoid misdiagnosis, we must resort to a relatively vague prior. With this in mind, the priors chosen for $\pi_{\mathcal{I}}$ are the same ones that were chosen in chapter 2:

$$\begin{aligned}
\theta_0 &\sim Gamma(2,1) \\
\theta_1 &\sim Gamma(2,1) \\
\theta_2 &\sim Gamma(10,1/20)\ \mathbb{1}\{\theta_2 > 0.16\} \\
G(0) &\sim \mathcal{N}(80,10000)\ \mathbb{1}\{G(0) \in [30,400]\}.
\end{aligned}$$

We consider these priors to be vague since their regions of high probability extend well beyond any estimations performed with real patients. θ_2 has been truncated for mathematical reasons (if θ_2 is too small, then from the system of ODEs in section 2.2, in (2.4) and (2.5), it will be possible for the glucose in the digestive system to begin with negative derivative, which is nonsense; see Christen et al., 2016) and $G(0)$ was truncated based on practical considerations: Any patient with an initial measurement anywhere near or below 30 or above $400d/mL$ will not be tested but instead will be placed into emergency care (a preliminary, instant, fingerstick blood test is conducted, for removal and immediate treatment of such cases).

π_I may be seen as an inadequate representation of our actual prior uncertainty, but using anything more informative can result in misdiagnosis of patients with unexpected glucose curves. Since this prior is needed to analyze data arising from all patients, we must then settle for this relatively vague prior.

Now, if we set $\pi_{\mathcal{D}}$ equal to this vague π_I our predictive prior will assign significant probability to regions that are not actually very likely scenarios. Our chosen design will therefore be tuned to take into consideration common situations as well as situations that occur infrequently, if they do. Our inference prior (π_I) was chosen for pragmatic reasons rather than based on an actual reflection of our uncertainty. For similar pragmatic reasons, it is not reasonable to use the same prior for design.

Moreover, as opposed to an informative inference prior, we do not expect the experimental design to have such a severe impact on misdiagnosis (this should

be tested of course, but we know that OGTTs have been used successfully with a poor design for years), so we consider it less dangerous to use a more informative design prior.

Choosing a prior for design is also a difficult issue. One practical alternative is simply to pick a prior which represents the available data reasonably, use it to select a good design, and then compare it to arbitrary designs. There are several ways to pick the data in order to make comparisons, but one fair choice is to generate the data from $\pi_{\mathcal{I}}$. If the design appears to work well for data that is generated from the inference prior as well then we can conclude that this design is a good choice regardless of the prior that was used to generate it.

We propose using an *extremely* informative prior for design. We have taken a sample of patients which represent typical scenarios, and have set our design prior to represent those specific patients. Our prior distribution gives an equal probability to each of the parameter combinations of these exact patients, and zero probability to anything else. This prior, of course, is not an adequate representation of our prior uncertainty either. If the design that is obtained when using our highly informative design prior proves to be useful for other patients as well, then this extreme prior will have served its purpose. In section 4.5 we will carefully examine how robust our results are, and whether our design proves suitable for other patients.

The reader might be inclined to take this approach of using a different prior for design and inference purposes as perhaps eccentric or strange. However, similar approaches have been studied in the context of reference priors, where the priors used for inference are different from the priors used for model selection, even when the context is the same (Pericchi, 2005). Discussions of the use of different priors for design and inference - in different contexts - can also be found in Berry and Kadane (1997) and also in Stone (1969). Of note, the circumstances which lead us to the selection of different priors for inference and design are not actually very unusual; inference priors are often selected with high entropy in order to avoid overwhelming the information contained in the data, specially when dealing with small sample sizes. In such a case, the use of a different prior for design may be something to consider. The extreme case is seen when using improper (reference) priors for inference, wherein there is really no choice at all since design priors must be proper for $U(d)$ to be well-defined. In such a case, $\pi_{\mathcal{I}}$ and $\pi_{\mathcal{D}}$ *must* be different.

4.5 Implementation and Results

The algorithm in section 4.3 was used to select a design for OGTT diagnosis. In order to calculate $G(t)$, the forward map was solved numerically using the LSODA package for ordinary differential equations and an MCMC was used to sample from the posterior distribution using the t-walk package (Christen and Fox, 2010). The t-walk is an MCMC algorithm which is designed to adjust to continuous posterior distributions without tuning, which is particularly useful for our purposes since it means the MCMC does not have to be tuned separately

for each patient.

The utility function used was the negative mean squared error of $G(t)$, integrated over the curve from $t = 0:00$ to $t = 3:00$ hours. The choice was made so as not to attach significant preference to any particular time or parameter. The utility function is estimated by numerically estimating the integrated squared error for each element of a posterior sample and averaging across the estimators. One problem with this estimator is that it is only asymptotically unbiased rather than unbiased for finite sample size (see section 4.3.3) so a large T_2 was used.

The selection of a design was done somewhat crudely, only comparing designs chosen with times at 15 minute intervals over a 2 hour period. A finer tuned selection would be significantly more expensive computationally, and it is not clear that it would be of much practical use since health professionals might not be able to take measurements at times which are specified with great precision while also keeping up with their other duties.

In order to decide how many measurement times are required, comparisons are not done sequentially but allowed to be inconclusive if the decision cannot be made with a large T_1. "Large" in this case means 600. For such situations, where an experiment is to be performed several times, this number is interpretable; each element represents a simulated patient. If a decision cannot be reached with $T_1 = n$ then this means that no difference is detectable when performing an OGTT test over a sample of n patients. The number of measurements was deemed sufficient when adding another measurement resulted only in inconclusive comparisons.

The Python 2.7 programming language was used, running the t-walk MCMC algorithm for 1500 iterations for each patient. One such run takes between 5 and 10 seconds on an Intel processor running at 1.7GHz. To compare designs, a sample of 600 patients is taken for each, (unless one of these designs already has samples available from a previous comparison). One such comparison takes about 15 minutes. The full process is computationally intensive, but not unreasonably so, and can be parallelized for additional efficiency if needed. In this particular case, considering 15 min intervals only, the full process took roughly 6 hours.

The resulting selection of times is $t = 0:00, t = 0:45, t = 1:15, t = 1:45$ and $t = 2:00$ hours. In table 4.1 we see the times from our newly proposed design next to the times from the conventional design which measures every hour. We also see a "Full" design that is sometimes used for research purposes. It is not practical to use this design in general, but it is used for validation purposes in section 8.1.2

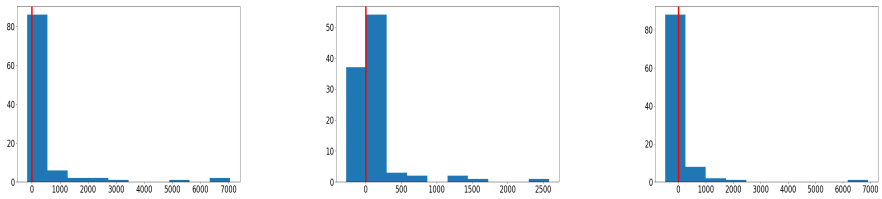
4.5.1 Validation

Comparison with arbitrary designs

In order to check how robust this design is across varying data structures, the following experiment was performed: We selected sample sizes of 4, 5 and 6

mins	0:00	0:15	0:30	0:45	1:00	1:15	1:30	1:45	2:00
Conventional	x				x				x
Proposed	x			x		x		x	x
Full	x	x	x	x	x	x	x	x	x

Table 4.1: Conventional times for glucose measurement in OGTTs and our proposed times. The "Full" times are used for validation purposes in section 4.5.1



(a) (b) (c)

Figure 4.1: Histograms of differences between the quality of our proposed design and of an arbitrary design on random data (arbitrary units). The vertical line indicates a difference of zero. The arbitrary design has one point less (a), the same number of points (b) and one more point (c) than our proposed design with 5 measuring points, seen in Table 4.1. Note that, with the considered sample sizes, including bigger designs all of these histograms have right tails and none of them have a left tail. This means that our proposed design is never significantly worse than the arbitrarily chosen alternative, and is sometimes much better.

data points (including measurement upon arrival). 100 designs were generated uniformly at random for each size. For each design a random "patient " was simulated, drawing ϑ from $\pi_{\mathcal{I}}$. For each simulated patient a sample was simulated for the random design and also for the proposed design. Inference was performed on each sample and the utility as described in section 4.5 was estimated.

Figure 4.1 shows the histograms of the differences in utility between the arbitrary design and our suggested design for each sample size (utility of suggested design minus utlity of arbitrary design). We see a general trend: For most values of the parameters, the design does not make a very big difference in the quality of inference, thus the differences cluster around zero. All of the histograms have a right tail, and none of them have a left tail: For some values the design is more important; in these situations our design significantly outperforms the arbitrary design, even when the arbitrary design has a larger sample size. We can therefore conclude that our design does appear to be a generally good choice.

Comparison with the conventional design

While it is a very good sign that our design outperforms random designs, it is also important to compare our results with the conventional OGTT testing design which is actually used in practice. In the conventional design measurements are taken at $t = 0, 1, 2$ hours. This design has two fewer measurements than our proposed design so it is reasonable to expect that our design will be better for that reason alone, but it also means that the design is more costly. Quantifying the improvement over the classical design is therefore necessary to understand if and when this extra cost pays off.

To compare our new design to the conventional one, we have a sample of 17 real (healthy) patients, obtained by AM, for whom OGTT measurements were taken every 15 minutes, resulting in information that is significantly more complete than what is usually available from OGTT tests. The conventional and proposed designs, as well as the full design were shown in table 4.1.

In order to compare the two designs, the utility function must be estimated, but since these are real patients, the true value of the parameters is unknown. It is therefore not possible to estimate the expected utility with the precision which was used before, but a surrogate utility can be written which behaves similarly using the inference from the full data. The true utility function can be written as

$$U(d) = -\int\int_0^3 (G_\theta(t) - G_{\hat{\theta}}(t))^2 dt \; \pi_{\mathcal{I}}(\hat{\theta}|y, d) d\hat{\theta}.$$

Since in this case the true parameters θ are unavailable we use their posterior distribution as calculated using the data from the full design. Our new surrogate utility is now

$$\hat{U}(d) = -\int\int\int_0^3 (G_\theta(t) - G_{\hat{\theta}}(t))^2 dt \; \pi_{\mathcal{I}}(\hat{\theta}|y, d) d\hat{\theta} \; \pi_{\mathcal{I}}(\theta|y_f) d\theta$$

where $\pi_{\mathcal{I}}(\theta|y_f)$ is the posterior distribution of the parameters θ using the full data y_f, that is, with measurements every 15min. This surrogate utility can be estimated using the available samples.

This was done for the available set of 17 patients and the estimates for the surrogate utilities were compared using the conventional design and using our proposal. It is not surprising that the new design is better than the conventional design since, to start with, it has more measurements, but we want to know how much better. In order to adequately represent the relative difference, a histogram of the *quotients* of these utilities can be seen in figure 4.2. There are two patients for whom the utility of the conventional design outperforms the new one. For no patient did the new design result in an estimated utility of less than 82% of the utility of the conventional design. For all other patients the new design outperforms the conventional design, usually by a factor of 2 or greater, and sometimes by a much wider margin.

As this example shows, the effect of choosing a better design can be dramatic. For the data tested, our proposed design has proven to be a significant

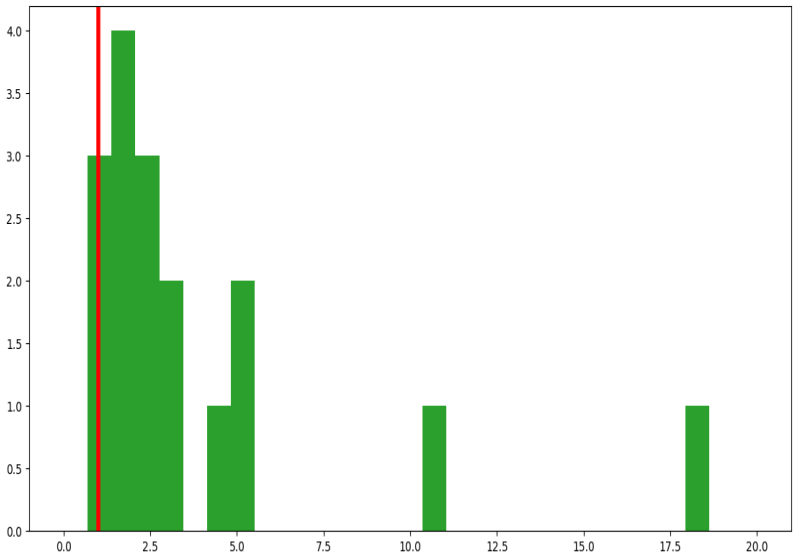


Figure 4.2: Histogram of the quotients of the surrogate utility functions for 17 real patients using the conventional and proposed designs (conventional divided by proposed: All values are negative, so large quotients mean the conventional design yields larger errors). The vertical line indicates a quotient of 1.

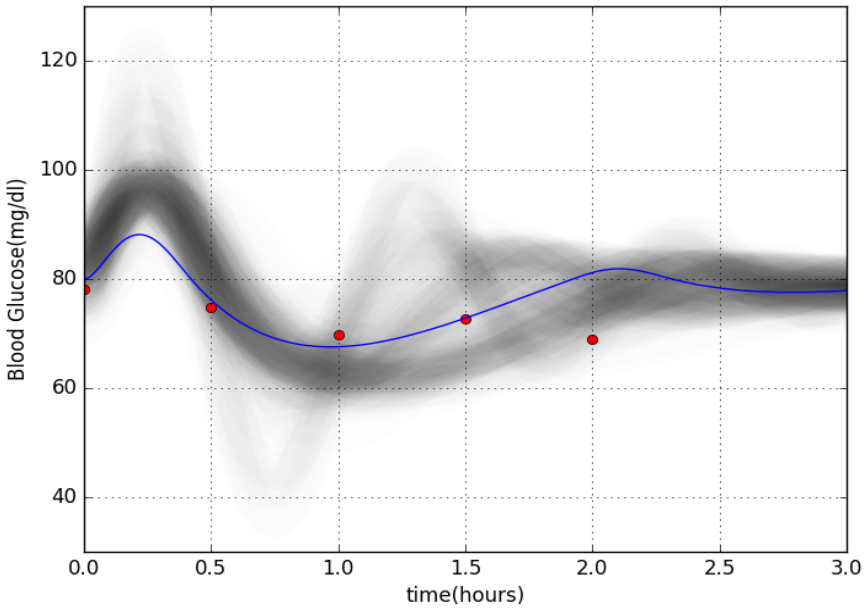


Figure 4.3: Simulated data for an extremely unusual situation where a patient's insulin response is 80 times stronger than the glucagon response. The true curve is in green and the simulated data in red. Although the data is extremely unusual, our design points yield the necessary information to obtain reasonable information about this strange behavior.

improvement; raising the number of measurements from 3 to 5 achieved more than twice the utility for most patients.

Simulation test to verify robustness

Another important test is to verify the behavior of this design when evaluating a particularly unusual set of data. Since the design was trained using typical scenarios, we should verify that highly unusual shapes can still be discovered. Data was simulated using a very strange set of parameters: $\theta_0 = 80$, $\theta_1 = 1$, $\theta_2 = 1.5$, $G_0 = 80$. This represents a patient whose insulin response is extremely violent, but whose glucagon production is not. We have never seen a patient like this one, with such a dramatic difference between the production of the two hormones. Data was simulated using the proposed design and a posterior sample can be seen in figure 4.3 with the true curve seen in green. As we can see, while the inference is imperfect, the design still works reasonably well in performing inference even on this extremely strange patient.

4.6 Discussion

In this chapter the model proposed in chapter 2 was used to suggest a better analysis and improved sampling protocols for Oral Glucose Tolerance Tests. The main objective was to use the model to redesign the OGTT test in a way that improves the quality of the information gathered. The chosen technique to achieve these purposes was Bayesian experimental design, in order to find an alternative set of times at which to perform the glucose measurements on the patient.

Although some techniques for Bayesian experimental design already exist, the specific properties of this problem lead us towards developing a different tool for comparison of experimental designs, which is computationally intensive, but which provides a fine control of uncertainty in the design process itself.

We used this new tool to select a design for the OGTT and the resulting choice was compared favorably both to the classical design (with real data) and to hypothetical arbitrary designs. The result is very promising and may lead to improved diagnosis techniques for patients who are at risk of type 2 diabetes.

From a mathematical perspective, there remains an issue regarding the algorithm for comparing designs in those cases when a decision should be forced (by increasing T_1). In section 4.3.2 we briefly discuss the notion of sequential comparisons in cases where the initial test proves inconclusive. While this was not necessary in our case, in most other cases the value of T_1 is not easily interpretable. If our method is to be generally applicable, an efficient algorithm for sequential testing should be developed. Although we performed some numerical experiments in sequential design, we have not come to any clear conclusions regarding how to do it efficiently.

The most innovative and potentially controversial issue in this chapter (and perhaps the thesis in general) is of course the explicit use of two separate priors over the parameter space, one for design and one for inference purposes. In section 4.4 we have discussed some pragmatic reasons why this may be a desirable - and in some cases necessary- option, but it may be possible to treat the subject more formally. Prior selection is -after all- a decision, and it may be possible to frame this instance of prior selection in the context of decision theory. The use of decision theoretic constructions to select priors has been studied in the context of reference analysis (see for example Bernardo and Smith, 1994) and a similar approach may potentially shed light on this context as well.

Chapter 5

Capillary and venous blood

5.1 Capillary Glucose

In this chapter we investigate a change to the OGTT protocol which alters the technology and infrastructure involved. In particular, we are interested in a change which lowers costs and discomfort for patients. While in the previous chapter we were interested in finding out whether our mathematical model could be used to fine tune OGTT tests for greater precision, in this chapter we will be interested in whether our model permits us to use equipment previously thought to be insufficiently precise.

While the ODE model we have been working with works well for standard OGTT tests, these tests are invasive and bothersome for the patient. They are also slow, expensive, and take up valuable time for the laboratory staff. There is, however, an alternative method for measuring blood glucose which is much more practical. It consists of the use of an apparatus known as a glucometer. A glucometer is an easily available device which measures blood glucose from capillaries rather than veins. Glucometers were designed for diabetic patients to measure their blood glucose at home. They are easy to use, yield instant results, and only require a single drop of blood from the tip of the patient's finger. The reason that glucometers are not commonly used for OGTT tests is that typically they are thought not to be accurate enough (Ginsberg, 2009).

While glucometer accuracy is indeed significantly less, and importantly, often shows a systematic bias, the use of our mathematical model was able to drastically increase the information gained from an OGTT. It may be reasonable to think that the gain in information from using our model might compensate for the less accurate measurements from the glucometer. In this chapter we explore whether or not this is the case.

This chapter is organized as follows. Section 5.2 shows the results of attempting to use the dynamic model directly with glucometer data. The results are less than satisfactory, so section 5.3 describes the efforts that go into adjusting the error model for glucometer data. The results of this investigation are

very promising, so in section 5.4 we continue to develop this idea by revising the model we were using for venous errors to begin with. This new revision results in a model which we believe allows for the serious consideration of glucometer measurements to perform OGTTs. Section 5.5 discusses the remaining issues present in the model and its limitations. Finally, section 5.6 concludes the chapter.

5.2 Testing Glucometer Data with the Dynamic Model

In order to test the value of glucometer data, our medical collaborator collected glucometer data from patients who were visiting the hospital for a normal OGTT test. Along with the usual OGTT measurements, these patients also had their blood tested with a glucometer at the same time as the blood sample used for the normal tests. The glucometer measurements are - essentially - a duplicate OGTT for the same patient, measured with a glucometer and capillary blood. In order to ensure the patient's safety, our medical collaborator also habitually tests venous blood with a glucometer before giving patients the glucose concentrate (this only takes a few seconds and ensures that if a patient's glucose is too high then he/she can be treated immediately). This is convenient because it provides a measurement which differentiates the effect of the blood (venous vs capillary) from the effect of the glucometer.

Attempting to directly use the initial model as-is with capillary glucose, however, is a failure. In figure 5.1 we see inference curves for this procedure for a real patient. The red curve represents the posterior using venous blood the usual way and the blue curve represents the posterior using glucometer measurements. As we can see, these results not only do not match, but they exhibit wildly different behavior and would likely lead to different diagnoses and mistreatment. It is clear that if capillary glucose is to be used, then some change is required for the analysis of the data.

5.3 Glucometer Error Model

Our first improvement from simply using the preexisting model is to adapt our error model for measurements. Since the device that is measuring is one that is known to be imprecise, the model should be adjusted to reflect this error. Literature about glucometer error indicates that the conditions of measurement affect the reliability of glucometer measurements. OGTT conditions are quite singular in that they are performed by a professional in a carefully controlled environment, but with a patient whose body is undergoing an unusual kind of stress.

From our collaborator's dataset, the proportional differences between glucometer and venous blood glucose measurements are charted in a histogram, seen in figure 5.2 as well.

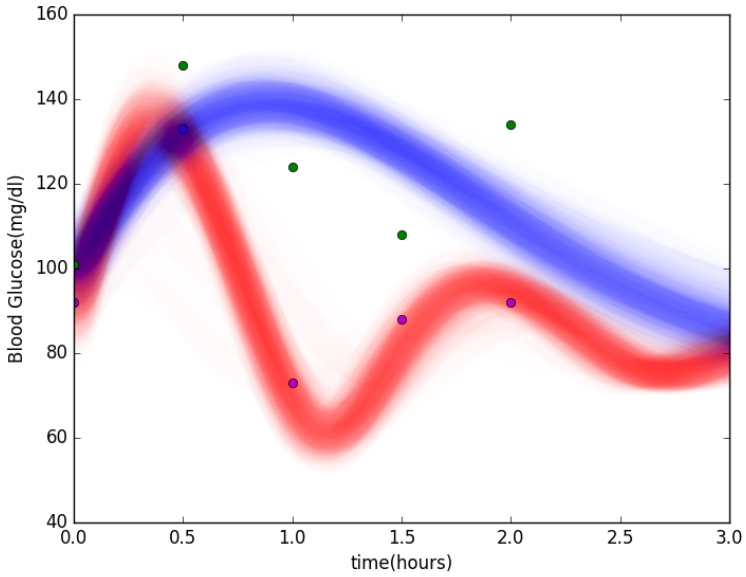


Figure 5.1: Inference curves for the same patient, using the unaltered dynamic model. Using venous data we obtain the magenta dots and red curves and using capillary data we obtain the green dots and blue curves. As we see, these curves do not match, and lead to very different inference about this patient. It is because of cases like this one that the model must be revised for capillary blood.

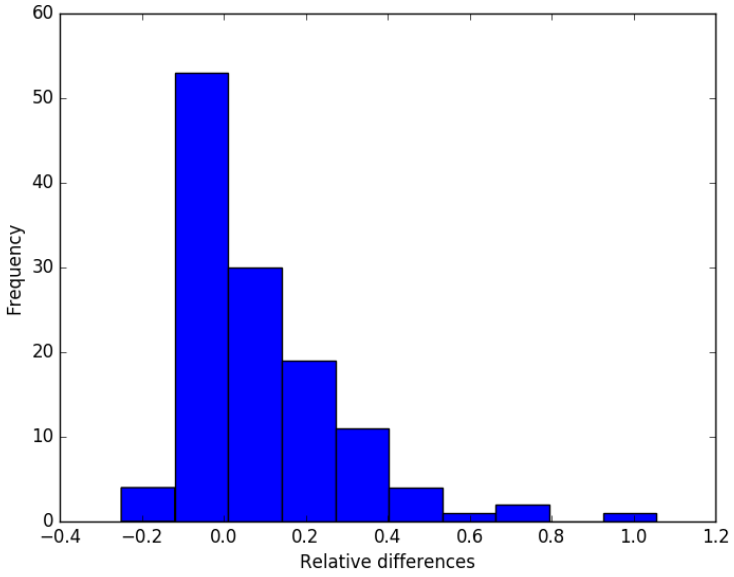


Figure 5.2: Histogram of the relative differences between venous blood glucose measurements and glucometer capillary blood measurements.

One important detail to note is that along with the wide spread of values, there is also a clearly visible bias. This bias makes medical sense as well, since blood in the capillaries delivers glucose to muscles before entering the veins. The errors of the venous test are assumed to be negligible in comparison to the errors of the glucometer test (this assumption is later confirmed when the errors of the venous test are studied carefully in section 5.4 of this chapter), so modelling the difference between the glucometer and the venous data is tantamount to modelling glucometer error.

Since literature on glucometer errors usually works with proportional errors, this version is chosen. A model was selected which gives the error as the sum of a gamma distributed bias due to the loss of glucose in the capillaries, and a normally distributed measurement error. Since the bias is driven by biological processes in the patient's body, it is assumed to be equal for all measurements across a single OGTT, but different from patient to patient.

Point estimates are obtained for the bias and error parameters and the model is updated. Our data now looks like

$$y_i = G(t_i) + b + \epsilon_i;\ b \sim Gamma()\ \epsilon_i \sim \mathcal{N}(0,)$$

The inference method was adapted to this error model, and the results can be seen in figure 5.3 for the same patient as figure 5.1. The improvement is very noticeable. Quite similar curves are obtained despite the data being extremely

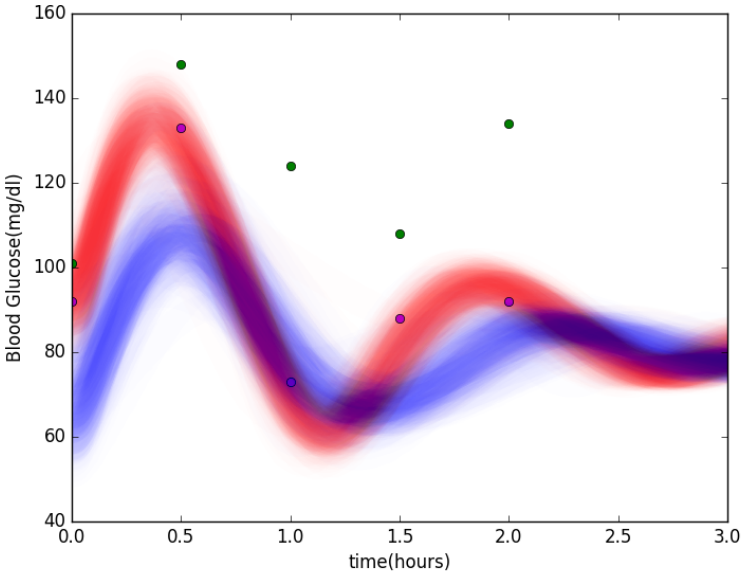


Figure 5.3: The same patient from figure 5.1, with the new error model. Although the data are very different, it is possible to recover similar information.

different. There is an almost magical quality to the difference between the glucometer measurements and the data obtained from them.

5.4 Venous Test Error Model

Although the patient whose data is in figure 5.3 is by no means unique, not all patient's data work quite as well. In figure 5.4 we see one patient where the adjustment still is not sufficient to explain the difference between the two methods for measuring the OGTT. Both curves still indicate approximately the same general degree of health, but they nonetheless represent somewhat different situations. If glucometer OGTTs are to be used for any any serious purpose, even more must be done.

One idea is to revisit the error model for venous tests. While gaussian noise with a five mg/dl standard deviation was the first proposed idea, it was accepted mostly on the basis of providing good results and not subjected to close scrutiny. Additional work might improve our model for venous test error, and hopefully shed more light on the apparent flaw in our analysis of glucometer results.

For this purpose our medical collaborator produced a series of 10 duplicate blood glucose measurements. For these measurements, a venous blood sample was taken, split in half, and its glucose content was measured twice (These

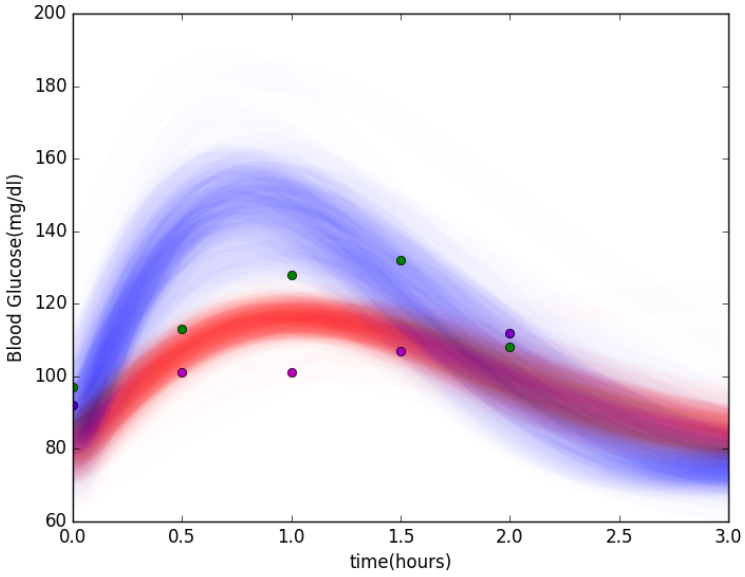


Figure 5.4: A case of a patient where the adjustment for capillary data was not sufficient to achieve similar results from venous and capillary blood.

are not full OGTT results, but simply 10 blood samples). There is no medical reason to run the same sample twice, so these data were collected for this purpose only. The samples showed that the variance in measurements is actually quite low, but they did include one outlier measurement which was more than two standard deviations from zero. This outlier was found regardless of whether the differences were treated as absolute differences or as proportional differences.

A second series of 5 duplicates was produced and it contained a second datum more than two standard deviations from the mean. While the amount of data available is still somewhat small, there is reason to believe that the variance of the error is even smaller than previously believed (our estimates show a standard deviation of around $0.027 * G$), but the errors have heavy tails.

For this reason, an alteration to the original venous blood model was proposed, replacing the Gaussian errors with a scaled t distribution with 4 degrees of freedom. Hence, our new error distribution for venous measurements has density function:

$$p(x|sigma) = \frac{\Gamma(\frac{5}{2})}{\Gamma\frac{5}{2}\sqrt{4\pi\sigma^2}} \left(1 + \frac{x^2}{4\sigma^2}\right)^{\frac{5}{2}}$$

with $\sigma = 0.021$.

Using this error distribution for venous blood, the original capillary and venous OGTTs were recalculated. Figure 5.5 shows the same patient from

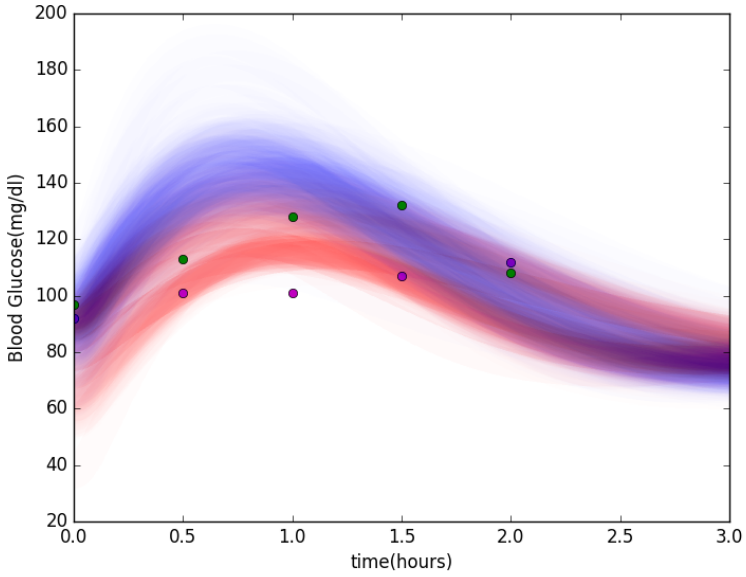


Figure 5.5: The same patient from figure 5.4 with the new adjustments for venous blood. The new model causes the venous posterior to become bimodal, and one of the modes matches the capillary inference quite closely.

figure 5.4. In this case we see that the posterior distribution for venous blood is bimodal, and that one of these modes matches well with the posterior obtained using capillary blood. This particular patient is by no means the only example of this phenomenon. In fact, the agreement between the two improves with this adjustment almost across the board. This is a clear indication that a heavy tailed error distribution improves the model used for venous blood. It is also an indication that the model used for capillary blood is better than it seemed previously.

5.5 Shortcomings

After the aforementioned adjustments, our model for capillary glucose is drastically improved, but it still does not quite manage the same power as venous OGTT tests. There are three main issues, the first of which is illustrated in figure 5.6 where we see a venous and capillary OGTT for the same patient, but although the capillary test produces a posterior which overlaps the posterior from the venous test, the posterior from the capillary OGTT has so much variance that it is very difficult to draw any conclusions about what it indicates about the patient. This is not too big of a problem. If a patient's results from

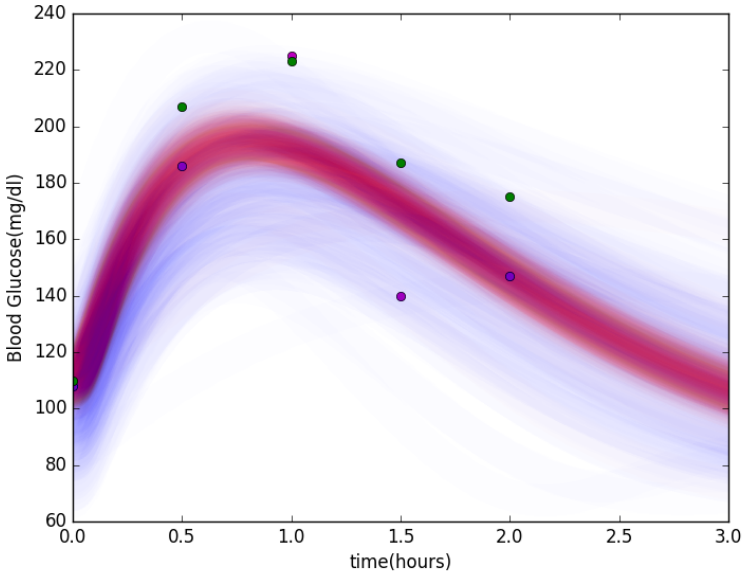


Figure 5.6: In this data set, the capillary inference (blue curves) has an enormous variance. In such a situation, although the information does match the venous test, it is so vague that it is impossible to draw any definite conclusions from it. This sort of situation is not such a big issue because if results are too vague then a second venous OGTT can be performed to increase certainty.

a glucometer test have too much variance then a second venous test can be performed to gain certainty.

The second issue is illustrated by figure 5.7. In this case, the glucometer tests produce a single wild outlier which completely skews the data. Changing the error model for capillary tests to a heavy tailed model does fix these issues, but the results from other tests suffer greatly. This is not such a big problem either since wild outliers of this kind can be easily spotted, and handled accordingly.

The third issue is illustrated by figure 5.8. In this case neither is there enormous variance in the glucometer OGTT nor is there a wild oulier. The issue is that the posteriors simply do not match. In figure 5.8 this is visible even directly from the data, since the data from the capillary OGTT and the venous OGTT simply exhibit different behavior. No adjustment to the error model will solve this issue. From 65 patients for whom this test was performed, only 3 of them produced this sort of unacceptable error.

Of note, upon first examination, 4 unacceptable cases were found, but the raw data for these cases was reviewed. After reviewing the raw data, one case was found to be an error when reading the output of the venous test. The way this patient was handled initially was incorrect, and without this careful study

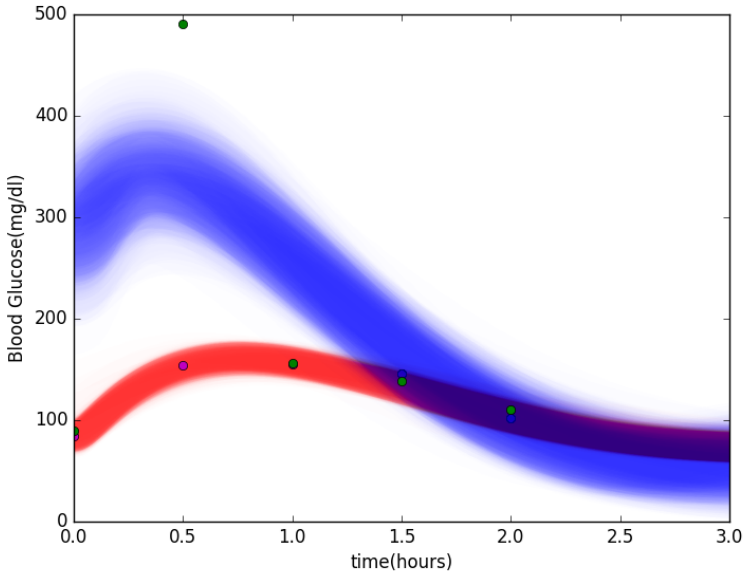


Figure 5.7: For this patient, we see that the capillary curves have a single wild outlier 30 minutes from the start of the test. This causes the curve to vary wildly from the venous data. This single outlier is fairly easy to spot, however, and can be handled. There are several known factors that can cause single extreme glucometer measurements, but our collaborator believes that this particular case is simply one of incorrectly transcribed data when performing the OGTT. Changing the glucometer error distribution to a heavy tailed one does solve this, but it worsens inference in most other cases.

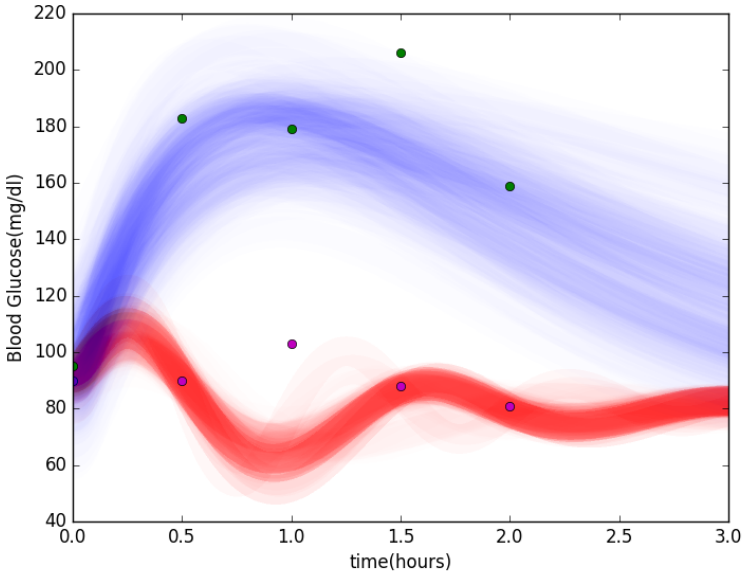


Figure 5.8: A case where no matter what model is used the venous and capillary curves do not match. This behavior is predictable simply by looking directly at the data, since they exhibit different behavior.

of the capillary data the error would not have been detected.

5.6 Conclusions

While the intent of this study was merely to investigate the possibility of using the ODE model to study capillary OGTTs, the investigation also led us to propose changes to the model for venous OGTTs as well. That improving the venous model made the results match the capillary model more closely makes a strong case for the value of capillary tests.

The new capillary model has produced unacceptable results in nearly 5% of the tested patients. While this does seem like an alarmingly high rate of error, it is worth noting that the original venous technique produced at least one patient with equally unacceptable results, and that it was precisely the capillary model which allowed the detection of this problem.

The practical gains to using capillary blood and glucometers are significant: Not only are venous tests expensive and bothersome, but for many people they may simply be unavailable because of lack of access to a medical facility with the proper equipment. Glucometers, on the other hand, are frequently sold in pharmacies, intended for home use, and are relatively inexpensive. With the analytical tools to infer from capillary tests, it is now possible for a health practitioner to perform an OGTT test using only a glucometer; the data could be analyzed using our model, and relevant software.

While data on error rates in diagnosis from classical OGTT analysis (without the ODE model) are not readily available, there is reason to suspect that errors may well occur even more often than with tests using the capillary model (Davidson et al., 2000). It is therefore worth considering capillary OGTT as a reasonable alternative for situations when venous OGTT tests are impractical.

Chapter 6

Conclusions

The primary purpose of this thesis is to thoroughly analyze various changes and improvements to the OGTT test for type 2 diabetes. The resulting contributions in this thesis are of two main kinds. The first class of contributions are those which are specifically related to the medical issue of OGTT tests. The second class of contributions are applicable to more general statistical theory, and these arise throughout the course of the study. While this second class of contributions were not the primary focus of the thesis, they add substantially to the significance of the research.

There are four main contributions of the first class. The first of these is the dynamic model itself, as seen in chapter 2. In general, diagnosis of type 2 diabetes is done based on full clinical histories rather than on the basis of any single test. While tests such as the OGTT are important indicators, they typically do not constitute the basis for diagnosis. This may in large part be due to the fact that the full information from OGTT tests is not put to good use by common methods of analysis. While the idea of using a dynamic model is not new, we are unaware of any serious attempts to fit other models to the data from real patients and to use the models as a means for analysis. We believe that the results from fitting the model to data make it self-evident that there is far more information in a typical OGTT test than what the usual analysis techniques can take advantage of. Whether or not this specific model is eventually used in practice, we are certain that some sort of similar idea should replace the old guidelines.

The second contribution to type 2 diabetes is the investigation of the data from the new breath tests, as seen in chapter 3. The contribution in this case did not result in any improvement in methods for diagnosis of type 2 diabetes, but the investigation is nonetheless worthwhile since the technique did appear promising at first.

The third contribution to type 2 diabetes is the design of proper times for testing, as studied in chapter 4. The times chosen depend significantly on the model, so this experimental protocol is only really recommended if the use of the model is adopted. Nonetheless, if the model is adopted, the implementation

of the change is almost trivial. We investigated the gain which results from using our design in contrast with ones that are in common use, and found that our design sometimes makes great improvements on the quality of the resulting information. As a result, if our dynamic model is adopted for regular use, the experimental design is an adaptation that has some significant advantages and which comes at almost no cost.

The fourth contribution to OGTT tests is our analysis of glucometer data, as studied in chapter 5. From a practical standpoint, this may be the most significant contribution to OGTT tests in the sense that it allows tests to be performed in situations where it may previously have been thought to be impossible. These adaptations do require some form of dynamic model, but the method for choosing them is independent of the model, so they can be used regardless of the specific analytical technique being used. This work does not in any way contest the previously common notion that glucometer measurements are insufficiently precise, and in fact we gather some further evidence to support this belief. On the other hand, however, we find that this belief is true for *individual* glucometer measurements, while through our model, the *joint use* of several of them – as in the context of an OGTT – can make great strides to improve the information, and in some cases to resolve both the imprecision and the bias inherent in this kind o data.

There are two contributions to statistical theory in general. One of them is found in the investigation of the breath test (chapter 3) and the other is in the selection of measurement times (chapter 4). They are fundamentally very different in nature. The breath test was a contribution to the methodology of variable selection. This is a broad and well explored topic, and our contribution comes not from the unavailability of applicable methods, but rather somewhat accidentally, from ideas which came up when deciding which method to use. The methodology developed is in line with the philosophy of Bayesian statistics, but it extends beyond the boundaries of Bayesian statistics and is useful in many situations. On the other hand, the other contribution - to the theory of Bayesian experimental design, is one that arises out of necessity. The available literature is sparse, and most of the available techniques rely on certain assumptions about the problem that we did not believe were met in our situation. The problem as described – and the related methods – are problems specifically related to Bayesian analysis.

Variable selection is a common problem, and one that has been studied extensively. That said, there are a multitude of subtle differences between the various scenarios where variable selection is desirable. We do not, in general, believe in a one size fits all approach to the problem, and the FATSO operator developed in chapter 3 is a good solution for situations when we wish to make use of a certain kind of external information. In general, the $p > n$ regression problem is one in which the necessary information is not found entirely in the data. All regularization techniques are ways in which - in one sense or another - subjective information is used to select one solution. This subjective information is typically not explicit, and usually interpretable only in the sense of very abstract guidelines, but these regularization methods are still tuned by the user

in some sense. FATSO can be thought of as a way to make the information more explicit and easier to interpret. When appropriate information is available, it would be ideal to tune regularization parameters in a way that uses it. In these situations, FATSO is recommendable.

Besides linear regression, where some common expected utilities are treatable, Bayesian experimental design seems to be an under-explored topic in general. The problems are typically computationally very complex, involving great amounts of calculation. They are also conceptually difficult, with some very subtle issues relating to the selection of utility functions, and several sources of error. Most of the prior work had been either applicable only under narrow circumstances, or had insufficient theoretical justification. The main objective of our work on experimental design was primarily to obtain a mathematically sound way to design Bayesian experiments without making any unreasonable demands on the behavior of the problem. Our resulting algorithm is slow, but it has received proper mathematical treatment and works under very general conditions. We also propose an unusual handling of utility functions and priors. While the idea is unconventional, we believe it makes good pragmatic sense. Fundamentally, inference and design are two different problems, and different problems sometimes require different models, which express different parts of the issue. Since priors are - at heart - models for our uncertainty, it should make a lot of sense to adjust them to properly model the problems at hand.

Overall, our primary objective was to make suggestions that improve the power of techniques for diagnosing diabetes, particularly in the context of OGTT tests. In this sense, our purpose was to provide a framework on which to build a clinical protocol for OGTT tests which may greatly improve the existing methods. Certainly, our approach has been centered around our dynamic model. While the latter is the backbone of the research, however, we are more interested in the idea of using a dynamic model in general than we are in insisting very strongly on this particular one. The model does provide significant improvements to the understanding of OGTT data, but some other similar models may also do so. The main point is that the use of a dynamic model in general is a good idea, capable of creating large improvements in OGTT analysis.

To the extent that was possible, some of our contributions in the analysis of the model have been tailored to be useful independent of exactly what specific model is being used. While the times selected in the design problem are model dependent, the method used for selecting those times is very general. We expect the algorithm, as well as the utility function, to be plausible options for finding good times regardless of what the model is. The adaptations for capillary glucose go beyond this, and while they were tested using the dynamic model, the adaptations themselves were generated without using the model at all, and we believe that the results will probably work well with any reasonable dynamic model.

Overall, this thesis propose a new framework for OGTT analysis. In the process, we have developed some new statistical techniques, and examined an alternative to OGTT tests. The proposed framework has been thoroughly examined and we have proposed several adaptations for practical use. While the

implementation of our proposal is a question for medicine rather than statistics, we believe that this thesis provides the backbone for a testing protocol worthy of consideration for widespread use.

Appendix A

Easy plotting of posterior distributions for functions using transparencies with a KDE justification

The graphs of OGTT inference in this thesis were produced using a simple and easy method which has an interesting theoretical justification. It can be used for Bayesian regression problems for which inference was performed by a posterior sampling algorithm, where the regressor is a function of a single variable (such as time) and where the object of interest is the posterior distribution of the regressor itself rather than any of the associated parameters.

Write the model as $Y = f_\theta(X) + \epsilon$. Once posterior inference has been performed using a Monte Carlo method, we have a posterior sample $\{\theta_i\}_{i=1}^T$. Essentially all that is required is to overlay graphs of $f_{\theta i}(x)$ for each element i of the posterior sample of θ, adjusted with some degree of transparency α. The resulting graph, in tones of grey, displays an estimate of the posterior distribution of the regressor in the following sense. For any vertical slice $x = c$, in tones of grey, the opacity of each pixel is proportional to a kernel density estimate (Silverman, 1986) of $\pi(f(c)|Y)$. In fact, a kernel density estimate using a uniform kernel with the bandwidth h equal to the width of the graphed line.

This is easy to see. If we write out the opacity of each pixel as

$$O(y) \propto \sum_i \mathbb{1}(|f_{\theta i}(c) - y| < h/2)$$

and note that the kernel density estimator is

$$\hat{l}(x) = \frac{1}{nh} \sum_i K\left(\frac{x - X_i}{h}\right)$$

and if we write $K(x) = \mathbb{1}(|x| < \frac{1}{2})$ then we note that the opacity is proportional to the appropriate kernel density estimate.

Of course, not all kernel density estimates are equally good. In particular, the selection of h greatly affects the quality of the estimate, and the selection of α affects the visibility and legibility of the resulting outcome. In some situations, it may be reasonable to tune these values by hand, but in our case we made a very large number of these graphs. For this reason, an automated way of selecting these values in some reasonable sense is desirable.

We set $\alpha = \frac{1}{n}$ where n is the size of the posterior sample. This assures us that no pixel will ever be at over 100% opacity. Since opacity is truncated at this point, any pixel at higher opacity would not be displaed correctly in the graph. It is frequently the case that no pixel actually reaches such high opacity, so a less stringent bound is theoretically plausible, but visual examination of the graphs indicates that -at least in our case- no graphs are ever too light to read, so this simple rule of thumb is effective.

The problem is harder in the case of h. In particular, the optimum bandwidth for a kernel density estimate is known to depend on the density, and within one graph, there is a different density of $f(c)$ for each value of c. Since varying the thickness of the lines throughout the graph defeats the purpose of having this very simple method, we choose to select the line width uniformly. This may produce some cases where some kernels are over or under-smoothed, but the objective is a visual representation, so subtle differences are acceptable. In our case, the disparity in distribution is not big enough to matter. We choose to use the distribution at $G(t)$ for $t = 3$ hours to select h and then apply this h throughout.

To select h, we use Silverman's rule of thumb (Silverman, 1986). This is the optimum h for estimating a gaussian distribution with a uniform kernel. Silerman's rule selects

$$h = \left(\frac{4\hat{\sigma}^5}{3n}\right)^{\frac{1}{5}} \approx 1.06\hat{\sigma}n^{-1/5}$$

where $\hat{\sigma}$ is the sample standard deviation. It is well known that this rule sometimes over-smooths, particularly for multimodal or heavy tailed distributions, but once again, the intent is only to find a reasonable automated selection which produces easily interpretable graphs, and in practice we have found that Silverman's rule is sufficient to achieve this.

Figure A.1 shows an example of the best possible case scenario for this technique. The sample is a linear model where the slope is known and the posterior for the intercept is a gaussian. In this case the variance of the posterior of $f(c)$ does not depend on c and Silverman's rule calculates the exact optimum bandwidth. This particular situation makes the technique almost unnecessary. For more organic examples of good cases to use this technique, see the rest of this thesis.

Figure A.2 shows an example of this technique failing badly. In this case it is also a simple linear model, but the intercept is fixed and the slope varies.

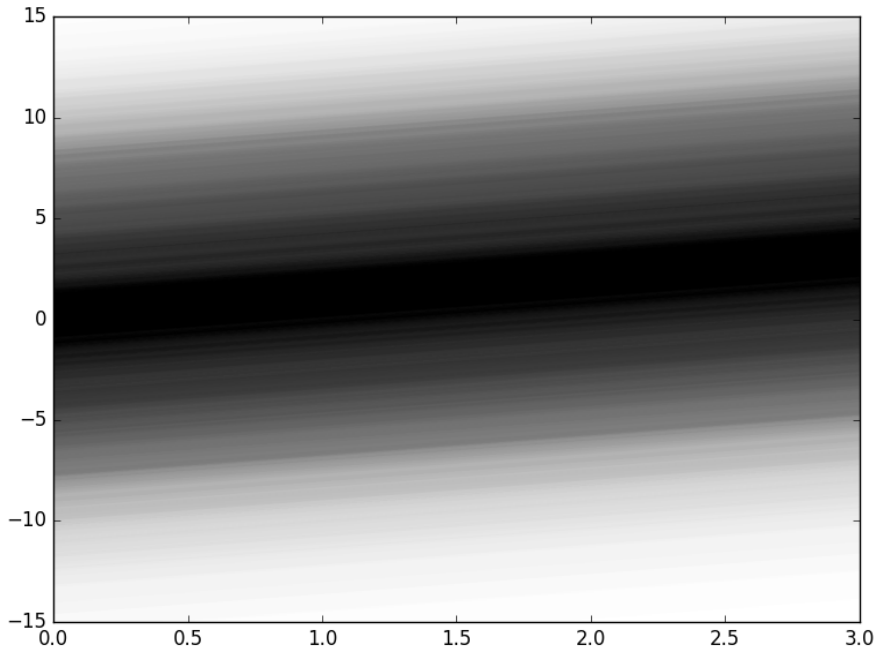


Figure A.1: The best case scenario for this technique has a uniform variance accross the window and has an exactly gaussian posterior. The illustrated situation is so simple that this technique is hardly necessary, but the resulting graph is very clear.

In this case the variance of the posterior of $f(c)$ at the $c = 0$ is exactly 0 and grows linearly with c. Silverman's rule was calculated midway through the window. Not only is the distribution drastically oversmoothed at $c = 0$ and undersmoothed at $c = 3$, but the great overlapping of curves near the origin produces graphical artifacts and interferes with the observed transparency of the lines in the graph.

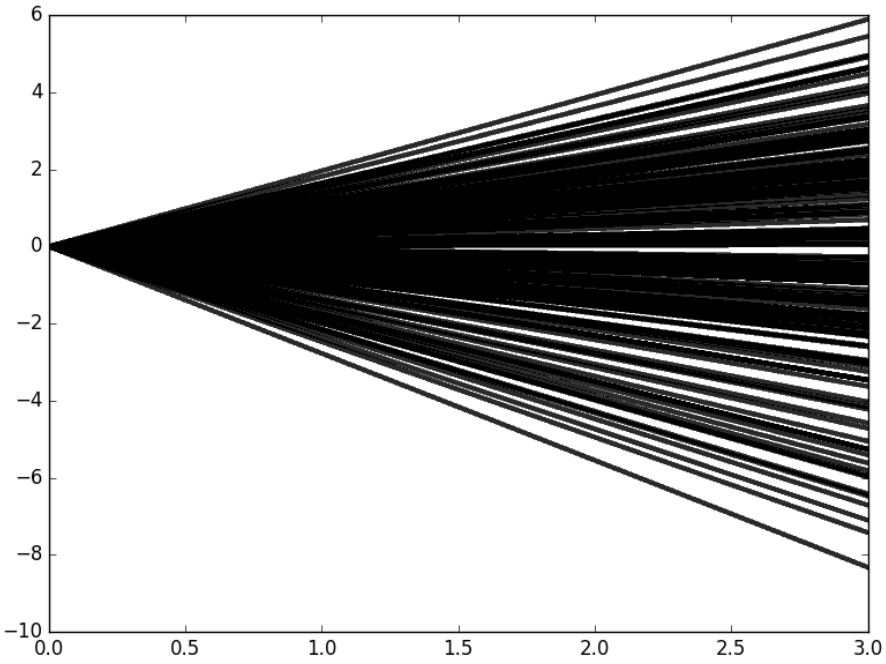


Figure A.2: A situation where the technique fails. The variance of $f(c)$ changes greatly with c. Silverman's rule was calculated towards the center of the region. Not only is the origin oversmoothed and the edges undersmoothed, but the overlapping of lines at the origin causes graphical artifacts whcih interfere witht he transparency of the lines.

Bibliography

Alexanderian, A., N. Petra, G. Stadler, and O. Ghattas (2016). A fast and scalable method for a-optimal design of experiments for infinite-dimensional bayesian nonlinear inverse problems. *SIAM Journal on Scientific Computing 38*, A243–A272.

American Medical Association, Boyd E., M. M. (2006). *American Medical Association Guide to Living with Diabetes: Preventing and Treating Type 2 Diabetes - Essential Information You and Your Family Need to Know* (1 ed.). Hoboken: John Wiley and Sons.

Anand, F. S., J. H. Lee, and M. J. Realff (2010). Optimal decision-oriented bayesian design of experiments. *Journal of Process Control 20*, 1084–1091.

Anderwald, C., A. Gastaldelli, A. Tura, M. Krebs, M. Promintzer-Schifferl, A. Kautzky-Willer, M. Stadler, R. A. DeFronzo, G. Pacini, and M. G. Bischof (2011). Mechanism and effects of glucose absorption during an oral glucose tolerance test among females and males. *J. Clin. Endocrinol. Metab. 96*(2), 515–524.

Berger, J. O. (1993). *Statistical decision theory and bayesian analysis* (2nd ed ed.). Springer series in statistics. New York: Springer-Verlag.

Bernardo, J. M. and A. F. M. Smith (1994). *Bayesian theory* (1st ed.). Wiley series in probability and mathematical statistics. Chichester, Eng. New York: Wiley.

Berry, S. M. and J. B. Kadane (1997). Optimal bayesian randomization. *Journal of the Royal Statistical Society Series B (Methodological) 59*, 813–819.

Bhattacharya, A., D. Pati, N. S. Pillai, and D. B. Dunson (2014). Dirichlet-laplace priors for optimal shrinkage. *Journal of the American Statistical Association 110*, 1479–1490.

Box, G. E. P. and G. C. Tiao (1992). *Bayesian inference in statistical analysis*. New York: Wiley.

Carvalho, C. M., N. G. Polson, and J. G. Scott (2010). The horseshoe estimator for sparse signals. *Biometrika 97*, 465–480.

Chaloner, K. and K. Larntz (1989). Optimal bayesian design applied to logistic regression experiments. *Journal of Statistical Planning and Inference 21*, 191–208.

Christen, J. A. and C. E. Buck (1998). Sample selection in radiocarbon dating. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 47*, 543–557.

Christen, J. A., M. Capistrán, A. Monroy, S. Alavez, S. Q. Vargas, H. A. Flores-Arguedas, and N. Kuschinski (2016). A Diabetes minimal model for Oral Glucose Tolerance Tests.

Christen, J. A. and C. Fox (2010). A general purpose sampling algorithm for continuous distributions (the t -walk). *Bayesian Analysis 5*(2), 263–281.

Davidson, M. B., D. L. Schriger, A. L. Peters, and B. Lorber (2000). Revisiting the oral glucose tolerance test criterion for the diagnosis of diabetes. *Journal of General Internal Medicine 15*, 551–555.

DeGroot, M. and M. Schervish (2011). *Probability and statistics* (4ed. ed.). Boston: AW.

Duckworth, W. C., R. G. Bennett, and F. G. Hamel (1998). Insulin degradation: Progress and potential. *Endocrine Reviews 19*(5), 608–624. PMID: 9793760.

Duflo, M. (1997). *Random Iterative Models* (1 ed.). Stochastic Modelling and Applied Probability 34. Berlin, Heidelberg: Springer-Verlag Berlin Heidelberg.

Eaton, M. (2007). *Multivariate statistics : a vector space approach*. Beachwood, Ohio: Institute of Mathematical Statistics.

Fox, C., H. Haario, and J. Christen (2013). Inverse problems. In P. Damien, P. Dellaportas, N. Polson, and D. Stephens (Eds.), *Bayesian Theory and Applications*, Chapter 31, pp. 619–643. Oxford University Press.

Gallego, A. M. (2016). Análisis de metabolitos presentes en el aliento: Determinación de la linea basal.

Gilmour, S. G. and L. A. Trinca (2012). Bayesian l-optimal exact design of experiments for biological kinetic models. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 61*, 237–251.

Ginsberg, B. H. (2009). Factors affecting blood glucose monitoring: Sources of errors in measurement. *Journal of Diabetes Science and Technology 3*, 903–913.

Guyon, I. and A. Elisseeff (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research 2003*, 1157–1182.

Hoerl, A. E. and R. W. Kennard (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics 12*(1), 55–67.

Huan, X. and Y. M. Marzouk (2014). Gradient-based stochastic optimization methods in Bayesian experimental design. *International Journal for Uncertainty Quantification 4*(6), 479–510.

Jansson, L., L. Lindskog, N. Nordén, S. Carlström, and B. Scherstén (1980). Diagnostic value of the oral glucose tolerance test evaluated with a mathematical model. *Computers and Biomedical Research 13*, 512–521.

Jiang, G. and B. B. Zhang (2003). Glucagon and regulation of glucose metabolism. *American Journal of Physiology-Endocrinology And Metabolism 284*(4), E671–E678.

Jones, E., T. Oliphant, P. Peterson, et al. (2001–). SciPy: Open source scientific tools for Python.

Kaipio, J. and E. Somersalo (2006). *Statistical and computational inverse problems*, Volume 160. New York: Springer Science & Business Media.

Li, Q. and N. Lin (2010). The bayesian elastic net. *Bayesian Analysis 5*, 151–170.

Liang, F., R. Paulo, G. Molina, M. A. Clyde, and J. O. Berger (2008). Mixtures of g priors for bayesian variable selection. *Journal of the American Statistical Association 103*, 410–423.

Lokhorst, J., B. Venables, B. T. port to R, and tests etc: Martin Maechler (2014). *lasso2: L1 constrained estimation aka 'lasso'*. R package version 1.2-19.

Obuchi, Tomoyuki; Kabashima, Y. (2016). Cross validation in lasso and its acceleration. *Journal of Statistical Mechanics: Theory and Experiment 2016*, 053304.

Palumbo, P., S. Ditlevsen, A. Bertuzzi, and A. D. Gaetano (2013). Mathematical modeling of the glucose–insulin system: A review. *Mathematical Biosciences 244*(2), 69 – 81.

Park, C. and Y. J. Yoon (2011). Bridge regression: Adaptivity and group selection. *Journal of Statistical Planning and Inference 141*, 3506–3519.

Park, T. and G. Casella (2008). The bayesian lasso. *Journal of the American Statistical Association 103*, 681–686.

Pericchi, L. R. (2005). *[Handbook of Statistics] Bayesian Thinking - Modeling and Computation Volume 25 —— Model Selection and Hypothesis Testing based on Objective Probabilities and Bayes Factors*, pp. 115–149. Amsterdam Boston: Elsevier.

Petersen, K. B., M. S. Pedersen, et al. (2008). The matrix cookbook. *Technical University of Denmark 7*(15), 510.

Petzold, L. (1983). Automatic selection of methods for solving stiff and nonstiff systems of ordinary differential equations. *SIAM journal on scientific and statistical computing 4*(1), 136–148.

Rencher, A. (2008). *Linear models in statistics*. Hoboken, N.J: Wiley-Interscience.

Ribbing, J., J. Nyberg, O. Caster, and E. N. Jonsson (2007). The lasso—a novel method for predictive covariate model building in nonlinear mixed effects models. *Journal of Pharmacokinetics and Pharmacodynamics 34*, 485–517.

Robbins, H. and S. Monro (1951). A stochastic approximation method. *The Annals of Mathematical Statistics 22*, 400–407.

Sardy, S. (2008). On the practice of rescaling covariates. *International Statistical Review 76*, 285–297.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis*. London: Routledge.

Solonen, A., H. Haario, and M. Laine (2012). Simulation-based optimal design using a response variance criterion. *Journal of Computational and Graphical Statistics 21*, 234–252.

Stamey, Thomas A.and Kabalin, J. N., M. Ferrari, and N. Yang (1989). Prostate specific antigen in the diagnosis and treatment of adenocarcinoma of the prostate. iv. anti-androgen treated patients. *The Journal of Urology 141*, 1088–1090.

Stone, M. (1969). The role of experimental randomization in bayesian statistics: Finite sampling and two bayesians. *Biometrika 56*, 681–683.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B (Methodological) 58*, 267–288.

Wald, A. (1945). Sequential Tests of Statistical Hypotheses. *The Annals of Mathematical Statistics 16*(2), 117–186.

Weaver, B. P., B. J. Williams, C. M. Anderson-Cook, and D. M. Higdon (2016). Computational enhancements to bayesian design of experiments using gaussian processes. *Bayesian Analysis 11*, 191–213.

West, M. (1987). On scale mixtures of normal distributions. *Biometrika 74*(3), 646–648.

Wild, S., G. Roglic, A. Green, R. Sicree, and H. King (2004). Global prevalence of diabetes: Estimates for the year 2000 and projections for 2030. `http://care.diabetesjournals.org/content/27/5/1047.full`.

Yuan, M. and Y. Lin (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 68*, 49–67.

Zhang, Yao; Meeker, W. Q. (2006). Bayesian methods for planning accelerated life tests. *Technometrics 48*, 49–60.

Zou, H. and T. Hastie (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67*, 301–320.